Journal of Multivariate Analysis 173 (2019) 204–228

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sparse model identification and learning for ultra-high-dimensional additive partially linear models

Xinyi Li [a], Li Wang [b,*], Dan Nettleton [b]

[a] SAMSI / Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, NC 27709, USA
[b] Department of Statistics, Iowa State University, Ames, IA 50011, USA

ARTICLE INFO

Article history:
Received 16 April 2018
Received in revised form 17 February 2019
Accepted 18 February 2019
Available online 1 March 2019

AMS 2010 subject classifications:
primary 62H12
secondary 62F12

Keywords:
Dimension reduction
Inference for ultra-high-dimensional data
Semiparametric regression
Spline-backfitted local polynomial
Structure identification
Variable selection

ABSTRACT

The additive partially linear model (APLM) combines the flexibility of nonparametric regression with the parsimony of regression models, and has been widely used as a popular tool in multivariate nonparametric regression to alleviate the "curse of dimensionality". A natural question raised in practice is the choice of structure in the nonparametric part, i.e., whether the continuous covariates enter into the model in linear or nonparametric form. In this paper, we present a comprehensive framework for simultaneous sparse model identification and learning for ultra-high-dimensional APLMs where both the linear and nonparametric components are possibly larger than the sample size. We propose a fast and efficient two-stage procedure. In the first stage, we decompose the nonparametric functions into a linear part and a nonlinear part. The nonlinear functions are approximated by constant spline bases, and a triple penalization procedure is proposed to select nonzero components using adaptive group LASSO. In the second stage, we refit data with selected covariates using higher order polynomial splines, and apply spline-backfitted local-linear smoothing to obtain asymptotic normality for the estimators. The procedure is shown to be consistent for model structure identification. It can identify zero, linear, and nonlinear components correctly and efficiently. Inference can be made on both linear coefficients and nonparametric functions. We conduct simulation studies to evaluate the performance of the method and apply the proposed method to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes for illustration.

1. Introduction

In the past three decades, flexible and parsimonious additive partially linear models (APLMs) have been extensively studied and widely used in many statistical applications, including biology, econometrics, engineering, and social science. Examples of recent work on APLMs include [19,20,22–24,28,29]. APLMs are natural extensions of classical parametric models with good interpretability and are becoming more and more popular in data analysis.

Suppose we observe $(Y_1, \mathbf{Z}_{(1)}, \mathbf{X}_{(1)}), \ldots, (Y_n, \mathbf{Z}_{(n)}, \mathbf{X}_{(n)})$. For subject $i \in \{1, \ldots, n\}$, Y_i is a univariate response, $\mathbf{Z}_{(i)} = (Z_{i1}, \ldots, Z_{ip_1})^\top$ is a p_1-dimensional vector of covariates that may be linearly associated with the response, and $\mathbf{X}_{(i)} = (X_{i1}, \ldots, X_{ip_2})^\top$ is a p_2-dimensional vector of continuous covariates that may have nonlinear associations with the

* Corresponding author.
E-mail address: lilywang@iastate.edu (L. Wang).

https://doi.org/10.1016/j.jmva.2019.02.010
0047-259X/

response. We assume that $(Y_1, \mathbf{Z}_{(1)}, \mathbf{X}_{(1)}), \ldots, (Y_n, \mathbf{Z}_{(n)}, \mathbf{X}_{(n)})$ form a random sample from the distribution of $(Y, \boldsymbol{Z}, \boldsymbol{X})$, satisfying the model

$$Y_i = \mu + \mathbf{Z}_{(i)}^{\top}\boldsymbol{\alpha} + \sum_{\ell=1}^{p_2} \phi_\ell(X_{i\ell}) + \varepsilon_i = \mu + \sum_{k=1}^{p_1} Z_{ik}\alpha_k + \sum_{\ell=1}^{p_2} \phi_\ell(X_{i\ell}) + \varepsilon_i, \tag{1}$$

where μ is the intercept, $\alpha_1, \ldots, \alpha_{p_1}$ are unknown regression coefficients, $\phi_1, \ldots, \phi_{p_2}$ are unknown smooth functions, and each ϕ_ℓ is centered with $\mathrm{E}\phi_\ell(X_{i\ell}) = 0$ to make model (1) identifiable. The $\mathbf{X}_{(i)}$ is a p_2-dimensional vector of zero-mean covariates having density with a compact support. Without loss of generality, we assume that each covariate $X_{i1}, \ldots, X_{p_2}$ can be rescaled into an interval $\chi = [a, b]$. The ε_i terms are iid random errors with mean zero and variance σ^2.

