Electronic Journal of Statistics
Vol. 11 (2017) 2397–2423
ISSN: 1935-7524
DOI: 10.1214/17-EJS1291

SiAM: A hybrid of single index models and additive models

Shujie Ma[*]

Department of Statistics
University of California
Riverside, CA 92521, USA
e-mail: shujie.ma@ucr.edu

Heng Lian[†]

Department of Mathematics
City University of Hong Kong
Kowloon Tong, Hong Kong
e-mail: henglian@cityu.edu.hk

Hua Liang[‡]

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

and

Raymond J. Carroll[§]

Department of Statistics
Texas A&M University
3143 TAMU, College Station, TX 77843-3143, USA
and School of Mathematical and Physical Sciences
University of Technology Sydney
Broadway NSW 2007, Australia
e-mail: carroll@stat.tamu.edu

Abstract: While popular, single index models and additive models have potential limitations, a fact that leads us to propose SiAM, a novel hybrid combination of these two models. We first address model identifiability under general assumptions. The result is of independent interest. We then develop an estimation procedure by using splines to approximate unknown functions and establish the asymptotic properties of the resulting estimators. Furthermore, we suggest a two-step procedure for establishing

[*]Ma's research was supported by NSF grant DMS 1306972.
[†]Lian's research was supported by City University of Hong Kong Start Up grant 7200521.
[‡]Corresponding author. Liang's research was partially supported by NSF grants DMS-1418042 and DMS-1620898, and by Award Number 11529101, made by National Natural Science Foundation of China.
[§]Carroll's research was supported by a grant from the National Cancer Institute U01-CA057030.

confidence bands for the nonparametric additive functions. This procedure enables us to make global inferences. Numerical experiments indicate that SiAM works well with finite sample sizes, and are especially robust to model structures. That is, when the model reduces to either single-index or additive scenario, the estimation and inference results are comparable to those based on the true model, while when the model is misspecified, the superiority of our method can be very great.

MSC 2010 subject classifications: Primary 62G08; secondary 62G20, 62J02, 62F12.

Keywords and phrases: Additive models, global inference, identifiability, misspecification, oracle estimator, partially linear single index models, regression spline, simultaneous confidence band.

Received August 2016.

1. Introduction

Because of the complexity of data sets in practice, there has been much interest in developing statistical analysis tools for problems involving high dimensional covariates. Examples of these models include additive models [AM, 13] and single index models [SiM, 15]. A common feature of these models is that they achieve dimension reduction [30] to circumvent the "curse of dimensionality" [1] while retaining flexibility of the nonparametric regression.

Additive models (AM) [AM, 13, 39, 40] and additive partially linear models [APLM, 13, 17, 19, 25, 32] corresponding to continuous response variables have been well studied in the literature [13]. The latter parsimoniously specifies the relationship between the response variable and some of the covariates in a linear function form, and the relationship between the dependent variable and the remaining covariates in a form of additive nonlinear unknown functions. The APLM enjoys the simplicity property of the linear model and the flexibility of the AM, due to the combination of parametric and nonparametric components. For estimation, [4] applied a backfitting procedure, proposed by [3], to approximate the additive components. [19] proposed to estimate the nonparametric components by polynomial splines [28, 30]. After the spline basis is chosen, the coefficients can be estimated by least squares, leading to great gains computationally when contrasted with backfitting.

Although the AM and APLM are flexible and widely used for data exploration [13, 14, 35], their limitations are also evident from their relatively special structures. For instance, they can be used only for the *additive* case and are unable to reflect interactions of two or more variables, which we may encounter in the analysis of complex biomedical data.

Single index models (SiM), another attempt to gain dimensional reduction, have attracted great attention for estimating a conditional mean function because they relax restrictive assumptions imposed on parametric models of conditional mean functions such as linear or generalized linear models [12, 16], and therefore gain more flexibility. There are various estimation procedures for single-index models. See [15] for a comprehensive survey and various applications

of single-index models. An advantage of the SiM and their various extensions over additive models is that they can take interactions of multiple variables into account, which are frequently encountered in the analysis of complex biomedical data, for example from gene regulatory networks, while still enjoying dimension reduction. However, SiM have their own limitations, in that they assume a common nonlinear structure for the linear combination of predictors with different weights, and can not reflect the nonlinear main effect of each predictor when this feature truly exists.

To overcome the limitations of AM and SiM but still enjoy their advantages, we propose SiAM, a combination of these two structures. However, such a combination immediately raises several concerns: (i) under what assumptions is the resulting model identifiable? (ii) The resulting model contains two classes of nonparametric functions; i.e., the first one for the single index part, and the second one for the individual components. Whether we can equally treat them and similar criteria can be applied is unclear. (iii) Are the resulting models robust? That is when the true model is a sole SiM or AM, does the hybrid model and associated methods for it perform the same (or almost the same) as the true model and its associated methods? Technically, it is much more challenging to develop estimation and inference procedures for such a combination due to the complexity of the model structure. As a result, establishing theory for these procedures, such as asymptotic properties, is much more difficult.

In this paper, we address these concerns, and provide an alternative but more flexible tool for data exploration. To further gain simplicity in the implementation, we apply spline approximation to estimate each unknown component functions. This strategy has been applied in the recent literature of estimation and inference for semiparametric models [33, 39]. Further, we suggest a two-step procedure for establishing confidence bands for the nonparametric additive functions. This procedure enables us to make global inference for the nonparametric functions.

The paper is organized as follows. Section 2 presents the modelling framework, and addresses model identifiability. Section 3 proposes estimation for the single index and nonparametric components. Section 4 establishes asymptotic properties for the resulting estimators. The asymptotic normality for index estimators and the rates of convergence for the nonparametric estimators are developed. Section 5 describes the two-step procedure and presents the simultaneous confidence band. Section 6 illustrates the numerical performance of the proposed method through simulation experiments. The last section provides remarks and discussions. All proofs are provided in an Appendix. Additional tables and graphs are also provided in the Supplemental Material [24].

2. The models and identifiability

To combine additive models and single index models, we propose single index additive models (SiAM), given by

$$E(Y|\mathbf{X}) = \widetilde{g}(\mathbf{X}^\top \boldsymbol{\alpha}) + \widetilde{m}_1(X_1) + \cdots + \widetilde{m}_p(X_p), \tag{1}$$

where $\mathbf{X} = (X_1, \cdots, X_p)^\top$ is the p-dimensional *continuous* covariates, $\mathbf{X}^\top\boldsymbol{\alpha}$ is referred to as the index component, and $\widetilde{g}(\cdot)$, $\widetilde{m}_1(\cdot), \cdots, \widetilde{m}_p(\cdot)$ are unknown smooth functions. We briefly discuss the incorporation of discrete covariates in Section 7.

Since single-index models and additive models are two special cases of SiAM, two obvious constraints for identifiability borrowed from these two special cases are $\|\boldsymbol{\alpha}\| = 1$ with $\alpha_1 > 0$, and $E\{m_j(X_j)\} = 0$, respectively. However, these two constraints alone are not sufficient for identifiability. To see that identifiability can fail even with these two constraints, a simple example is that $2\{(X_1 + X_2)/\sqrt{2}\}^2 - \{X_1^2 - E(X_1^2)\} - \{X_2^2 - E(X_2^2)\} = -2\{(X_1 - X_2)/\sqrt{2}\}^2 + 2E(X_1^2) + 2E(X_2^2) + \{X_1^2 - E(X_1^2)\} + \{X_2^2 - E(X_2^2)\}$.

To achieve identifiability, we first need to decompose m_j as the sum of a linear term and a term orthogonal to the space of linear functions. That is, we write $m_j(x)$ as $a_j + \beta_j x + \widetilde{m}_j(x)$ with $E\{\widetilde{m}_j(X_j)\} = E\{X_j\widetilde{m}_j(X_j)\} = 0$. Since $g(\mathbf{X}^\top\boldsymbol{\alpha}) + \sum_j a_j + \mathbf{X}^\top\boldsymbol{\beta} = g(\mathbf{X}^\top\boldsymbol{\alpha}) + \sum_j a_j + (\boldsymbol{\beta}^\top\boldsymbol{\alpha})\mathbf{X}^\top\boldsymbol{\alpha} + \mathbf{X}^\top\{\boldsymbol{\beta} - (\boldsymbol{\beta}^\top\boldsymbol{\alpha})\boldsymbol{\alpha}\} = \widetilde{g}(\mathbf{X}^\top\boldsymbol{\alpha}) + \mathbf{X}^\top\widetilde{\boldsymbol{\beta}}$ where $\widetilde{g}(x) = g(x) + \sum_j a_j + (\boldsymbol{\beta}^\top\boldsymbol{\alpha})x$, $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} - (\boldsymbol{\beta}^\top\boldsymbol{\alpha})\boldsymbol{\alpha}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, with $\widetilde{\boldsymbol{\beta}}$ orthogonal to $\boldsymbol{\alpha}$, the conditional expectation in an SiAM can be written as

$$g(\mathbf{X}^\top\boldsymbol{\alpha}) + \mathbf{X}^\top\boldsymbol{\beta} + \textstyle\sum_{j=1}^p m_j(X_j), \tag{2}$$

with $\|\boldsymbol{\alpha}\| = 1, \alpha_1 > 0, \boldsymbol{\alpha} \perp \boldsymbol{\beta}, E\{m_j(X_j)\} = E\{X_j m_j(X_j)\} = 0$. We call (2) the canonical form of SiAM. It is natural to require the canonical form to be unique. The following theorem gives sufficient conditions for the parameters to be identified. Although these are not necessary conditions, in view of our previous counterexample these conditions are reasonably weak.

Theorem 2.1. *Suppose X_j $(1 \le j \le p)$ has a density function supported on an interval $\mathcal{S}_j \subseteq R$ and $\mathbf{X}$ has a joint positive density on the interior of $\prod_j \mathcal{S}_j$. Consider (2), with $\|\boldsymbol{\alpha}\| = 1, \alpha_1 > 0, \boldsymbol{\alpha} \perp \boldsymbol{\beta}, E\{m_j(X_j)\} = E\{X_j m_j(X_j)\} = 0$. There are two situations.*

- *(i) g and m_j, $j = 1, \ldots, p$ are second order differentiable. $g'' \not\equiv 0$ on the support of $\mathbf{X}^\top\boldsymbol{\alpha}$. $\boldsymbol{\alpha}$ has at least three nonzero components.*
- *(ii) g and m_j, $j = 1, \ldots, p$ are second order differentiable. g'' is a nonconstant continuous function on the support of $\mathbf{X}^\top\boldsymbol{\alpha}$. $\boldsymbol{\alpha}$ has at least two nonzero components.*

Under either (i) or (ii), $(g, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{m_j\}_{j=1}^p)$ is unique.

3. Estimation

The full SiAM model is

$$Y = g(\mathbf{X}^\top\boldsymbol{\alpha}^0) + \mathbf{X}^\top\boldsymbol{\beta}^0 + \textstyle\sum_{j=1}^p m_j(X_j) + \varepsilon, \tag{3}$$

with $\|\boldsymbol{\alpha}^0\| = 1, \alpha_1^0 > 0, \boldsymbol{\alpha}^0 \perp \boldsymbol{\beta}^0, E\{m_j(X_j)\} = E\{X_j m_j(X_j)\} = 0$, and ε is the error term satisfying $E(\varepsilon\,|\mathbf{X}) = 0$. Define

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{m},g) = E\{Y - \textstyle\sum_{j=1}^{p} m_j(X_j) - g(\mathbf{X}^\top\boldsymbol{\alpha}) - \mathbf{X}^\top\boldsymbol{\beta}\}^2, \tag{4}$$

where $\mathbf{m} = \{m_j(\cdot), j = 1,\ldots,p\}$. Denote $\widetilde{Y} = Y - \sum_{j=1}^{p} m_j(X_j)$, and define

$$\varphi_\alpha(u) = E(\widetilde{Y}\,|\mathbf{X}^\top\boldsymbol{\alpha} = u), \Gamma_\alpha(u) = E(\mathbf{X}\,|\mathbf{X}^\top\boldsymbol{\alpha} = u). \tag{5}$$

By simple calculation,

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{m},g) = L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{m}) = E[\widetilde{Y} - \varphi_\alpha(\mathbf{X}^\top\boldsymbol{\alpha}) - \boldsymbol{\beta}^\top\{\mathbf{X} - \Gamma_\alpha(\mathbf{X}^\top\boldsymbol{\alpha})\}]^2.$$

In what follows, for any vector $\zeta = (\zeta_1,\ldots,\zeta_s)^\top \in R^s$, denote $\|\zeta\|_\infty = \max_{1\le\ell\le s}|\zeta_\ell|$ and $\|\zeta\|_2 = (|\zeta_1|^2 + \cdots + |\zeta_s|^2)^{1/2}$. For any symmetric matrix $\mathbf{A}_{s\times s}$, denote its L_r norm as $\|\mathbf{A}\|_r = \max_{\zeta\in\mathbf{s},\zeta\neq\mathbf{0}} \|\mathbf{A}\zeta\|_r \|\zeta\|_r^{-1}$. For any matrix $\mathbf{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le s}\sum_{j=1}^{t}|A_{ij}|$.