The APLM is particularly convenient when $\boldsymbol{Z}$ is a vector of categorical or discrete variables, and in this case, the components of $\boldsymbol{Z}$ enter the linear part of model (1) automatically, and the continuous variables usually enter the model nonparametrically. In practice, we might have reasons to believe that some of the continuous variables should enter the model linearly rather than nonparametrically. A natural question is how to determine which continuous covariates have a linear effect and which continuous covariates have a nonlinear effect. If the choice of linear components is correctly specified, then the biases in the estimation of these components are eliminated and root-n convergence rates can be obtained for the linear coefficients. However, such prior knowledge is rarely available, especially when the number of covariates is large. Thus, structure identification, or linear and nonlinear detection, is an important step in the process of building an APLM from high-dimensional data.

When the number of covariates in the model is fixed, structure identification in additive models (AMs) has been studied in the literature. Zhang et al. [33] proposed a penalization procedure to identify the linear components in AMs in the context of smoothing splines ANOVA. They demonstrated the consistency of the model structure identification and established the convergence rate of the proposed method specifically under the tensor product design. Huang et al. [13] proposed another penalized semiparametric regression approach using a group minimax concave penalty to identify the covariates with linear effects. They showed consistency in determining the linear and nonlinear structure in covariates, and obtained the convergence rate of nonlinear function estimators and asymptotic properties of linear coefficient estimators; but they did not perform variable selection at the same time.

For high-dimensional AMs, Lian et al. [18] proposed a double penalization procedure to distinguish covariates that enter the nonparametric and parametric parts and to identify significant covariates simultaneously. They demonstrated the consistency of the model structure identification, and established the convergence rate of nonlinear function estimators and asymptotic normality of linear coefficient estimators. Despite the nice theoretical properties, their method heavily relies on the local quadratic approximation in [9], which is incapable of producing naturally sparse estimates. In addition, employing the local quadratic approximation can be extremely expensive because it requires the repeated factorization of large matrices, which becomes infeasible when the number of covariates is very large.

Note that all the aforementioned papers [13,18,33] about structure identification focus on the AM with continuous explanatory variables. However, in many applications, a canonical partitioning of the variables exists. In particular, if there are categorical or discrete explanatory variables, as in the case of the SAM data studies (see the details in Section 5) and in many genome-wide association studies, we may want to keep discrete explanatory variables separate from the other design variables and let discrete variables enter the linear part of the model directly. In addition, if there is some prior knowledge of certain parametric forms for some specific covariates, such as a linear form, we may lose efficiency if we simply model all the covariates nonparametrically.

The above practical and theoretical concerns motivate our further investigation of the simultaneous variable selection and structure selection problem for flexible and parsimonious APLMs, in which the features of the data suitable for parametric modeling are modeled parametrically and nonparametric components are used only where needed. We consider the setting where both the dimension of the linear components and the dimension of nonlinear components is ultra-high. We propose an efficient and stable penalization procedure for simultaneously identifying linear and nonlinear components, removing insignificant predictors, and estimating the remaining linear and nonlinear components. We prove the proposed *S*parse *M*odel *I*dentification, *L*earning and *E*stimation (referred to as SMILE) procedure is consistent. We propose an iterative group coordinate descent approach to solve the penalized minimization problem efficiently. Our algorithm is very easy to implement because it only involves simple arithmetic operations with no complicated numerical optimization steps, matrix factorizations, or inversions. In one simulation example with $n = 500$ and $p_1 = p_2 = 5000$, it takes less than one minute to complete the entire model identification and variable selection process on a regular PC.