Let $\omega(\boldsymbol{\alpha}) = E[\{\mathbf{X} - \Gamma_\alpha(\mathbf{X}^\top\boldsymbol{\alpha})\}^{\otimes 2}]$ and $\nu(\boldsymbol{\alpha}) = E[\{\mathbf{X} - \Gamma_\alpha(\mathbf{X}^\top\boldsymbol{\alpha})\}\{\widetilde{Y} - \varphi_\alpha(\mathbf{X}^\top\boldsymbol{\alpha})\}]$, where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$ for any matrix $\mathbf{A}$. For given $\boldsymbol{\alpha}$ and $\mathbf{m}$, the corresponding $\boldsymbol{\beta}$ which minimizes $L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{m})$ is $\boldsymbol{\beta}_\alpha = \{\omega(\boldsymbol{\alpha})\}^-\nu(\boldsymbol{\alpha})$, where $\{\omega(\boldsymbol{\alpha})\}^-$ is a generalized inverse of $\omega(\boldsymbol{\alpha})$. According to Theorem 2 of [38], by assuming that $\mathbf{X}$ has a joint positive density function on an open convex subset in R^p, for given $\mathbf{m}$, the minimum point of $L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{m})$ with $\boldsymbol{\alpha}\perp\boldsymbol{\beta}$ is unique at $\boldsymbol{\alpha}^0$ and

$$\boldsymbol{\beta}^0 = \boldsymbol{\beta}_{\alpha^0} = \{\omega(\boldsymbol{\alpha}^0)\}^+\nu(\boldsymbol{\alpha}^0), \tag{6}$$

where $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$ are the true parameters in model (3), and $\{\omega(\boldsymbol{\alpha})\}^+$ is the Moore-Penrose inverse of $\omega(\boldsymbol{\alpha})$.

We approximate the nonparametric functions $g(\cdot)$ and $m_j(\cdot)$ by means of B-splines, and the estimators of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $g(\cdot)$ and $m_j(\cdot)$ can be obtained by minimizing an objective function, which is the sample analog of (4). However, to simultaneously obtain those estimators is computationally very challenging given that the estimates of the parameters and the nonparametric functions intrinsically depend on each other. We then apply an iterative algorithm to minimize the objective function with respect to one parameter vector and fixing the others. The iterative algorithm has been commonly used for estimation in partially linear single-index models (PLSiMs) and it converges well [5, 20, 31, 36, 38]. When the nonparametric functions are given, estimation of the parameters follows the same procedure as given in [38]. When the parameters are fixed, we deal with estimation of nonparametric additive functions instead of nonparametric univariate function in PLSiMs, but they should have the same computational convergence property. Let $\mathbf{X}_i = (X_{ij}, 1\le j\le p)^\top$ and Y_i be the i^{th} realization of $\mathbf{X}$ and Y. The estimation is achieved in three steps:

Step I. For given $\boldsymbol{\alpha}$ and $m_j(\cdot)$, the estimate of $\boldsymbol{\beta}$ is obtained by a sample estimate of (6). We first estimate $\varphi_\alpha(u)$ and $\Gamma_\alpha(u)$ given in (5) by means of B-splines. Let $\mathbf{B}(u) = \{B_1(u),\ldots,B_K(u)$ be a set of q-th order B-spline basis functions, in which K is the number of basis functions which increases with the sample size n, and thus the number of interior knots is $K - q$. Moreover, the distance between

neighboring knots satisfies the assumptions given in [26]. Then the spline estimators of $\varphi_\alpha(u)$ and $\Gamma_\alpha(u)$ are given as $\widehat{\varphi}_\alpha(u) = \mathbf{B}(u)^\top \widehat{\lambda}$ and $\widehat{\Gamma}_\alpha(u) = \mathbf{B}(u)^\top \widehat{\boldsymbol{\Theta}}$, where $\widehat{\lambda}$ and $\widehat{\boldsymbol{\Theta}}$ are spline coefficient estimates obtained from least squares estimation with responses $\widetilde{Y}_i$ and $\mathbf{X}$, respectively. Then the estimate of $\boldsymbol{\beta}$ is given as

$$\widehat{\boldsymbol{\beta}} = \{\sum_{i=1}^{n} (\triangle \mathbf{X}_i)^{\otimes 2}\}^{+} \{\sum_{i=1}^{n} (\triangle \mathbf{X}_i)(\triangle \widetilde{Y}_i)\},$$

where $\triangle \mathbf{X}_i = \mathbf{X}_i - \widehat{\Gamma}_\alpha(\mathbf{X}_i^\top \boldsymbol{\alpha})$ and $\triangle \widetilde{Y}_i = \widetilde{Y}_i - \widehat{\varphi}_\alpha(\mathbf{X}_i^\top \boldsymbol{\alpha})$.

Step II. The estimate of $\boldsymbol{\alpha}$, denoted as $\widehat{\boldsymbol{\alpha}}$, is obtained by minimizing

$$\textstyle\sum_{i=1}^{n} \{Y_i - \sum_{j=1}^{p} \widehat{m}_j(X_j) - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}} - \widehat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha})\}^2,$$

where $\widehat{m}_j(\cdot)$ and $\widehat{g}(\cdot)$ are the spline estimates of $m_j(\cdot)$ and $g(\cdot)$ from the previous step of the iteration. Then we let $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}/\|\widehat{\boldsymbol{\alpha}}\|$.

Step III. For given $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, we estimate $g(u)$ and $m_j(x_j)$ by the spline estimates $\widehat{g}(u) = \mathbf{B}(u)^\top \widehat{\boldsymbol{\gamma}}$ and $\widehat{m}_j(x_j) = \mathbf{b}_j(x_j)^\top \widehat{\boldsymbol{\delta}}_j$ with $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\delta}}_j$ minimizing

$$\textstyle\sum_{i=1}^{n} \{Y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}} - \mathbf{B}\left(\mathbf{X}_i^\top \widehat{\boldsymbol{\alpha}}\right)^\top \boldsymbol{\gamma} - \sum_{j=1}^{p} \mathbf{b}_j(X_{ij})^\top \boldsymbol{\delta}_j\}^2,$$

where $\mathbf{b}_j(x_j) = \{b_{j1}(x_j), \ldots, b_{jL}(x_j)\}^\top$ are sets of basis functions for $j = 1, \ldots, p$ defined as follows. Let

$$\mathbf{B}_j(x_j) = \{B_{j1}(x_j), \ldots, B_{jL}(x_j)\}^\top \tag{7}$$

be sets of q-th order B-spline basis functions for $j = 1, \ldots, p$. To ensure that $E_n\{\widehat{m}_j(X_j)\} = E_n\{X_j \widehat{m}_j(X_j)\} = 0$, where $E_n(\cdot)$ denotes the empirical average, we let $b_{j\ell}(x_j) = B_{j\ell}(x_j) + a_\ell + b_\ell x_j$, with

$$\begin{aligned} b_\ell &= \frac{n^{-1}\sum_{i=1}^{n} B_{j\ell}(X_{ij}) \sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} B_{j\ell}(X_{ij}) X_{ij}}{\sum_{i=1}^{n} X_{ij}^2 - n^{-1}(\sum_{i=1}^{n} X_{ij})^2}, \\ a_\ell &= -n^{-1}\textstyle\sum_{i=1}^{n} B_{j\ell}(X_{ij}) - b_\ell n^{-1} \sum_{i=1}^{n} X_{ij}. \end{aligned} \tag{8}$$

We iterate Steps I–III until convergence. The initial estimates $\widehat{\boldsymbol{\alpha}}^{\text{ini}}$, $\widehat{\boldsymbol{\beta}}^{\text{ini}}$ and $\widehat{g}^{\text{ini}}(\cdot)$ of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $g(\cdot)$ are obtained by fitting the partially linear single-index model: $Y = g(\mathbf{X}^\top \boldsymbol{\alpha}) + \mathbf{X}^\top \boldsymbol{\beta} + \varepsilon_1^*$ by the method used in [38]. The initial estimates of m_j are obtained by fitting the additive model: $Y - \widehat{g}^{\text{ini}}(\mathbf{X}^\top \widehat{\boldsymbol{\alpha}}^{\text{ini}}) - \mathbf{X}^\top \widehat{\boldsymbol{\beta}}^{\text{ini}} = \sum_{j=1}^{p} m_j(X_j) + \varepsilon_2^*$,

4. Asymptotic properties

In this section, we study the large-sample properties of the SiAM parameter estimators, which are obtained by minimizing the objective function

$$L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \textstyle\sum_{i=1}^{n} [Y_i - \mathbf{B}^*(\mathbf{X}_i)^\top \boldsymbol{\delta} - \mathbf{B}\left(\mathbf{X}_i^\top \boldsymbol{\alpha}\right)^\top \boldsymbol{\gamma} - \mathbf{X}_i^\top \boldsymbol{\beta}]^2, \tag{9}$$

where $\mathbf{B}^*(\mathbf{X}_i) = \{\mathbf{b}_1(X_{i1})^\top, \ldots, \mathbf{b}_p(X_{ip})^\top\}^\top$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_p^\top)^\top$. For given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the spline coefficient estimators are $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\widehat{\boldsymbol{\delta}}_j(\boldsymbol{\alpha}, \boldsymbol{\beta}), 1 \le j \le p\}^\top$, where

$$\left\{\widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}, \boldsymbol{\beta})^\top, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta})^\top\right\}^\top = \{\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i(\boldsymbol{\alpha})\boldsymbol{\Phi}_i(\boldsymbol{\alpha})^\top\}^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i(\boldsymbol{\alpha})(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}), \tag{10}$$

and

$$\boldsymbol{\Phi}_i(\boldsymbol{\alpha}) = \{\mathbf{B}^*(\mathbf{X}_i)^\top, \mathbf{B}(\mathbf{X}_i^\top \boldsymbol{\alpha})^\top)^\top, \tag{11}$$

and the estimators of the nonparametric functions are $\widehat{g}(u; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{B}(u)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\widehat{m}_j(x_j; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{b}_j(x_j)^\top \widehat{\boldsymbol{\delta}}_j(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We obtain the estimators $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as the minimizers of

$$L_n^*(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \textstyle\sum_{i=1}^n \{Y_i - \mathbf{B}^*(\mathbf{X}_i)^\top \widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{B}\left(\mathbf{X}_i^\top \boldsymbol{\alpha}\right)^\top \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{X}_i^\top \boldsymbol{\beta}\}^2.$$

The spline estimators of the nonparametric functions $g(u)$ and $m_j(x_j)$ are given as $\widehat{g}\left(u; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}\right)$ and $\widehat{m}_j(x_j; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$, respectively.

We next introduce the following notation and definitions. Let $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$ be the true parameters in model (3). Now define the Hilbert space $\mathcal{H}$ as a collection of additive functions with finite L_2 norm on $\mathcal{S}_1 \times \cdots \times \mathcal{S}_p$, where $\mathcal{S}_j$ is the support of X_j, $j = 1, \ldots, p$, by

$$\begin{aligned} \mathcal{H} = \Big\{ & h(\mathbf{x}) = h_0(\mathbf{x}^\top \boldsymbol{\alpha}^0) + \textstyle\sum_{j=1}^p h_j(x_j), Eh_0\left(\mathbf{X}^\top \boldsymbol{\alpha}^0\right)^2 < \infty, \\ & Eh_j(X_j)^2 < \infty, Eh_j(X_j) = 0, EX_j h_j(X_j) = 0 \Big\}, \end{aligned} \tag{12}$$

where $\mathbf{x} = (x_1, \ldots, x_p)^\top$. Moreover, define

$$\mathbb{P}(X_j) = \arg\min_{h \in \mathcal{H}} E\{X_j - h(\mathbf{X})\}^2,$$

and

$$\mathbb{P}(Z_j) = \arg\min_{h \in \mathcal{H}} E\{Z_j - h(\mathbf{X})\}^2,$$

where $Z_j = X_j \dot{g}(\mathbf{X}^\top \boldsymbol{\alpha}^0)$ for $j = 1, \ldots, p$ and $\dot{g}(u) = \partial g(u)/\partial u$. Let

$$\begin{aligned} \mathbb{P}(\mathbf{X}) &= \{\mathbb{P}(X_1), \ldots, \mathbb{P}(X_p)\}^\top, \\ \mathbb{P}(\mathbf{Z}) &= \{\mathbb{P}(Z_1), \ldots, \mathbb{P}(Z_p)\}^\top, \end{aligned} \tag{13}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$, and $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbb{P}(\mathbf{X})$ and $\widetilde{\mathbf{Z}} = \mathbf{Z} - \mathbb{P}(\mathbf{Z})$. Denote $Z_{ij} = X_{ij} \dot{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0)$, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$, $\widetilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbb{P}(\mathbf{X}_i)$ and $\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i - \mathbb{P}(\mathbf{Z}_i)$. For any positive numbers a_n and b_n, let $a_n \ll b_n$ denote that $a_n/b_n = o(1)$, and $a_n \asymp b_n$ means that $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero constant. Let r with $r \ge 2$ be the smoothness order of the coefficient functions $m_\ell(\cdot)$ as given in Condition (C2) in the Appendix. Denote $\mathrm{var}(Y | \mathbf{X} = \mathbf{x}) = \sigma^2(\mathbf{x})$. We assume that $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are in a neighborhood of $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$.

Theorem 4.1. *Under Conditions (C1)–(C4) in the Appendix, and* $K \asymp L \asymp K_n$, $n^{1/(2r+2)} \ll K_n \ll n^{1/4}$, *we have*

$$\sqrt{n}\left\{(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^\top, (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)^\top\right\}^\top = \left\{n^{-1}\sum_{i=1}^n (\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)^\top (\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)\right\}^{-1} \times \left[n^{-1/2}\sum_{i=1}^n \left\{Y_i - g(\mathbf{X}_i^\top \boldsymbol{\alpha}_0) - \mathbf{X}_i^\top \boldsymbol{\beta}_0 - \textstyle\sum_{j=1}^p m_j(X_j)\right\}(\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)^\top\right] + o_p(1).$$

Consequently, $\sqrt{n}\left\{(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^\top, (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)^\top\right\}^\top \to_d Normal(\mathbf{0}, \Sigma)$, *as* $n \to \infty$, *where*

$$\Sigma = \left[E\left\{(\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)^\top (\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)\right\}\right]^{-1} \left[E\left\{\sigma^2(\mathbf{X})(\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)^\top (\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)\right\}\right] \left[E\left\{(\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)^\top (\widetilde{\mathbf{Z}}^\top, \widetilde{\mathbf{X}}^\top)\right\}\right]^{-1}. \tag{14}$$

Remark 1. *The asymptotic expansion stated in Theorem 4.1 can be used to conduct inferences for the parameters such as constructing confidence intervals and Wald test statistics. We estimate* $\mathbb{P}(X_{ij})$, $\mathbb{P}(Z_{ij})$, $1 \le j \le p$, $1 \le i \le n$, *by the spline estimator*

$$\begin{aligned}\widehat{\mathbb{P}}(X_{ij}) &= \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})^\top \{\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})\boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})^\top\}^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}}) X_{ij},\\ \widehat{\mathbb{P}}(Z_{ij}) &= \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})^\top \{\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})\boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}})^\top\}^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i(\widehat{\boldsymbol{\alpha}}) Z_{ij},\end{aligned}$$

respectively, and the residuals are estimated by

$$\widehat{\varepsilon}_i = Y_i - \textstyle\sum_{j=1}^p \widehat{m}_j(X_{ij}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \widehat{g}(\mathbf{X}_i^\top \widehat{\boldsymbol{\alpha}}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}. \tag{15}$$

Thus, the covariance matrix Σ *given in (14) is estimated by*

$$\widehat{\Sigma} = \left\{n^{-1}\sum_{i=1}^n (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)^\top (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)\right\}^{-1} \left\{n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i^2 (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)^\top (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)\right\} \times \left\{n^{-1}\sum_{i=1}^n (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)^\top (\widehat{\mathbf{Z}}_i^\top, \widehat{\mathbf{X}}_i^\top)\right\}^{-1},$$

where $\widehat{\mathbf{X}}_i = \mathbf{X}_i - \widehat{\mathbb{P}}(\mathbf{X}_i)$, $\widehat{\mathbf{Z}}_i = \mathbf{Z}_i - \widehat{\mathbb{P}}(\mathbf{Z}_i)$, $\widehat{\mathbb{P}}(\mathbf{X}_i) = \left\{\widehat{\mathbb{P}}(X_{i1}), \ldots, \widehat{\mathbb{P}}(X_{ip})\right\}^\top$, *and* $\widehat{\mathbb{P}}(\mathbf{Z}_i) = \left\{\widehat{\mathbb{P}}(Z_{i1}), \ldots, \widehat{\mathbb{P}}(Z_{ip})\right\}^\top$.

Let $\mathcal{S}_0$ be the support of $\mathbf{X}^\top \boldsymbol{\alpha}^0$. The following theorem presents the global convergence rates of the estimators for the nonparametric functions.

Theorem 4.2. *Under Conditions (C1)–(C4) in the Appendix, and* $K \asymp L \asymp K_n$, $n^{1/(2r+2)} \ll K_n \ll n^{1/4}$, *we have that (i)* $\int_{\mathcal{S}_0} \{\widehat{g}(u; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - g(u)\}^2 du = O_p\left(n^{-1}K_n + K_n^{-2r}\right)$ *and*
(ii) $\sum_{j=1}^p \int_{\mathcal{S}_j} \{\widehat{m}_j(x_j; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - m_j(x_j)\}^2 dx_j = O_p\left(n^{-1}K_n + K_n^{-2r}\right)$.

5. Inference for nonparametric functions

5.1. Goal

Although the one-step spline approximation can quickly estimate multiple nonparametric functions, according to [29], no asymptotic distribution is available for the resulting estimators. In this section, we adopt a refined two-step spline estimation procedure as proposed for additive models in [27] and [21], based on which asymptotic confidence bands are further constructed for global inferences of the nonparametric functions.

5.2. Oracle nonparametric estimator

Next we will describe the oracle estimator for $m_j(x_j)$, and the oracle estimator for $g(u)$ can be defined accordingly. By "oracle" here, we mean the estimation of one of the component functions of the SiAM model when all parameters and all the other functions are known.

Without loss of generality, we let $j = 1$. By assuming that $g(u)$, $m_j(x_j)$ for $j \geq 2$, $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$ are known, we rewrite model (3) as

$$Y_i^1 = Y_i - g(\mathbf{X}_i^\top \boldsymbol{\alpha}^0) - \textstyle\sum_{j\geq 2} m_j(X_{ij}) - \mathbf{X}_i^\top \boldsymbol{\beta}^0 = m_1(X_{i1}) + \varepsilon_i.$$

Thus we obtain the oracle estimator of $m_1(x_1)$ as the least squares spline estimator given as $\widehat{m}_1^{\mathrm{OR}}(x_1) = \widetilde{\mathbf{B}}(x_1)^\top \widehat{\boldsymbol{\delta}}_1^{\mathrm{OR}}$, where $\widetilde{\mathbf{B}}(x_1) = \{\widetilde{B}_{11}(x_1), \ldots, \widetilde{B}_{1\widetilde{L}}(x_1)\}^\top$ is a set of B-spline basis functions with the same spline order as $\mathbf{B}(x_1)$ given in (7) but different number of basis functions $\widetilde{L}$ and

$$\widehat{\boldsymbol{\delta}}_1^{\mathrm{OR}} = \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1} \sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1}) Y_i^1.$$

We propose a smooth simultaneous confidence band (SCB) for $m_1(\cdot)$ by studying the asymptotic behavior of maximum of the normalized deviation of the spline functional estimate. To construct asymptotic SCBs for $m_1(\cdot)$ over the interval $x_1 \in S_1$ with confidence level $100(1-\alpha)\%$, $\alpha \in (0,1)$, we need to find two functions $l_n(x_1)$ and $u_n(x_1)$ such that

$$\lim_{n\to\infty} P(l_n(x_1) \leq m_1(x_1) \leq u_n(x_1) \text{ for all } x_1 \in S_1) = 1 - \alpha. \tag{16}$$

In practice, we consider a variant of (16) and construct SCBs over a finite subset $S_{n,1}$ of S_1 with $S_{n,1}$ becoming denser as $n \to \infty$. Without loss of generality, we let $S_1 = [a, b]$ where a and b are two finite numbers. Thus, we partition $[a, b]$ according to N_n equally spaced points $a < \xi_0 < \xi_1 < \cdots < \xi_{N_n} < \xi_{N_n+1} = b$ where $N_n \to \infty$ as $n \to \infty$. Let $S_{n,1} = (\xi_0, \ldots, \xi_{N_n})$. Define $d_{N_n}(\alpha) = 1 - \{2\log(N_n+1)\}^{-1}[\log\{-(1/2)\log(1-\alpha)\} + (1/2)\{\mathrm{loglog}(N_n+1) + \log(4\pi)\}]$, and $Q_{N_n}(\alpha) = \{2\log(N_n+1)\}^{1/2} d_{N_n}(\alpha)$.

Theorem 5.1. *Under Conditions (C1)–(C4) in the Appendix, and* $N_n \asymp \widetilde{L} \asymp n^{1/(2r+1)}$, *we have*

$$\lim_{n\to\infty} P\left\{\sup_{x_1\in S_{n,1}} \left|\sigma_n(x_1)^{-1}\{\widehat{m}_1^{OR}(x_1) - m_1(x_1)\}\right| \le Q_{N_n}(\alpha)\right\} = 1-\alpha.$$

and thus an asymptotic $100(1-\alpha)\%$ *confidence band for* $m_1(x_1)$ *over* $x_1 \in S_{n,1}$ *is*

$$\widehat{m}_1^{OR}(x_1) \pm \sigma_n(x_1) Q_{N_n}(\alpha), \tag{17}$$

where $\sigma_n^2(x_1) = \widetilde{\mathbf{B}}(x_1)^\top \mathbf{\Xi}_n \widetilde{\mathbf{B}}(x_1)$ *with*

$$\begin{aligned}\mathbf{\Xi}_n \quad = \quad & \{\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\{\textstyle\sum_{i=1}^n \sigma^2(\mathbf{X}_i)\,\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\} \\ & \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}.\end{aligned} \tag{18}$$

Remark 2. *We estimate* $\mathbf{\Xi}_n$ *by*

$$\begin{aligned}\widehat{\mathbf{\Xi}}_n = & \{\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\{\textstyle\sum_{i=1}^n \widehat{\varepsilon}_i^2 \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\} \\ & \times \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1},\end{aligned}$$

where $\widehat{\varepsilon}_i$ *is given in (15).*

Remark 3. *Compared to the pointwise CI with width* $2Z_{1-\alpha/2}\sigma_n(x_1)$, *the width of the confidence bands (17) is inflated by* $\{2log(N_n+1)\}^{1/2} d_{N_n}(\alpha)/Z_{1-\alpha/2}$. *Moreover,* $\sigma_n(x_1) \asymp n^{-r/(2r+1)}\{1+o_p(1)\}$ *uniformly in* $x_1 \in S_{n,1}$.

Remark 4. *To construct the SCB based on Theorem 5.1, we propose a finite sample approximation scheme to compute the cutoff value* $Q_{N_n}(\alpha)$ *as follows. Let* $\eta(\xi_J), 1 \le J \le N_n+1$ *have jointly normal distribution with* $E\{\eta(\xi_J)\} = 0$, $var\{\eta(\xi_J)\} = 1$ *and covariance* $cov\{\eta(\xi_J), \eta(\xi_{J'})\} = \sigma_n^{-1}(\xi_J)\sigma_n^{-1}(\xi_{J'})\{\widetilde{\mathbf{B}}(\xi_J)^\top \mathbf{\Xi}_n \widetilde{\mathbf{B}}(\xi_{J'})\}$ *for* $1 \le J \ne J' \le N_n+1$. *We propose the finite sample cutoff value* $Q^*_{N_n}(\alpha)$ *defined by* $P\left\{\sup_{1\le J\le N_n+1} |\eta(\xi_J)| \le Q^*_{N_n}(\alpha)\right\} = 1-\alpha$. *Thus the cutoff value* $Q^*_{N_n}(\alpha)$ *is the* $100(1-\alpha)^{th}$ *percentile of the absolute maxima distribution of* $\eta(\xi_J)$.

5.3. Two-step estimator

Since $g(u)$, $m_j(x_j)$ for $j \ge 2$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown in reality, we replace the true functions and parameters by their estimators $\widehat{g}(u)$, $\widehat{m}_j(x_j)$ for $j \ge 2$, $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ from Section 3 to obtain the two-step estimator of $m_1(x_1)$, denoted as $\widehat{m}_1^{\mathrm{SS}}(x_1)$. The following theorem gives the uniform efficiency of the two-step estimator $\widehat{m}_1^{\mathrm{SS}}(x_1)$.

Theorem 5.2. *Under Conditions (C1)–(C4) in the Appendix, and* $\widetilde{L} \asymp n^{1/(2r+1)}$, $K \asymp L \asymp K_n$, *and* $n^{1/(2r+1)} \ll K_n \ll n^{1/4}$, *we have*

$$\sup_{x_1 \in S_1} |\widehat{m}_1^{SS}(x_1) - \widehat{m}_1^{OR}(x_1)| = O_p\{(log(n)/n)^{1/2} + K_n^{-r}\} = o_p(n^{-r/(2r+1)}).$$

Remark 5. *Based on the uniform rate given in Theorem 5.2, the difference between the two-step and the oracle estimator is asymptotically negligible, so that the asymptotic* $100(1-\alpha)\%$ *confidence band for* $m_1(x_1)$ *is given as* $\widehat{m}_1^{SS}(x_1) \pm \sigma_n(x_1)\{2log(N_n+1)\}^{1/2} d_{N_n}(\alpha)$. *Moreover, to have the result in Theorem 5.2, we need the spline estimator in the first step to be undersmoothed with the number of basis functions satisfying* $L \gg n^{1/(2r+1)}$. *The number of basis functions in the second step has the optimal order given as* $\widetilde{L} \asymp n^{1/(2r+1)}$.