After variable selection and structure detection, we would like to provide an inferential tool for the linear and nonparametric components. The spline method is fast and easy to implement; however, the rate of convergence is only established in mean squares sense, and there is no asymptotic distribution or uniform convergence, so no measures of confidence can be assigned to the estimators. In this paper, we propose a two-step spline-backfitted local-linear smoothing (SBLL) procedure for APLM estimation, model selection and simultaneous inference for all the components. In the first stage, we approximate the nonparametric functions $\phi_1, \ldots, \phi_{p_2}$ with undersmoothed constant spline functions. We perform model selection for the APLM using a triple penalized procedure to select important variables and identify the linear vs. nonlinear structure for the continuous covariates, which is crucial to obtain efficient estimators for the non-zero components. We show that the proposed model selection and structure identification for both parametric and

nonparametric terms are consistent, and the estimators of the nonzero linear coefficients and nonzero nonparametric functions are both L_2 norm consistent. In the second stage, we refit the data with covariates selected in the first step using higher-order polynomial splines to achieve root-n consistency of the coefficient estimators in the linear part, and apply a one-step local-linear backfitting to the projected nonparametric components obtained from the refitting. Asymptotic normality for both linear coefficient estimators and nonlinear component estimators, as well as simultaneous confidence bands (SCBs) for all nonparametric components, are provided.

The rest of the paper is organized as follows. In Section 2, we describe the first-stage spline smoothing and propose a triple penalized regularization method for simultaneous model identification and variable selection. The theoretical properties of selection consistency and rates of convergence for the coefficient estimators and nonparametric estimators are developed. Section 3 introduces the spline-backfitted local-linear estimators and SCBs for the nonparametric components. The performance of the estimators is assessed by simulations in Section 4 and illustrated by application to the SAM data in Section 5. Some concluding remarks are given in Section 6. Appendix A evaluates the effect of different smoothing parameters on the performance of the proposed method. Technical details are provided in Appendix B.

2. Methodology

2.1. Model setup

In the following, the functional form (linear vs. nonlinear) for each continuous covariate in model (1) is assumed to be unknown. In order to decide the form of ϕ_ℓ, for each $\ell \in \{1, \ldots, p_2\}$, we can decompose ϕ_ℓ into a linear part and a nonlinear part: $\phi_\ell(x) = \beta_\ell x + g_\ell(x)$, where g_ℓ is some unknown smooth nonlinear function; see Assumption (A1) in Section 2.3. For model identifiability, we assume that $\mathrm{E}(X_{i\ell}) = 0$, $\mathrm{E}\{g_\ell(X_{i\ell})\} = 0$ and $\mathrm{E}\{g'_\ell(X_{i\ell})\} = 0$. The first two constraints $\mathrm{E}(X_{i\ell}) = 0$ and $\mathrm{E}\{g_\ell(X_{i\ell})\} = 0$, are required to guarantee identifiability for the APLM, i.e., $\mathrm{E}\{\phi_\ell(X_{i\ell})\} = 0$. The constraint $\mathrm{E}\{g'_\ell(X_{i\ell})\} = 0$ ensures there is no linear form in nonlinear function g_ℓ. Note that these constraints are also in accordance with the definition of nonlinear contrast space in [33], which is a subspace of the orthogonal decomposition of RKHS. In the following, we assume Y_i values are centered so that we can express the APLM in (1) without an intercept parameter as

$$Y_i = \sum_{k=1}^{p_1} Z_{ik}\alpha_k + \sum_{\ell=1}^{p_2} X_{i\ell}\beta_\ell + \sum_{\ell=1}^{p_2} g_\ell(X_{i\ell}) + \varepsilon_i. \tag{2}$$

In the following, we define predictor variable Z_k as irrelevant in model (2), if and only if $\alpha_k = 0$, and X_ℓ as irrelevant if and only if $\beta_\ell = 0$ and $g_\ell(x_\ell) = 0$ for all x_ℓ on its support. A predictor variable is defined as relevant if and only if it is not irrelevant. Suppose that only an unknown subset of predictor variables is relevant. We are interested in identifying such subsets of relevant predictors consistently while simultaneously estimating their coefficients and/or functions.

For covariates $\boldsymbol{Z}$, we define

$$\text{Active index set for } \boldsymbol{Z} : \mathcal{S}_z = \{k = 1, \ldots, p_1 : \alpha_k \neq 0\},$$
$$\text{Inactive index set for } \boldsymbol{Z} : \mathcal{N}_z = \{k = 1, \ldots, p_1 : \alpha_k = 0\}.$$

For continuous covariate X_ℓ, we say it is a linear covariate if $\beta_\ell \neq 0$ and $g_\ell(x_\ell) = 0$ for all x_ℓ on its support, and X_ℓ is a nonlinear covariate if $g_\ell(x_\ell) \neq 0$. Explicitly, we define the following index sets for $\boldsymbol{X}$:

$$\text{Active pure linear index set for } \boldsymbol{X} : \mathcal{S}_{x,PL} = \{\ell = 1, \ldots, p_2 : \beta_\ell \neq 0,\ g_\ell \equiv 0\},$$
$$\text{Active nonlinear index set for } \boldsymbol{X} : \mathcal{S}_{x,N} = \{\ell = 1, \ldots, p_2 : g_\ell \neq 0\},$$
$$\text{Inactive index set for } \boldsymbol{X} : \mathcal{N}_x = \{\ell = 1, \ldots, p_2 : \beta_\ell = 0,\ g_\ell \equiv 0\}.$$

Note that the active nonlinear index set for $\boldsymbol{X}$, $\mathcal{S}_{x,N}$, can be decomposed as $\mathcal{S}_{x,N} = \mathcal{S}_{x,LN} \cup \mathcal{S}_{x,PN}$, where $\mathcal{S}_{x,LN} = \{\ell = 1, \ldots, p_2 : \beta_\ell \neq 0,\ g_\ell \neq 0\}$ is the index set for covariates whose linear and nonlinear terms in (2) are both nonzero, and $\mathcal{S}_{x,PN} = \{\ell = 1, \ldots, p_2 : \beta_\ell = 0,\ g_\ell \neq 0\}$ is the index set for active pure nonlinear index set for $\boldsymbol{X}$.

Therefore, the model selection problem for model (2) is equivalent to the problem of identifying $\mathcal{S}_z$, $\mathcal{N}_z$, $\mathcal{S}_{x,PL}$, $\mathcal{S}_{x,LN}$, $\mathcal{S}_{x,PN}$ and $\mathcal{N}_x$. To achieve this, we propose to minimize

$$\sum_{i=1}^{n}\left\{Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{\ell=1}^{p_2} X_{i\ell}\beta_\ell - \sum_{\ell=1}^{p_2} g_\ell(X_{i\ell})\right\}^2 + \sum_{k=1}^{p_1} p_{\lambda_{n1}}(|\alpha_k|) + \sum_{\ell=1}^{p_2} p_{\lambda_{n2}}(|\beta_\ell|) + \sum_{\ell=1}^{p_2} p_{\lambda_{n3}}(\|g_\ell\|_2), \tag{3}$$

where $\|g_\ell\|_2^2 = \mathrm{E}\{g_\ell^2(X_\ell)\}$, and $p_{\lambda_{n1}}$, $p_{\lambda_{n2}}$ and $p_{\lambda_{n3}}$ are penalty functions explained in detail in Section 2.3. The tuning parameters λ_{n1}, λ_{n2} and λ_{n3} decide the complexity of the selected model. The smoothness of predicted nonlinear functions is controlled by λ_{n3}, and λ_{n1}, λ_{n2} and λ_{n3} go to infinity as $n \to \infty$.

2.2. Spline basis approximation

We approximate the smooth functions $\{g_\ell : \ell = 1, \ldots, p_2\}$ in (2) by polynomial splines for their simplicity in computation. For example, for each $\ell \in \{1, \ldots, p_2\}$, let $\upsilon_{0,\ell}, \ldots, \upsilon_{N_n+1,\ell}$ be knots that partition $[a, b]$ with $a = \upsilon_{0,\ell} < \upsilon_{1,\ell} < \cdots < \upsilon_{N_n,\ell} < \upsilon_{N_n+1,\ell} = b$. The space of polynomial splines of order $d \geq 1$, $\mathbb{B}_\ell^{(d)}[a, b]$, consisting of functions s satisfying (i) the restriction of s to subintervals $[\upsilon_{J,\ell}, \upsilon_{J+1,\ell}), J = 1, \ldots, N_n + d$, and $\left[\upsilon_{N_n,\ell}, \upsilon_{N_n+1,\ell}\right]$, is a polynomial of $(d-1)$-degree (or less); (ii) for $d \geq 2$ and $0 \leq d' \leq d-2$, s is d' times continuously differentiable on $[a, b]$. Below we denote $b_{1,\ell}^{(d)}, \ldots, b_{N_n+d,\ell}^{(d)}$ the basis functions of $\mathbb{B}_\ell^{(d)}[a, b]$.