Remark 6. *As suggested by a reviewer, since the effect of any single covariate appears in both the single-index part and the additive part, it is of interest to make inferences on the function*

$$h(x_1) := g(x_1\alpha_1 + x_2^*\alpha_2 + \cdots + x_p^*\alpha_p) + x_1\beta_1 + x_2^*\beta_2 + \cdots + x_p^*\beta_p + m_1(x_1) + \sum_{j=2}^{p} m_j(x_j^*) \tag{19}$$

where $x_2^*, \cdots, x_p^*$ *are fixed values of* $x_2, \ldots, x_p$. *By an analogy with the inferences for* m_1 *above, we can construct the confidence bands for* (19) *using*

$$\hat{h}_1(x_1) \pm \sigma_h(x_1) Q_{N_n}(\alpha)$$

where $\sigma_h^2(x_1) = \breve{\mathbf{B}}(x_1)^\top \breve{\mathbf{\Xi}}_n \breve{\mathbf{B}}(x_1)$ *with* $\breve{\mathbf{\Xi}}_n = \{\sum_{i=1}^n \breve{\mathbf{B}}(X_{i1})\breve{\mathbf{B}}(X_{i1})^\top\}^{-1}\{\sum_{i=1}^n \sigma^2(\mathbf{X}_i)\breve{\mathbf{B}}(X_{i1})\breve{\mathbf{B}}(X_{i1})^\top\}\{\sum_{i=1}^n \breve{\mathbf{B}}(X_{i1})\breve{\mathbf{B}}(X_{i1})^\top\}^{-1}$, $\breve{\mathbf{B}}(x_1) = (\mathbf{b}(x_1)^\top, \mathbf{b}(x_2^*)^\top, \ldots, \mathbf{b}(x_p^*)^\top, x_1, x_2^*, \ldots, x_p^*, \mathbf{B}(x_1\hat{\alpha}_1 + x_1^*\hat{\alpha}_2 + \cdots + x_p^*\hat{\alpha}_p))^\top$, *and* $\breve{\mathbf{B}}(X_i) = (\mathbf{B}^*(\mathbf{X}_i)^\top, \mathbf{X}_i^\top, \mathbf{B}(\mathbf{X}_i^\top\hat{\boldsymbol{\alpha}})^\top))^\top$. *However, unlike the previous inferences for* m_1, *theoretical investigation of this requires* joint asymptotics *of different estimated components and we are not able to provide a rigorous justification in this paper and will only demonstrate this using simulations.*

6. Simulations

We conducted simulation studies to investigate the finite sample performance of SiAM, the additive model (AM) and the partially linear single-index model (PLSiM). The algorithm is implemented in R and the code can be obtained from the second author upon request. The first example is

Example 1.

$$Y_i = g(\mathbf{X}_i^\top\boldsymbol{\alpha}) + \sum_{j=1}^{4}\{m_j(X_{ij}) - Em_j(X_{ij})\} + \varepsilon_i, i = 1, \ldots, n,$$

with $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$. We set $\boldsymbol{\alpha} = (2, 1, -2, -1)^\top / \sqrt{10}$, $g(x) = 10x^2$, $m_1(x) = 2\sin(5x)/\{2 - \sin(5x)\}$, $m_2(x) = 4x\exp(-x^2)$, $m_3(x) = \exp(5x - 4)$, $m_4(x) = 3x\sin(4x - 4)$. The above generating model is not presented in the canonical form. The associated canonical form (2) can be found numerically. For example, we find by numerical integration that $\boldsymbol{\beta} = (-1.236, 2.034, 0.938, -2.313)^\top$.

We generated the covariates from a multivariate Gaussian distribution with $\text{cov}(X_{ij}, X_{ij'}) = \rho^{|j-j'|}$ and marginally transformed the covariates into $[0, 1]$ using the cumulative distribution function of the standard normal distribution. We use cubic B-splines with the number of internal knots equal to $\lfloor n^{1/9} \rfloor$, which is the theoretically optimal order when using cubic splines, and $\lfloor a \rfloor$ denotes the largest integer no greater than a. Although it is possible to choose the number of internal knots in a more data-adaptive way, the strategy of using such fixed choice is much more convenient and even a small number of internal knots can provide a flexible fit to various functions and is thus commonly adopted in the literature of regression splines [21, 27, 34].

We let $\rho = 0.2$ and choose $n \in \{200, 500, 1000\}$ and $\sigma = 0.5$, a total of 3 settings. In each setting, 100 data sets are generated and fitted using SiAM.

TABLE 1
Estimated standard errors of the parameters for data simulated in Example 1 in Section 6.

(n, σ)		$\boldsymbol{\alpha}$				$\boldsymbol{\beta}$			
$(200, 0.5)$	Est	0.038	0.025	0.036	0.028	0.174	0.132	0.136	0.165
	Emp	0.045	0.028	0.041	0.031	0.218	0.169	0.162	0.206
$(500, 0.5)$	Est	0.025	0.016	0.024	0.017	0.119	0.083	0.085	0.116
	Emp	0.027	0.018	0.026	0.019	0.138	0.102	0.108	0.130
$(1000, 0.5)$	Est	0.019	0.010	0.017	0.012	0.098	0.067	0.073	0.091
	Emp	0.020	0.012	0.019	0.014	0.102	0.072	0.079	0.095

First we consider the estimation of standard errors for the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. It is easy to obtain standard error estimates based on the asymptotic normality results. On each generated data set, we can get an estimate of standard errors and the average of these over 100 data sets are reported in Table 1, on rows indicated by "Est". The sample standard errors of the estimated parameter values based on 100 data sets are taken as the empirical standard errors, reported on rows indicated by "Emp". It is seen that the estimated standard errors are reasonably close to the empirical values, especially for large sample size.

For an illustration of the construction of the confidence band, Figure 1 and the Supplemental Material Figures 6 and 7 show visually the 95% confidence bands obtained on one data set for $\sigma = 1$, for functions $g, m_1, \ldots, m_4$, as well as h defined in (19) with $x_2^* = \cdots = x_p^* = 1/2$. To construct these bands, except for h which only uses a one-step estimator, we use $\lfloor 2n^{1/7} \rfloor + 1$ internal knots for the first-stage estimator and use $\lfloor n^{1/9} \rfloor + 1$ internal knots in the second stage, which takes a similar form as recommended in previous works such as [27]. We set $N_n = 20$. To investigate the coverage of the confidence bands, 500 data sets are generated in each parameter setting and the results are reported in Table 2. The coverage improves with sample size. The more severe under-coverage of the band for m_1 with $n = 200$ is possibly due to the relatively large bias in

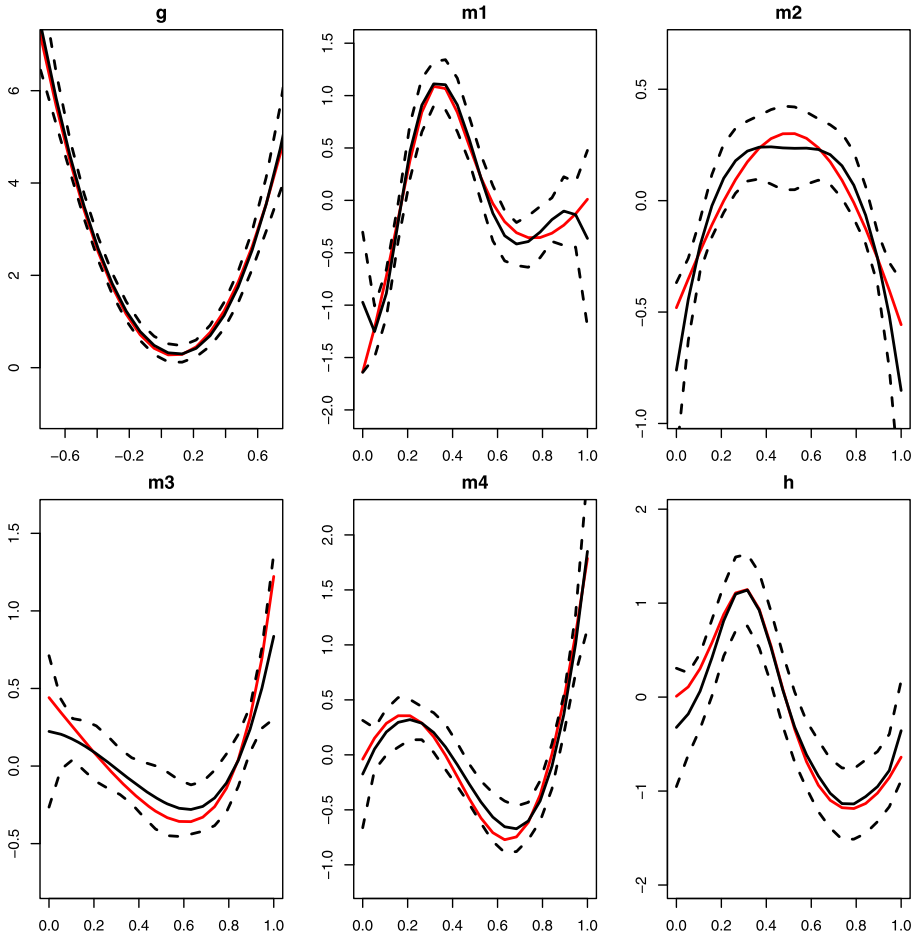

FIG 1. *95% confidence band for the nonparametric functions for Example 1 in Section 6 when* $n = 500$.

TABLE 2
Coverage of 95% *confidence bands for data simulated in Example 1.*

(n, σ)	g	m_1	m_2	m_3	m_4	h
$(200, 0.5)$	0.842	0.764	0.914	0.826	0.878	0.834
$(500, 0.5)$	0.918	0.812	0.916	0.890	0.902	0.874
$(1000, 0.5)$	0.930	0.926	0.932	0.900	0.926	0.918

estimation for this sample size, which can also be seen in Figures 3–5 of the Supplemental Material, for example.

We now use this example to illustrate that the performances actually critically depend on the choice of a good initial estimator. As mentioned before, our initial estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained as in [38]. Under the setting of $n = 500$, we add independent normal perturbation errors with standard deviation $\sigma_p =$

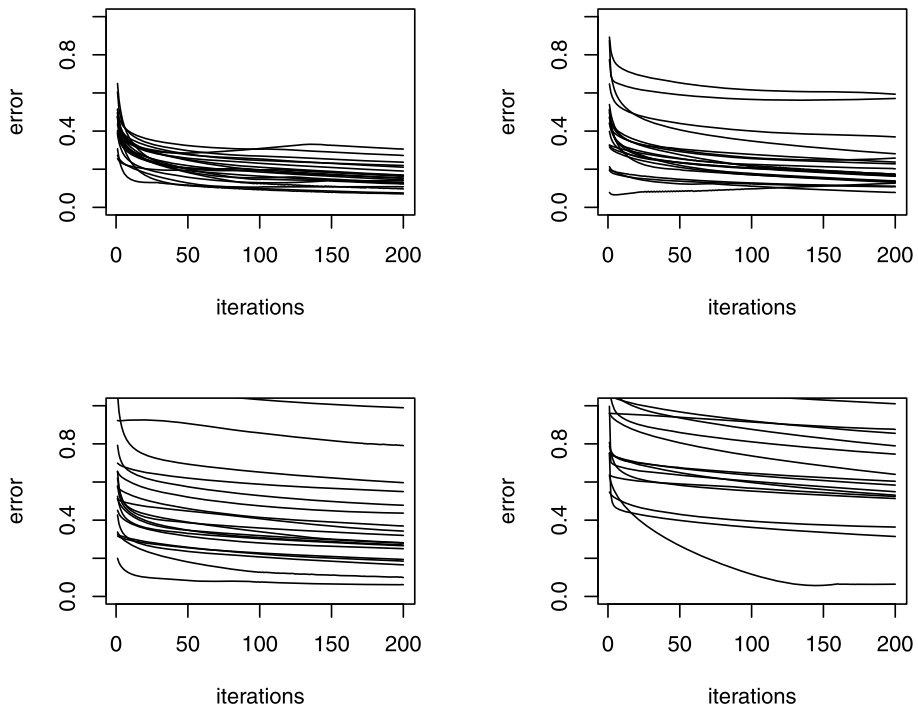

FIG 2. *Trajectory of estimation error versus iterations. Upper left, upper right, lower left and lower right panels show results for* $\sigma_p = 0, 0.1, 0.2, 0.4$, *respectively.*

$0.1, 0.2, 0.4$ to each component of the initial values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, normalize $\boldsymbol{\alpha}$ and orthogonalize $\boldsymbol{\beta}$ with respect to $\boldsymbol{\alpha}$, and then use these as initial values instead. The trajectory of the estimator error $\sqrt{\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2/p + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p}$ varying with iterations are shown in Figure 2, together with the results using initial estimator based on [38] without perturbation. It is seen that even with a small perturbation $\sigma_p = 0.1$ the results become worse and do not seem to converge to the correct value. Thus a good selection of initial values are important in the estimation, and one may even use multiple starting values to safeguard estimation.