To ensure $\mathrm{E}\{g_\ell(X_{i\ell})\} = 0$ and $\mathrm{E}\{g_\ell'(X_{i\ell})\} = 0$, we consider the following normalized first-order B-splines, referred to as piecewise constant splines. We define for any $\ell \in \{1, \ldots, p_2\}$ the piecewise constant B-spline function as the indicator function $I_{J,\ell}(x_\ell)$ of the $(N_n + 1)$ equally-spaced subintervals of $[a, b]$ with length $H = H_n = (b - a)/(N_n + 1)$, i.e., for all $J \in \{0, \ldots, N_n - 1\}$,

$$I_{J,\ell}(x_\ell) = \begin{cases} 1 & \text{if } a + JH \leq x_\ell < a + (J+1)H, \\ 0 & \text{otherwise} \end{cases}, \quad I_{N_n,\ell}(x_\ell) = \begin{cases} 1 & \text{if } a + N_n H \leq x_\ell \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Define the following centered spline basis. For all $J \in \{1, \ldots, N_n\}$ and $\ell \in \{1, \ldots, p_2\}$, set

$$b_{J,\ell}^{(1)}(x_\ell) = I_{J,\ell}(x_\ell) - (\|I_{J,\ell}\|_2 / \|I_{J-1,\ell}\|_2) I_{J-1,\ell}(x_\ell),$$

with the standardized version given, for any $\ell \in \{1, \ldots, p_2\}$, and $J \in \{1, \ldots, N_n\}$, by

$$B_{J,\ell}^{(1)}(x_\ell) = b_{J,\ell}^{(1)}(x_\ell) / \|b_{J,\ell}^{(1)}\|_2. \tag{4}$$

So $\mathrm{E}\{B_{J,\ell}^{(1)}(X_{i\ell})\} = 0$, $\mathrm{E}\{B_{J,\ell}^{(1)}(X_{i\ell})\}^2 = 1$. In practice, we use the empirical distribution of $X_{1\ell}, \ldots, X_{n\ell}$ to perform the centering and scaling in the definitions of $b_{J,\ell}^{(1)}(x_\ell)$ and $B_{J,\ell}^{(1)}(x_\ell)$.

For each $\ell \in \{1, \ldots, p_2\}$, we approximate the nonparametric function $g_\ell(x_\ell)$ using the above normalized piecewise constant splines

$$g_\ell(x_\ell) \approx g_{\ell s}(x_\ell) = \sum_{J=1}^{N_n} \gamma_{J,\ell} B_{J,\ell}^{(1)}(x_\ell) = \mathbf{B}_\ell^{(1)\top}(x_\ell) \boldsymbol{\gamma}_\ell, \tag{5}$$

where $\mathbf{B}_\ell^{(1)}(x_\ell) = (B_{1,\ell}^{(1)}(x_\ell), \ldots, B_{N_n,\ell}^{(1)}(x_\ell))^\top$, and $\boldsymbol{\gamma}_\ell = \left(\gamma_{1,\ell}, \ldots, \gamma_{N_n,\ell}\right)^\top$ is a vector of the spline coefficients. By using the centered constant spline basis functions, we can guarantee that $\sum_{i=1}^n g_{\ell s}(X_{i\ell}) = 0$, and $\sum_{i=1}^n g_{\ell s}'(X_{i\ell}) = 0$ except at the location of the knots.

Denote a length N_n vector $\mathbf{B}_{i\ell}^{(1)} = (B_{1,\ell}^{(1)}(X_{i\ell}), \ldots, B_{N_n,\ell}^{(1)}(X_{i\ell}))^\top$. For any vector $\boldsymbol{a} \in \mathbb{R}^p$, denote $\|\boldsymbol{a}\| = (\sum_{\ell=1}^p a_\ell^2)^{1/2}$ as the L_2 norm of $\boldsymbol{a}$. Following from (5), to minimize (3), it is approximately equivalent to consider the problem of minimizing

$$\sum_{i=1}^n \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{\ell=1}^{p_2} X_{i\ell}\beta_\ell - \sum_{\ell=1}^{p_2} \mathbf{B}_{i\ell}^{(1)} \boldsymbol{\gamma}_\ell \right\}^2 + \sum_{k=1}^{p_1} p_{\lambda_{n1}}(|\alpha_k|) + \sum_{\ell=1}^{p_2} p_{\lambda_{n2}}(|\beta_\ell|) + \sum_{\ell=1}^{p_2} p_{\lambda_{n3}}(\|\boldsymbol{\gamma}_\ell\|).$$

2.3. Adaptive group LASSO regularization

We use adaptive LASSO [35] and adaptive group LASSO [12] for variable selection and estimation. Other popular choices include methods based on the Smoothly Clipped Absolute Deviation penalty [9] or the minimax concave penalty [32]. Specifically, we start with group LASSO estimators obtained from the following minimization:

$$(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^n \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{\ell=1}^{p_2} X_{i\ell}\beta_\ell - \sum_{\ell=1}^{p_2} \mathbf{B}_{i\ell}^{(1)} \boldsymbol{\gamma}_\ell \right\}^2 + \widetilde{\lambda}_{n1} \sum_{k=1}^{p_1} |\alpha_k| + \widetilde{\lambda}_{n2} \sum_{\ell=1}^{p_2} |\beta_\ell| + \widetilde{\lambda}_{n3} \sum_{\ell=1}^{p_2} \|\boldsymbol{\gamma}_\ell\|. \tag{6}$$

Then, let $w_k^\alpha = |\widetilde{\alpha}_k|^{-1}\mathbf{1}(|\widetilde{\alpha}_k| > 0) + \infty \times \mathbf{1}(|\widetilde{\alpha}_k| = 0)$, $w_\ell^\beta = |\widetilde{\beta}_\ell|^{-1}\mathbf{1}(|\widetilde{\beta}_\ell| > 0) + \infty \times \mathbf{1}(|\widetilde{\beta}_\ell| = 0)$, $w_\ell^\gamma = \|\widetilde{\boldsymbol{\gamma}}_\ell\|^{-1}\mathbf{1}(\|\widetilde{\boldsymbol{\gamma}}_\ell\| > 0) + \infty \times \mathbf{1}(\|\widetilde{\boldsymbol{\gamma}}_\ell\| = 0)$, where by convention, $\infty \times 0 = 0$. The adaptive group LASSO objective function is defined as

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}; \lambda_{n1}, \lambda_{n2}, \lambda_{n3}) = & \sum_{i=1}^n \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{\ell=1}^{p_2} X_{i\ell}\beta_\ell - \sum_{\ell=1}^{p_2} \mathbf{B}_{i\ell}^{(1)} \boldsymbol{\gamma}_\ell \right\}^2 \\ & + \lambda_{n1} \sum_{k=1}^{p_1} w_k^\alpha |\alpha_k| + \lambda_{n2} \sum_{\ell=1}^{p_2} w_\ell^\beta |\beta_\ell| + \lambda_{n3} \sum_{\ell=1}^{p_2} w_\ell^\gamma \|\boldsymbol{\gamma}_\ell\|. \end{aligned} \tag{7}$$

The adaptive group LASSO estimators are minimizers of (7), denoted by

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}; \lambda_{n1}, \lambda_{n2}, \lambda_{n3}).$$

The model structure selected is defined by

$$\widehat{\mathcal{S}}_z = \{1 \le k \le p_1 : |\widehat{\alpha}_k| > 0\}, \quad \widehat{\mathcal{S}}_{x,PL} = \{\ell : |\widehat{\beta}_\ell| > 0, \|\widehat{\boldsymbol{\gamma}}_\ell\| = 0, 1 \le \ell \le p_2\},$$
$$\widehat{\mathcal{S}}_{x,LN} = \{\ell : |\widehat{\beta}_\ell| > 0, \|\widehat{\boldsymbol{\gamma}}_\ell\| > 0, 1 \le \ell \le p_2\}, \quad \widehat{\mathcal{S}}_{x,PN} = \{\ell : |\widehat{\beta}_\ell| = 0, \|\widehat{\boldsymbol{\gamma}}_\ell\| > 0, 1 \le \ell \le p_2\}.$$