The Supplemental Material has additional results. We report the estimation errors of $\boldsymbol{\alpha}, \boldsymbol{\beta}, g, m_1, \ldots, m_4$ in Table 4 (the quantities here refer to those in the canonical form). For $\boldsymbol{\alpha}$ the estimation error is defined as $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|$ and similarly for $\boldsymbol{\beta}$. The estimation error of g is defined by $\sqrt{\sum_{j=1}^{200}\{\widehat{g}(t_j) - g(t_j)\}^2/200}$, where $t_1 < t_2 < \cdots < t_{200}$ are equally spaced grid points on the range of $\mathbf{X}_i^\top \widehat{\boldsymbol{\alpha}}$. Similarly the estimation error for m_j is $\sqrt{\sum_{j=1}^{200}\{\widehat{m}(x_j) - m(x_j)\}^2/200}$ with 200 grid points on $[0, 1]$. From Table 4 of the Supplemental Material, we see that as sample size increases or the noise decreases, the estimation errors become smaller, as expected. For $\sigma = 1$, in Figures 3–5 of the Supplemental Material, we show the estimated nonparametric curves for 20 generated data sets, to-

gether with the truth (solid red curve), to visually illustrate how the estimation accuracy improves with sample size.

Our next aim is to compare the performance of SiAM with two of its special cases, PLSiM and AM. We generated data from the following three examples. Examples 2 and 3 correspond to PLSiM and AM, respectively, and Example 4 represents a more general model that is actually not within the class of SiAM.

Example 2. $Y_i = g(\mathbf{X}_i^\top \boldsymbol{\alpha}) + \mathbf{X}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$, where $g(x) = 10x^2$, $\boldsymbol{\alpha} = (2, 1, -2, -1)^\top / \sqrt{10}$, and $\boldsymbol{\beta} = (2, 2, 2, 2)^\top$.

Example 3. $Y_i = 1 + \sum_{j=1}^4 \{m_j(X_{ij}) - Em_j(X_{ij})\} + \varepsilon_i$, $i = 1, \ldots, n$, where $m_j, j = 1, \ldots, 4$ are the same as in Example 1.

Example 4. $Y_i = -4 + \{\mathbf{X}_i^\top \boldsymbol{\alpha} + \sum_{j=1}^4 2(X_{ij} - 0.2)^2\}^2 + \varepsilon_i$, $i = 1, \ldots, n$, where $\boldsymbol{\alpha} = (2, 1, -2, -1)^\top / \sqrt{10}$.

We consider the same parameter settings as before and in each case generate 100 data sets. Whichever example is the true generating model, we fit the data using the three different models: SiAM, PLSiM and AM. Of course we expect that the estimation would be best when the model used in fitting matches the true generating model. However, calculating the estimation errors is generally not appropriate in comparing different models. For example if the true model is SiAM while an AM is applied in model fitting, it is expected that the estimator is consistent for estimating the "best approximation" of SiAM using AM, which is not necessarily the additive part in the true generating model. In particular, it is difficult to find numerically what quantity the AM is trying to estimate when the true model is SiAM.

Thus we compare the performance of different methods in terms of their prediction accuracy by generating independently 500 observations from the true model. The prediction error is define to be $\sqrt{\sum_{i=1}^{500}(Y_i - \widehat{Y}_i)^2/500}$ where $\widehat{Y}_i$ is the predicted response value and Y_i is the generated true response. The prediction errors are reported in Table 3. We can see that among three different fitting methods, the prediction errors are smallest when the true model is used in data fitting. However, the prediction errors for SiAM are close to the best fitting model for Examples 2 and 3, and much smaller than AM and PLSiM in Example 1. This illustrates that the cost of overfitting using a more flexible model is relatively small compared to the cost of misspecification. Finally, in Example 4 for which all fitting models are misspecified, SiAM still has by far the smallest prediction error, which is more obvious in the low-noise setting.

7. Discussion

In this paper, we have proposed a new model, SiAM, that combines the additive model (AM) and single index model (SiM), and have developed statistical theory for model identifiability. We have further developed a two-step procedure for making global inferences for nonparametric functions. In brief, the model and the proposed methods have the following properties: (i) the estimators of the

TABLE 3
Prediction errors (average errors with standard deviations inside brackets on 100 simulated data sets) by fitting three different models when the data are generated from Examples 1–4 in Section 6.

(n, σ)	Fitting Method		
	SiAM	PLSiM	AM
Example 1			
$(200, 0.5)$	0.56(0.012)	1.01(0.023)	1.13(0.034)
$(500, 0.5)$	0.54(0.005)	0.99(0.011)	1.09(0.017)
$(1000, 0.5)$	0.50(0.003)	0.99(0.007)	1.06(0.012)
Example 2			
$(200, 0.5)$	0.52(0.019)	0.50(0.010)	1.11(0.035)
$(500, 0.5)$	0.50(0.010)	0.49(0.004)	1.07(0.017)
$(1000, 0.5)$	0.49(0.004)	0.49(0.005)	1.06(0.013)
Example 3			
$(200, 0.5)$	0.59(0.106)	1.04(0.131)	0.55(0.010)
$(500, 0.5)$	0.59(0.117)	1.03(0.135)	0.53(0.004)
$(1000, 0.5)$	0.55(0.157)	1.03(0.203)	0.49(0.002)
Example 4			
$(200, 0.5)$	0.62(0.016)	1.03(0.029)	1.07(0.027)
$(500, 0.5)$	0.60(0.007)	1.00(0.014)	1.03(0.012)
$(1000, 0.5)$	0.59(0.006)	0.99(0.011)	1.02(0.008)

index parameters have been shown to be asymptotically normal, and the estimators of the nonparametric functions have optimal rates of convergence; (ii) the two-step estimators have the oracle property; (iii) the proposed methods show promising performance in finite sample situations; and (iv) the implemented algorithm is computationally efficient. Using regression splines, the implementation of the method is much simpler than that of the backfitting-based or profile-based estimation.

Because SiAM contains the single index component and additive components, it can detect interactions among the covariates as well as uncover possible nonlinear main effects, while SiM or AM can only achieve one or the other. Thus, the proposed model is more flexible than those two models.

As a starting point, SiAM can be used for flexible exploratory analysis. If no main effect is detected, one may simplify the model to a SiM, and if the single index component is not significant, one may reduce the model to an additive one. It appears possible that SiAM can be modified by using penalization to develop a variable selection procedure for identifying which elements should enter in the index component, and which ones can be treated as additive components: we will consider this in future work. Moreover, as a future research topic, it would be interesting to develop a method for adaptively choosing among the AM, SiM and our proposed SiAM. Such a model selection problem can be tackled by the "structural adaptation" strategy proposed in [11]. The detailed method and the associated theories need a further investigation.

In this article, we have focused on modeling with fixed dimensional covariates. It is of interest to extend the methods to high dimensional SiAM. However, the theory and implementation of such an extension is much more complicated and warrants further study.

In this article we developed our theory assuming that all predictors are continuous. If some predictors are discrete, they can be added to the linear part of the model: it is straightforward to extend our theory to this slightly more general case, with a few complications of notation only.

Software

The algorithm is implemented in R and the code can be obtained from the second author upon request.

Appendix

A.1. An example about identifiability

Here we give an example to illustrate that if data are generated from the SiAM model, while we use an additive model to fit the data, then the results will be totally misleading.

Suppose the data are generated from $E(Y|\mathbf{X}) = 2X_1X_2 + 2X_2X_3 + 2X_1X_3 - 2X_1 - 2X_2 - 2X_3 + 3/2$, which can be re-expressed as $(X_1 + X_2 + X_3)^2 - (X_1^2 + X_2^2 + X_3^2 + 2X_1 + 2X_2 + 2X_3 - 3/2)$, with $X_1, X_2, X_3 \overset{i.i.d.}{\sim} Unif(0,1)$. This is a SiAM. If we try to use an additive model to fit the data instead of SiAM. We will see what exactly can we obtain.

Note that for any functions m_j, $j = 1, 2, 3$ with $E\{m_j^2(X_j)\} < \infty$, we have

$$\begin{aligned} & E[\{(X_1 - 1/2)(X_2 - 1/2) - m_1(X_1) - m_2(X_2) - m_3(X_3)\}^2] \\ = \; & E\{(X_1 - 1/2)^2(X_2 - 1/2)^2\} + E[\{m_1(X_1) + m_2(X_2) + m_3(X_3)\}^2] \\ & \quad -2\sum_{j=1}^{3} E\{(X_1 - 1/2)(X_2 - 1/2)m_j(X_j)\} \\ = \; & E\{(X_1 - 1/2)^2(X_2 - 1/2)^2\} + E[\{m_1(X_1) + m_2(X_2) + m_3(X_3)\}^2] \\ \geq \; & E\{(X_1 - 1/2)^2(X_2 - 1/2)^2\}, \end{aligned}$$

where the second equality above used the fact that the covariates are independent and $EX_1 = EX_2 = 1/2$. The above means that the best additive approximation of $(X_1 - 1/2)(X_2 - 1/2)$ is zero function and thus the best additive approximation of X_1X_2 is $X_1/2 + X_2/2 - 1/4$. From this, we immediately see the best additive approximation of $E(Y|\mathbf{X}) = (X_1 + X_2 + X_3)^2$ is $X_1^2 + X_2^2 + X_3^2 + 2X_1 + 2X_2 + 2X_3 - 3/2$. As a result, we could mistakenly use a the zero function to approximate $E(Y|\mathbf{X}) = (X_1 + X_2 + X_3)^2 - (X_1^2 + X_2^2 + X_3^2 + 2X_1 + 2X_2 + 2X_3 - 3/2)$ if we try to use additive model to fit the data instead of SiAM in the limit of $n \to \infty$.

A.2. Assumptions

Denote the space of r-th 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 $\mathcal{X}_w$, i.e.,

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

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) The density function $f_{U(\alpha)}(\cdot)$ of the random variable $U(\alpha) = \mathbf{X}^\top \alpha$ is bounded away from 0 on $\mathcal{S}_U$ and $f_{U(\alpha)}(\cdot) \in C^{0,1}(\mathcal{S}_U)$ for α in the neighborhood of α^0, where $\mathcal{S}_U = \{\mathbf{X}^\top \alpha, \mathbf{X} \in S\}$ and S is a compact support set of $\mathbf{X}$, and the density function $f_{X_j}(x_j)$ of random variable X_j is bounded away from 0 on the support S_j of X_j for $j = 1, \ldots, p$.

(C2) The nonparametric functions $g \in C^{(r)}(\mathcal{S}_U)$ and $m_j \in C^{(r)}(S_j)$, $1 \leq j \leq p$, for some integer $r \geq 2$, and the spline order q satisfies $q \geq r$.

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

(C4) The functions h_0 and h_j given in (12) satisfy $h_0 \in C^{(1)}(\mathcal{S}_U)$ and $h_j \in C^{(1)}(\mathcal{S}_j)$, $j = 1, \ldots, p$.

Conditions (C1)–(C4) are commonly used in the nonparametric smoothing literature; for example, see [7] and [41].

A.3. Proof of Theorems 2.1

Suppose we have other variables $(f, \boldsymbol{\theta}, \boldsymbol{\eta}, \{f_j\}_{j=1}^p)$ satisfying the same constraints. We have

$$g(\mathbf{X}^\top \boldsymbol{\alpha}) + \mathbf{X}^\top \boldsymbol{\beta} + m_1(X_1) + \cdots + m_p(X_p) = f(\mathbf{X}^\top \boldsymbol{\theta}) + \mathbf{X}^\top \boldsymbol{\eta} + h_1(X_1) + \cdots + h_p(X_p). \tag{A.1}$$

Taking second derivatives with respect to $\mathbf{X}$, we get

$$g''(\mathbf{X}^\top \boldsymbol{\alpha}) \boldsymbol{\alpha} \boldsymbol{\alpha}^\top + \text{diag}\{m_1''(X_1), \ldots, m_p''(X_p)\} = f''(\mathbf{X}^\top \boldsymbol{\theta}) \boldsymbol{\theta} \boldsymbol{\theta}^\top + \text{diag}\{h_1''(X_1), \ldots, h_p''(X_p)\}.$$

The above displayed equation means $g''(\mathbf{X}^\top \boldsymbol{\alpha}) \boldsymbol{\alpha} \boldsymbol{\alpha}^\top - f''(\mathbf{X}^\top \boldsymbol{\theta}) \boldsymbol{\theta} \boldsymbol{\theta}^\top$ is a diagonal matrix.

First consider assumption (i). Without loss of generality we assume $\alpha_1 > 0, \alpha_2 \neq 0, \alpha_3 \neq 0$. By looking at the off-diagonal entries of the 3×3 principal submatrix of $g''(\mathbf{X}^\top \boldsymbol{\alpha}) \boldsymbol{\alpha} \boldsymbol{\alpha}^\top - f''(\mathbf{X}^\top \boldsymbol{\theta}) \boldsymbol{\theta} \boldsymbol{\theta}^\top$, we get

$$g''(\mathbf{X}^\top \boldsymbol{\alpha}) \alpha_1 \alpha_2 \quad = \quad f''(\mathbf{X}^\top \boldsymbol{\theta}) \theta_1 \theta_2,$$

$$\begin{aligned} g''(\mathbf{X}^\top\boldsymbol{\alpha})\alpha_2\alpha_3 &= f''(\mathbf{X}^\top\boldsymbol{\theta})\theta_2\theta_3, \\ g''(\mathbf{X}^\top\boldsymbol{\alpha})\alpha_1\alpha_3 &= f''(\mathbf{X}^\top\boldsymbol{\theta})\theta_1\theta_3. \end{aligned} \tag{A.2}$$

Take $\mathbf{X}$ such that $g''(\mathbf{X}^\top\boldsymbol{\alpha}) \neq 0$. From the above, the assumption that $\alpha_j \neq 0, j = 1,2,3$ implies $\theta_j \neq 0, j = 1,2,3$. We also have $\alpha_1/\theta_1 = \alpha_2/\theta_2 = \alpha_3/\theta_3$. Furthermore, looking at other off-diagonal entries of $g''(\mathbf{X}^\top\boldsymbol{\alpha})\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - f''(\mathbf{X}^\top\boldsymbol{\theta})\boldsymbol{\theta}\boldsymbol{\theta}^\top$, we get $g''(\mathbf{X}^\top\boldsymbol{\alpha})\alpha_1\alpha_j = f''(\mathbf{X}^\top\boldsymbol{\theta})\theta_1\theta_j$, which implies either $\alpha_j = \theta_j = 0$ or $\alpha_1/\theta_1 = \alpha_j/\theta_j$, for all j. By the constraint $\|\boldsymbol{\alpha}\| = 1$ and $\alpha_1 > 0$ (similarly for $\boldsymbol{\theta}$), we get $\boldsymbol{\alpha} = \boldsymbol{\theta}$. By (A.2) this also implies $g''(\mathbf{X}^\top\boldsymbol{\alpha}) = f''(\mathbf{X}^\top\boldsymbol{\alpha})$. We also get $m_j''(X_j) = h_j''(X_j)$. Considering the constraints $Em_j(X_j) = EX_jm_j(X_j) = 0$, this implies $m_j = h_j$. Now (A.1) implies $g(\mathbf{X}^\top\boldsymbol{\alpha}) + \mathbf{X}^\top\boldsymbol{\beta} = f(\mathbf{X}^\top\boldsymbol{\theta}) + \mathbf{X}^\top\boldsymbol{\eta}$. The identifiability follows from that of partially linear single-index models [18, 37].

Now consider assumption (ii). Without loss of generality we assume $\alpha_1 > 0, \alpha_2 \neq 0$. In this case similar to (i) we have

$$g''(\mathbf{X}^\top\boldsymbol{\alpha})\alpha_1\alpha_2 = f''(\mathbf{X}^\top\boldsymbol{\theta})\theta_1\theta_2. \tag{A.3}$$

By our assumption that g'' is nonconstant and the identifiability of single-index models, we know $\boldsymbol{\alpha} = \boldsymbol{\theta}$ which immediately implies $g''(\mathbf{X}^\top\boldsymbol{\alpha}) = f''(\mathbf{X}^\top\boldsymbol{\alpha})$. The rest of the proof is the same as for (i).

A.4. Proofs of Theorems 4.1 and 4.2

By Bernstein's inequality in [2], we can show that

$$\sup_{1\le\ell\le L}\left|n^{-1}\textstyle\sum_{i=1}^n B_{j\ell}(X_{ij}) - E\{B_{j\ell}(X_j)\}\right| = O_p(\sqrt{\log n/(nL)}) = o_p(1).$$

Thus, for the b_ℓ and a_ℓ defined in (8), we have $\sup_{1\le\ell\le L}|b_\ell| = O_p(L^{-1})$ and $\sup_{1\le\ell\le L}|a_\ell| = O_p(L^{-1})$, so that the basis functions $\mathbf{b}_j(x_j) = \{b_{j1}(x_j), \ldots, b_{jL}(x_j)\}^\top$ with $b_{j\ell}(x_j) = B_{j\ell}(x_j) + a_\ell + b_\ell x_j$ are asymptotically equivalent to the B-spline basis functions $\mathbf{B}_j(x_j) = \{B_{j1}(x_j), \ldots, B_{jL}(x_j)\}^\top$.

The proposition below presents the convergence rate of the estimators $\widehat{g}(u;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$, $\widehat{m}_j(x_j;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$ and $\widehat{\dot{g}}(u;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$, where $\widehat{\dot{g}}(u;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0) = \dot{\mathbf{B}}(u)^\top\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$ which is the estimator of the first derivative of $g(u)$.

Proposition A.1. *Under Conditions (C1)–(C4), and $K \asymp L \asymp K_n$ and $K_n \to \infty$ and $nK_n^{-1} \to \infty$, as $n \to \infty$, one has (i) $\left|\widehat{g}(u;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0) - g(u)\right| = O_p\left(n^{-1/2}K_n^{1/2} + K_n^{-r}\right)$ uniformly for any $u \in \mathcal{S}_U$; (ii) $\left|\widehat{m}_j(x_j;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0) - m_j(x_j)\right| = O_p\left(n^{-1/2}K_n^{1/2} + K_n^{-r}\right)$ uniformly for any $x_j \in S_j$; and (iii) under $K_n \to \infty$ and $nK_n^{-3} \to \infty$, as $n \to \infty$, $\left|\widehat{\dot{g}}(u;\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0) - g(u)\right| = O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-r+1}\right)$ uniformly for any $u \in \mathcal{S}_U$.*

Proof. The proofs follow similar procedures as given in [41] and [42], and thus are omitted. □

Lemma A.1. *Under Conditions (C1)–(C4) in the Appendix, and* $K \asymp L \asymp K_n$, $n^{1/(2r+2)} \ll K_n \ll n^{1/4}$, *we have*

$$\left\| -\dot{L}_n^*(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - 2\sum_{i=1}^n \{Y_i - \sum_{j=1}^p m_j(X_{ij}) - g(\mathbf{X}_i^\top \boldsymbol{\alpha}^0) - \mathbf{X}_i^\top \boldsymbol{\beta}^0\}(\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)^\top \right\| = o_p\left(n^{1/2}\right),$$

where $\dot{L}_n^*(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) = \partial \dot{L}_n^*(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)/\partial(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$.

Proof. Let $\boldsymbol{\theta}^0 = (\boldsymbol{\alpha}^{0\top}, \boldsymbol{\beta}^{0\top})^\top$ and $\widehat{\varpi}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{\widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}, \boldsymbol{\beta})^\top, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta})^\top\right\}^\top$. By the definition of $L_n^*(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we have

$$\begin{aligned} -(1/2)\,\dot{L}_n^*(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) = \sum_{i=1}^n [\{\partial \widehat{\varpi}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)^\top/\partial \boldsymbol{\theta}\}\boldsymbol{\Phi}_i(\boldsymbol{\alpha}^0) \\ + \{\mathbf{X}_i^\top \hat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}^\top] \\ \times [Y_i - \mathbf{B}^*(\mathbf{X}_i)^\top \widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - \mathbf{B}\left(\mathbf{X}_i^\top \boldsymbol{\alpha}^0\right)^\top \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - \mathbf{X}_i^\top \boldsymbol{\beta}^0]. \end{aligned}$$

By Proposition A.1, we have for every $1 \le i \le n$,

$$\begin{aligned} |\Lambda_{i1}| &= \Big|\{Y_i - \mathbf{B}^*(\mathbf{X}_i)^\top \widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - \mathbf{B}\left(\mathbf{X}_i^\top \boldsymbol{\alpha}^0\right)^\top \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - \mathbf{X}_i^\top \boldsymbol{\beta}^0\} \\ &\quad - \{Y_i - \sum_{j=1}^p m_j(X_{ij}) - g(\mathbf{X}_i^\top \boldsymbol{\alpha}^0) - \mathbf{X}_i^\top \boldsymbol{\beta}^0\}\Big| \\ &= O_p(\sqrt{K_n/n} + K_n^{-r}). \end{aligned} \tag{A.4}$$

In the following, we will show that

$$\begin{aligned} \|\Lambda_{i2}\| &= \|\{\partial \widehat{\varpi}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)^\top/\partial \boldsymbol{\theta}\}\boldsymbol{\Phi}_i(\boldsymbol{\alpha}^0) + \{\mathbf{X}_i^\top \hat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}^\top \\ &\quad - (\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)^\top\| = O_p\left(n^{-1/2} K_n^{3/2} + K_n^{-1}\right). \end{aligned} \tag{A.5}$$

According to the result on page 149 of [8], there are $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_p^\top)^\top \in R^{pK}$ and $\boldsymbol{\gamma} \in R^L$, such that $\widetilde{m}_j(x_j) = \mathbf{b}_j(x_j)^\top \boldsymbol{\delta}_j$ and $\widetilde{g}(u) = \mathbf{B}(u)^\top \boldsymbol{\gamma}$ satisfying

$$\sup_{x_j \in S_j} |\widetilde{m}_j(x_j) - m_j(x_j)| = O\left(K_n^{-r}\right), \quad \sup_{u \in S_U} |\widetilde{g}(u) - g(u)| = O\left(K_n^{-r}\right). \tag{A.6}$$

Let $\varpi = \left(\boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top\right)^\top$. Let $\boldsymbol{\Phi}_i = \boldsymbol{\Phi}_i(\boldsymbol{\alpha}^0)$. By (10), we have

$$\begin{aligned} \boldsymbol{\Phi}_i^\top \{\partial \widehat{\varpi}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\theta}^\top\} &= \boldsymbol{\Phi}_i^\top [\partial\{\widehat{\varpi}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0) - \varpi\}/\partial \boldsymbol{\theta}^\top] \\ &= \boldsymbol{\Phi}_i^\top [\partial(\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^\top)^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^0 - \boldsymbol{\Phi}_i^\top \varpi)/\partial \boldsymbol{\theta}^\top] \\ &= \Upsilon_{1i} + \Upsilon_{2i} + \Upsilon_{3i}, \end{aligned}$$

where

$$\begin{aligned} \Upsilon_{1i} &= \boldsymbol{\Phi}_i^\top \left\{(\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^\top)^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i \partial(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^0 - \boldsymbol{\Phi}_i^\top \varpi)/\partial \boldsymbol{\theta}^\top\right\}, \\ &= -\boldsymbol{\Phi}_i^\top (n^{-1} \textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^\top)^{-1} n^{-1} \sum_{i=1}^n \boldsymbol{\Phi}_i \{\mathbf{X}_i^\top \hat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}, \end{aligned}$$

$$\Upsilon_{2i} = \mathbf{\Phi}_i^\top \left\{ (n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1} n^{-1} \sum_{i=1}^n (\partial \mathbf{\Phi}_i/\partial \boldsymbol{\theta}^\top)(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^0 - \mathbf{\Phi}_i^\top \varpi) \right\},$$

$$\Upsilon_{3i} = \mathbf{\Phi}_i^\top \left\{ \partial (n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1}/\partial \boldsymbol{\theta}^\top \right\} n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^0 - \mathbf{\Phi}_i^\top \varpi).$$

Let

$$\widetilde{\Upsilon}_{1i} = -\mathbf{\Phi}_i^\top (n^{-1} \sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1} n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i \{\mathbf{X}_i^\top \dot{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}.$$

By Theorem 5.4.2 of [10], [9] and Condition (C1), we can show that

$$\left\|(n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1}\right\|_\infty = O_p(K_n). \tag{A.7}$$

By Proposition A.1, (A.7) and $\left\|n^{-1}\sum_{i=1}^n \mathbf{\Phi}_i\right\|_\infty = O_p(K_n^{-1})$, we have

$$\begin{aligned}
&\|\Upsilon_{1i} - \widetilde{\Upsilon}_{1i}\|_2 \\
&\quad \le \left\|(n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1}\right\|_\infty \left\|n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\right\|_\infty O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-r+1}\right) \\
&\quad = O_p(K_n) O_p(K_n^{-1}) O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-r+1}\right) \\
&\quad = O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-r+1}\right).
\end{aligned} \tag{A.8}$$

Under Condition (C4), we have

$$\left\|\widetilde{\Upsilon}_{1i} + \{\mathbb{P}(\mathbf{Z}_i)^\top, \mathbb{P}(\mathbf{X}_i)^\top\}\right\|_2 = O_p\left(n^{-1/2}K_n^{1/2} + K_n^{-1}\right), \tag{A.9}$$

where $\mathbb{P}(\mathbf{X})$ and $\mathbb{P}(\mathbf{Z})$ are defined in (13). By Proposition A.1, we have

$$\left\|\{\mathbf{X}_i^\top \widehat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}^\top - (\mathbf{Z}_i^\top, \mathbf{X}_i^\top)^\top\right\|_2 = O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-r+1}\right). \tag{A.10}$$

Therefore, by (A.8), (A.9) and (A.10), we have

$$\begin{aligned}
&\left\|\Upsilon_{1i} + \{\mathbf{X}_i^\top \widehat{g}(\mathbf{X}_i^\top \boldsymbol{\alpha}^0; \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0), \mathbf{X}_i^\top\}^\top - (\widetilde{\mathbf{Z}}_i^\top, \widetilde{\mathbf{X}}_i^\top)\right\|_2 \\
&\qquad = O_p\left(n^{-1/2}K_n^{3/2} + K_n^{-1}\right).
\end{aligned} \tag{A.11}$$

Moreover,

$$\begin{aligned}
\|\Upsilon_{2i}\|_2 &\le \left\|(n^{-1}\textstyle\sum_{i=1}^n \mathbf{\Phi}_i\mathbf{\Phi}_i^\top)^{-1}\right\|_\infty \left\|n^{-1}\textstyle\sum_{i=1}^n (\partial \mathbf{\Phi}_i/\partial \boldsymbol{\theta}^\top)\right\|_\infty O(K_n^{-r}) \\
&= O_p(K_n) O_p(K_n^{-1}) O(K_n^{-r}) = O_p(K_n^{-r}),
\end{aligned} \tag{A.12}$$

and similarly we have

$$\|\Upsilon_{3i}\|_2 = O_p(K_n^{-r}). \tag{A.13}$$

Thus (A.5) is proved by (A.11), (A.12) and (A.13). By (A.4) and (A.5), we have

$$\begin{aligned} -(1/2)\,\dot{L}_n^*(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0) &= \textstyle\sum_{i=1}^n\{(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top+\Lambda_{i2}\}(\varepsilon_i+\Lambda_{i1}) \\ &= \textstyle\sum_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top\varepsilon_i+\sum_{i=1}^n\Lambda_{i2}\varepsilon_i+\sum_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top\Lambda_{i1} \\ &\quad +\textstyle\sum_{i=1}^n\Lambda_{i2}\Lambda_{i1}, \end{aligned}$$

where $\varepsilon_i = Y_i-\sum_{j=1}^p m_j(X_{ij})-g(\mathbf{X}_i^\top\boldsymbol{\alpha}^0)-\mathbf{X}_i^\top\boldsymbol{\beta}^0$. By the weak law of large numbers, (A.4) and (A.5), we have $\|\sum_{i=1}^n\Lambda_{i2}\varepsilon_i\|_2 = O_p\{(n^{1/2})\left(n^{-1/2}K_n^{3/2}+K_n^{-1}\right)\}$, $\left\|\sum_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top\Lambda_{i1}\right\|_2 = O_p\{(n^{1/2})\left(\sqrt{K_n/n}+K_n^{-r}\right)\}$, and

$$\textstyle\|\sum_{i=1}^n\Lambda_{i2}\Lambda_{i1}\|_2 = O_p\{n(\sqrt{K_n/n}+K_n^{-r})\times O_p\left(n^{-1/2}K_n^{3/2}+K_n^{-1}\right)\}.$$

Therefore, for $n^{1/(2r+2)}\ll K_n\ll n^{1/4}$, we have

$$\left\|-(1/2)\,\dot{L}_n^*(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)-\textstyle\sum_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top\varepsilon_i\right\|_2 = o_p(n^{1/2}). \qquad \square$$

Proof of Theorem 4.1. By Lemma A.1, we have

$$\left\|1/n\ddot{L}_n^*(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)-2/n\textstyle\sum_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)\right\| = o_p(1).$$

By Taylor expansion, for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ in a neighborhood of $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$, we have

$$\begin{aligned} \sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0\\ \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\end{pmatrix} &= \left\{n^{-1}\sum\nolimits_{i=1}^n(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)\right\}^{-1} \\ &\quad \left\{n^{-1/2}\sum\nolimits_{i=1}^n\varepsilon_i(\widetilde{\mathbf{Z}}_i^\top,\widetilde{\mathbf{X}}_i^\top)^\top\right\}+o_p(1), \end{aligned}$$

and thus the results in Theorem 4.1 follow. $\square$

Proof of Theorem 4.2. The results in Theorem 4.2 follow from Proposition A.1 and Theorem 4.1 directly. $\square$

A.5. Proof of Theorem 5.1

We decompose $\widehat{m}_1^{\mathrm{OR}}(x_1)$ into

$$\widehat{m}_1^{\mathrm{OR}}(x_1) = \widehat{m}_{1,m}^{\mathrm{OR}}(x_1)+\widehat{m}_{1,\varepsilon}^{\mathrm{OR}}(x_1), \tag{A.14}$$

where $\widehat{m}_{1,m}^{\mathrm{OR}}(x_1) = \widetilde{\mathbf{B}}(x_1)^\top\widehat{\boldsymbol{\delta}}_{1,m}^{\mathrm{OR}}$ and $\widehat{m}_{1,\varepsilon}^{\mathrm{OR}}(x_1) = \widetilde{\mathbf{B}}(x_1)^\top\widehat{\boldsymbol{\delta}}_{1,\varepsilon}^{\mathrm{OR}}$ with

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_{1,m}^{\mathrm{OR}} &= \{\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})m_1(X_{i1}), \\ \widehat{\boldsymbol{\delta}}_{1,\varepsilon}^{\mathrm{OR}} &= \{\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\varepsilon_i. \end{aligned}$$

Following the same reasons as Lemma A.7 in [23], by the strong approximation lemma given in Theorem 2.6.7 of [6], we can prove that

$$\sup_{x_1\in[a,b]}\left|\widehat{m}_{1,\varepsilon}^{\mathrm{OR}}(x_1)-\widehat{m}_{1,\varepsilon}^{0}(x_1)\right|=o(n^t)\ \ \text{a.s.} \tag{A.15}$$

for some $t<-r/(2r+1)<0$, where

$$\widehat{m}_{1,\varepsilon}^{0}(x_1)=\widetilde{\mathbf{B}}(x_1)^\top\{\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\sum_{i=1}^n\sigma(\mathbf{X}_i)\,\widetilde{\mathbf{B}}(X_{i1})Z_i,$$

where $Z_i, 1\le i\le n$, are *i.i.d.* $N(0,1)$ independent of X_{i1}. Define $\eta(x_1)=\sigma_n^{-1}(x_1)\widehat{m}_{1,\varepsilon}^{0}(x_1)$, where $\sigma_n^2(x_1)$ is defined in (18) and $\sigma_n(x_1)\asymp(\widetilde{L}/n)^{1/2}\{1+o_p(1)\}$ uniformly in $x_1\in S_{n,1}$. It is apparent that $\mathcal{L}\{\eta(\xi_J)\,|\mathbf{X}_i,1\le i\le n\}=N(0,1)$, so $\mathcal{L}\{\eta(\xi_J)\}=N(0,1)$ for $0\le J\le N_n$. Moreover, the eigenvalues of $\{E\widetilde{\mathbf{B}}(X_1)\widetilde{\mathbf{B}}(X_1)^\top\}^{-1}\asymp\widetilde{L}$. Then with probability approaching 1, for $J\ne J'$,

$$\begin{aligned}|E\{\eta(\xi_J)\eta(\xi_{J'})\}| &\asymp (n/\widetilde{L})n^{-1}\left|\widetilde{\mathbf{B}}(\xi_J)^\top\{E\widetilde{\mathbf{B}}(X_1)\widetilde{\mathbf{B}}(X_1)^\top\}^{-1}\widetilde{\mathbf{B}}(\xi_{J'})\right|\\ &\asymp \left|\widetilde{\mathbf{B}}(\xi_J)^\top\widetilde{\mathbf{B}}(\xi_{J'})\right|=\textstyle\sum_{\ell=1}^{\widetilde{L}}\widetilde{B}_\ell(\xi_J)\widetilde{B}_\ell(\xi_{J'}),\end{aligned}$$

and $\sum_{\ell=1}^{\widetilde{L}}\widetilde{B}_\ell(\xi_J)\widetilde{B}_\ell(\xi_{J'})\asymp C$ for a constant $0<C<\infty$ when $|\ell_J-\ell_{J'}|\le(q-1)$ and $\sum_{\ell=1}^{\widetilde{L}}\widetilde{B}_\ell(\xi_J)\widetilde{B}_\ell(\xi_{J'})=0$ when $|\ell_J-\ell_{J'}|>(q-1)$, in which ℓ_J denotes the index of the knot closest to ξ_J from the left. Therefore, by $N_n\asymp\widetilde{L}$, there exist constants $0<C_1<\infty$ and $0<C_2<\infty$ such that with probability approaching 1, for $J\ne J'$, $|E\{\eta(\xi_J)\eta(\xi_{J'})\}|\le C_1^{-|\ell_J-\ell_{J'}|}\le C_2^{-\left|J-J'\right|}$. By Lemma A1 given in [22], we have

$$\lim_{n\to\infty}P\{\sup_{0\le J\le N_n}|\eta(\xi_J)|\le\{2\log(N_n+1)\}^{1/2}d_{N_n}(\alpha)\}=1-\alpha. \tag{A.16}$$

Further, according to the result on page 149 of [8], we can show that

$$\begin{aligned}&\sup_{x_1\in S_1}\left|\{\log(N_n+1)\}^{-1/2}\sigma_n^{-1}(x_1)\{\widehat{m}_{1,m}^{\mathrm{OR}}(x_1)-m_1(x_1)\}\right|\\ =\;& O_p(\{\log(N_n+1)\}^{-1/2}(n/\widetilde{L})^{1/2}\widetilde{L}^{-r})=o_p(1).\end{aligned} \tag{A.17}$$

Therefore, by (A.14), (A.15), (A.16) and (A.17), we have

$$\begin{aligned}&\lim_{n\to\infty}P\{\sup_{0\le J\le N_n}\left|\sigma_n^{-1}(\xi_J)\{\widehat{m}_{1,m}^{\mathrm{OR}}(\xi_J)-m_1(\xi_J)\}\right|\le\{2\log(N_n+1)\}^{1/2}d_{N_n}(\alpha)\}\\ &\quad=1-\alpha,\end{aligned}$$

thus proving Theorem 5.1.

A.6. Proof of Theorem 5.2

Denote $\widehat{g}\left(\mathbf{X}_i^\top\boldsymbol{\alpha}\right)=\widehat{g}\left(\mathbf{X}_i^\top\boldsymbol{\alpha};\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0\right)$, $\widehat{\boldsymbol{\delta}}=\widehat{\boldsymbol{\delta}}(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$, $\widehat{\boldsymbol{\gamma}}=\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)$, and $U_i=\mathbf{X}_i^\top\boldsymbol{\alpha}^0$. By the definitions of $\widehat{m}_1^{\mathrm{SS}}(x_1)$ and $\widehat{m}_1^{\mathrm{OR}}(x_1)$, we have

$$\widehat{m}_1^{\mathrm{SS}}(x_1)-\widehat{m}_1^{\mathrm{OR}}(x_1)$$

$$
\begin{aligned}
&= \widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1} \sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1}) \\
&\quad \times [\{g(\mathbf{X}_i^\top \boldsymbol{\alpha}^0) - \widehat{g}\left(\mathbf{X}_i^\top \widehat{\boldsymbol{\alpha}}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}\right)\} \\
&\qquad + \sum_{j\geq 2}\{m_j(X_{ij}) - \widehat{m}_j(X_{ij})\} + \mathbf{X}_i^\top(\boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}})] \\
&= -\widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1} \\
&\quad \times \textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\{\boldsymbol{\Phi}_{i,-1}^\top (\widehat{\boldsymbol{\delta}}_{-1}^\top, \widehat{\boldsymbol{\gamma}}^\top)^\top - g(U_i) - \sum_{j\geq 2} m_j(X_{ij})\} + O(n^{-1/2}) \\
&= -\{\Psi_{n,1}(x_1) + \Psi_{n,2}(x_1)\} + O_p(n^{-1/2}), \qquad \text{(A.18)}
\end{aligned}
$$

where $\boldsymbol{\Phi}_{i,-1} = \{\boldsymbol{\Phi}_{i,-1,\ell}, 1 \leq \ell \leq L(p-1)+K\}^\top = \{\mathbf{B}_{-1}^*(\mathbf{X}_i)^\top, \mathbf{B}(U_i)^\top\}^\top$ with $\mathbf{B}_{-1}^*(\mathbf{X}_i) = \{\mathbf{b}_2(X_{i1})^\top, \ldots, \mathbf{b}_p(X_{ip})^{\mathrm{T}}\}^\top$, $\widehat{\boldsymbol{\delta}}_{-1} = (\widehat{\boldsymbol{\delta}}_2^\top, \ldots \widehat{\boldsymbol{\delta}}_p^\top)^\top$, and

$$
\begin{aligned}
\Psi_{n,1}(x_1) &= \widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1} \sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\boldsymbol{\Phi}_{i,-1}^\top (\widehat{\boldsymbol{\delta}}_{-1,e}^\top, \widehat{\boldsymbol{\gamma}}_e^\top)^\top \\
\Psi_{n,2}(x_1) &= \widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1} \\
&\qquad \times \textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\{\boldsymbol{\Phi}_{i,-1}^\top (\widehat{\boldsymbol{\delta}}_{-1,m}^\top, \widehat{\boldsymbol{\gamma}}_m^\top)^\top - g(U_i) - \sum_{j\geq 2} m_j(X_{ij})\},
\end{aligned}
$$

in which

$$
\begin{aligned}
(\widehat{\boldsymbol{\delta}}_e^\top, \widehat{\boldsymbol{\gamma}}_e^\top)^\top &= (\widehat{\boldsymbol{\delta}}_{1,e}^\top, \ldots, \widehat{\boldsymbol{\delta}}_{p,e}^\top, \widehat{\boldsymbol{\gamma}}_e^\top)^\top = (\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^\top)^{-1}\sum_{i=1}^n \boldsymbol{\Phi}_i\varepsilon_i; \\
(\widehat{\boldsymbol{\delta}}_m^\top, \widehat{\boldsymbol{\gamma}}_m^\top)^\top &= (\widehat{\boldsymbol{\delta}}_{1,m}^\top, \ldots, \widehat{\boldsymbol{\delta}}_{p,m}^\top, \widehat{\boldsymbol{\gamma}}_m^\top)^\top \\
&= (\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^\top)^{-1}\sum_{i=1}^n \boldsymbol{\Phi}_i\{g(U_i) + \sum_{j\geq 2} m_j(X_{ij})\},
\end{aligned}
$$

and $\widehat{\boldsymbol{\delta}}_{-1,e} = (\widehat{\boldsymbol{\delta}}_{2,e}^\top, \ldots, \widehat{\boldsymbol{\delta}}_{p,e}^\top)^\top$ and $\widehat{\boldsymbol{\delta}}_{-1,m} = (\widehat{\boldsymbol{\delta}}_{2,m}^\top, \ldots, \widehat{\boldsymbol{\delta}}_{p,m}^\top)^\top$. With probability approaching 1,

$$
\begin{aligned}
&\sup_{x_1\in S_1} E\{\Psi_{n,1}(x_1)\,|\mathbf{X}_i, 1\leq i\leq n\}^2 \asymp (\widetilde{L}/n)^2(K_n/n)^2 \\
&\quad \times \sup_{x_1\in S_1} \Big|\widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\boldsymbol{\Phi}_{i,-1}^\top\}(\sum_{i=1}^n \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^\top) \\
&\qquad \times \{\textstyle\sum_{i=1}^n \boldsymbol{\Phi}_{i,-1}\widetilde{\mathbf{B}}(X_{i1})^\top\}\widetilde{\mathbf{B}}(x_1)\Big| \\
&\asymp (\widetilde{L}/n)^2(K_n/n) \\
&\quad \times \sup_{x_1\in S_1} \Big|\widetilde{\mathbf{B}}(x_1)^\top \{\textstyle\sum_{i=1}^n \widetilde{\mathbf{B}}(X_{i1})\boldsymbol{\Phi}_{i,-1}^\top\}\{\sum_{i=1}^n \boldsymbol{\Phi}_{i,-1}\widetilde{\mathbf{B}}(X_{i1})^\top\}\widetilde{\mathbf{B}}(x_1)\Big|.
\end{aligned}
$$

By Bernstein's inequality [2], we can prove that $\sup_{\ell,\ell'}\left|\sum_{i=1}^n \widetilde{B}_\ell(X_{i1})\boldsymbol{\Phi}_{i,-1,\ell'}^\top\right| = O_{a.s.}(n\widetilde{L}^{-1}K_n^{-1})$. Thus, with probability approaching 1,

$$
\sup_{x_1\in S_1} E\{\Psi_{n,1}(x_1)\,|\mathbf{X}_i, 1\leq i\leq n\}^2
$$

$$\asymp (\widetilde{L}/n)^2(K_n/n)(n^2\widetilde{L}^{-2}K_n^{-2})K_n \sup_{x_1\in S_1}\left|\widetilde{\mathbf{B}}(x_1)^\top\widetilde{\mathbf{B}}(x_1)\right| = O\left(n^{-1}\right).$$

Therefore, by Bernstein's inequality in [2], we have

$$\sup_{x_1\in S_1}|\Psi_{n,1}(x_1)| = O_p(n^{-1/2}\sqrt{\log n}). \tag{A.19}$$

Moreover, by (A.6),

$$\begin{aligned}
&\sup_{x_1\in S_1}|\Psi_{n,2}(x_1)|\\
&\quad\le \sup_{x_1\in S_1}\Big|\widetilde{\mathbf{B}}(x_1)^\top\{\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\\
&\qquad\times\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\{\mathbf{\Phi}_{i,-1}^\top(\boldsymbol{\delta}_{-1}^\top,\boldsymbol{\gamma}^\top) - g(U_i) - \sum_{j\ge 2}m_j(X_{ij})\}\Big| + O(K_n^{-r})\\
&\quad\le \sup_{x_1\in S_1}\left|\widetilde{\mathbf{B}}(x_1)^\top\{\textstyle\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\widetilde{\mathbf{B}}(X_{i1})^\top\}^{-1}\sum_{i=1}^n\widetilde{\mathbf{B}}(X_{i1})\right| O(K_n^{-r})\\
&\qquad + O(K_n^{-r})\\
&\quad= O_p(K_n/n)O_p(n/K_n)O(K_n^{-r}) + O(K_n^{-r}) = O_p(K_n^{-r}).
\end{aligned}\tag{A.20}$$

Therefore, Theorem 5.2 follows from (A.18), (A.19) and (A.20).

Supplementary Material

Additional Results for Simulation Studies
(doi: 10.1214/17-EJS1291SUPP; .pdf).

References

[1] Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press, Princeton, N.J. MR0803258

[2] Bosq, D. (1998), *Nonparametric Statistic for Stochastic Process*, Vol. 10 of *Lecture notes in Statistics*, 2 edn, Springer, New York. MR0994249

[3] Breiman, L. and Friedman, J. H. (1985), 'Estimating optimal transformations for multiple regression and correlations (with discussion)', *Journal of the American Statistical Association* **80**, 580–619. MR1467842

[4] Buja, A., Hastie, T. and Tibshirani, R. (1989), 'Linear smoothers and additive models', *The Annals of Statistics* **17**, 453–510. MR0666546

[5] 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**(438), 477–489. MR2850216

[6] Csörgö, M. and Révész, P. (1981), *Strong Approximations in Probability and Statistics*, Academic Press, New York-London. MR1900298

[7] Cui, X., Härdle, W. and Zhu, L.-X. (2011), 'The EFM approach for single-index models', *The Annals of Statistics* **39**, 1658–1688. MR0862236

[8] de Boor, C. (2001), *A Practical Guide to Splines*, Vol. 27 of *Applied Mathematical Sciences*, revised edn, Springer-Verlag, New York. MR1261635
[9] Demko, S. (1986), 'Spectral bounds for $\left|a^{-1}\right|_{\infty}$', *Journal of Approximation Theory* **48**, 207–212. MR1212171
[10] DeVore, R. A. and Lorentz, G. G. (1993), *Constructive approximation*, Vol. 303 of *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, Berlin. MR2061786
[11] Goldenshluger, A. and Lepski, O. (2009), 'Structural adaptation via L_p-norm oracle inequalities', *Probability Theory and Related Fields* **1**43(1-2), 41–71. MR2449122
[12] Härdle, W., Hall, P. and Ichimura, H. (1993), 'Optimal smoothing in single-index models', *The Annals of Statistics* **21**, 157–178. MR1082147
[13] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Springer-Verlag, New York. MR2535631
[14] Hastie, T. and Tibshirani, R. (1990) , *Generalized Additive Models*, Monographs on statistics and applied probability, 1st edn, Chapman and Hall, London; New York. MR1230981
[15] Horowitz, J. L. (2009), *Semiparametric and Nonparametric Methods in Econometrics*, Springer, New York. MR2443182
[16] Ichimura, H. (1993), 'Semiparametric least squares (SLS) and weighted SLS estimation of single-index models', *Journal of Econometrics* **58**, 71–120. MR2380574
[17] Liang, H., Thurston, S., Ruppert, D., Apanasovich, T. and Hauser, R. (2008), 'Additive partial linear models with measurement errors', *Biometrika* **95**, 667–678. MR2827522
[18] Lin, W. and Kulasekera, K. B. (2007), 'Identifiability of single-index models and additive-index models', *Biometrika* **94**(2), 496–501. MR3097965
[19] Liu, X., Wang, L. and Liang, H. (2011), 'Estimation and variable selection for semiparametric additive partial linear models', *Statistica Sinica* **21**, 1225–1248. MR2780816
[20] Lu, X. and Cheng, T. (2007), 'Randomly censored partially linear single-index models', *Journal of Multivariate Analysis* **9**8, 1895–1922. MR2396946
[21] Ma, S. (2012), 'Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes', *The Annals of Statistics* **40**, 2943–2972. MR2933169
[22] Ma, S. and Yang, L. (2011), 'A jump-detecting procedure based on polynomial spline estimation', *Journal of Nonparametric Statistics* **23**, 67–81.
[23] Ma, S., Yang, L. and Carroll, R. (2012), 'Simultaneous confidence band for sparse longitudinal regression', *Statistica Sinica* **22**, 95–122.
[24] Ma, S., Lian, H., Liang H. and Carroll, R. J. (2017). Supplemental Material: Additional Results for Simulation Studies. DOI: 10.1214/17-EJS1291SUPP.
[25] Opsomer, J. and Ruppert, D. (1999), 'A root-n consistent backfitting estimator for semiparametric additive modeling', *Journal of Computational and Graphical Statistics* **8**, 715–732. MR2671198
[26] Shen, X., Wolfe, D. A. and Zhou, S. (1998), 'Local asymptotics for regres-

sion splines and confidence regions', *Annals of Statistics* **26**, 1760–1782. MR0840516

[27] Song, Q. and Yang, L. (2010), 'Oracally efficient spline smoothing of nonlinear additive autoregression model with simultaneous confidence band', *Journal of Multivariate Analysis* **101**, 2008–2025. MR0790566

[28] Stone, C. (1986), 'The dimensionality reduction principle for generalized additive models', *The Annals of Statistics* **14**, 590–606. MR1272079

[29] Stone, C. J. (1985), 'Additive regression and other nonparametric models', *The Annals of Statistics* **13**, 689–705. MR2589322

[30] Stone, C. J. (1994), 'The use of polynomial splines and their tensor products in multivariate function estimation', *The Annals of Statistics* **22**, 118–184. MR2893854

[31] Wang, J.-L., Xue, L., Zhu, L. and Chong, Y. S. (2010), 'Estimation for a partial-linear single-index model', *The Annals of Statistics* **38**, 246–274. MR3210980

[32] Wang, L., Liu, X., Liang, H. and Carroll, R. (2011), 'Estimation and variable selection for generalized additive partial linear models', *The Annals of Statistics* **39**, 1827–1851. MR2382655

[33] Wang, L., Xue, L., Qu, A. and Liang, H. (2014), 'Estimation and model selection in generalized additive partial linear models for high-dimensional correlated data', *The Annals of Statistics* **42**, 592–624. MR2206355

[34] Wang, L. and Yang, L. (2007), 'Spline-backfitted kernel smoothing of nonlinear additive autoregression model', *The Annals of Statistics* **35**, 2474–2503. MR2276153

[35] Wood, S. N. (2006), *Generalized Additive Models*, Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton, FL. MR1731489

[36] Xia, Y. C. and Härdle, W. (2006), 'Semi-parametric estimation of partially linear single-index models', *Journal of Multivariate Analysis* **97**, 1162–1184. MR1741980

[37] Xia, Y. and Li, W. K. (1999), 'On single-index coefficient regression models', *Journal of the American Statistical Association* **94**, 1275–1285. MR2796568

[38] Xia, Y., Tong, H. and Li, W. K. (1999), 'On extended partially linear single-index models', *Biometrika* **86**, 831–842. MR2327498

[39] Xue, L., Qu, A. and Zhou, J. (2010), 'Consistent model selection for marginal generalized additive model for correlated data', *Journal of the American Statistical Association* **105**(492), 1518–1530. MR1673277

[40] Xue, L. and Yang, L. (2006), 'Additive coefficient modeling via polynomial spline', *Statistica Sinica* **16**, 1423–1446. MR1742102

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

[42] Zhou, S. and Wolfe, D. A. (2000), 'On derivative estimation in spline regression', *Statistica Sinica* **10**, 93–105.