The spline estimators of each component function are

$$\widehat{g}_\ell(x_\ell) = \sum_{J=1}^{N_n} \widehat{\gamma}_{J,\ell} B_{J,\ell}^{(1)}(x_\ell) - \frac{1}{n}\sum_{i=1}^{n}\sum_{J=1}^{N_n} \widehat{\gamma}_{J,\ell} B_{J,\ell}^{(1)}(X_{i\ell}).$$

Accordingly, the spline estimators for the original component functions ϕ_ℓ are $\widehat{\phi}_\ell(x_\ell) = \widehat{\beta}_\ell x_\ell + \widehat{g}_\ell(x_\ell)$.

The following theorems establish the asymptotic properties of the adaptive group LASSO estimators. Theorem 1 shows the proposed method can consistently distinguish nonzero components from zero components. Theorem 2 gives the convergence rates of the estimators. We only state the main results here. To facilitate the development of the asymptotic properties, we assume the following sparsity condition:

(A1) (*Sparsity*) The numbers of nonzero components $|\mathcal{S}_z|$, $|\mathcal{S}_{x,PL}|$ and $|\mathcal{S}_{x,N}|$ are fixed, and there exist positive constants c_α, c_β and c_g such that $\min_{k\in\mathcal{S}_z}|\alpha_{0k}| \ge c_\alpha$, $\min_{\ell\in\mathcal{S}_{x,PL}}|\beta_{0\ell}| \ge c_\beta$, and $\min_{\ell\in\mathcal{S}_{x,N}}\|g_{0\ell}\|_2 \ge c_g$.

Other regularity conditions and proofs are provided in Appendix B.1–B.2.

Theorem 1. *Suppose that Assumptions (A1), (A2)–(A6) in Appendix B.1 hold. As $n \to \infty$, we have $\widehat{\mathcal{S}}_z = \mathcal{S}_z$, $\widehat{\mathcal{S}}_{x,PL} = \mathcal{S}_{x,PL}$, $\widehat{\mathcal{S}}_{x,LN} = \mathcal{S}_{x,LN}$ and $\widehat{\mathcal{S}}_{x,PN} = \mathcal{S}_{x,PN}$ with probability approaching* 1.

In the following, to avoid confusion, we use $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0p_1})^\top$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_2})^\top$ to denote the true parameters in model (2), and $\boldsymbol{g}_0 = (g_{01}, \ldots, g_{0p_2})^\top$ to denote the nonlinear functions in model (2). Let $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{0,\mathcal{S}_z}^\top, \boldsymbol{\alpha}_{0,\mathcal{N}_z}^\top)^\top$, where $\boldsymbol{\alpha}_{0,\mathcal{S}_z}$ consists of all nonzero components of $\boldsymbol{\alpha}_0$, and $\boldsymbol{\alpha}_{0,\mathcal{N}_z} = \mathbf{0}$ without loss of generality; similarly, let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,\mathcal{S}_{x,L}}^\top, \boldsymbol{\beta}_{0,\mathcal{N}_x}^\top)^\top$, where $\boldsymbol{\beta}_{0,\mathcal{S}_{x,L}}$ consists of all nonzero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{0,\mathcal{N}_x} = \mathbf{0}$ without loss of generality.

Theorem 2. *Suppose that Assumptions (A1), (A2)–(A6) in Appendix B.1 hold. Then*

$$\sum_{k\in\mathcal{S}_z}|\widehat{\alpha}_k - \alpha_{0k}|^2 + \sum_{\ell\in\mathcal{S}_{x,L}}|\widehat{\beta}_\ell - \beta_{0\ell}|^2 = O_P(n^{-1}N_n) + O(N_n^{-2}) + O_P\left(n^{-2}\sum_{j=1}^{3}\lambda_{nj}^2\right),$$

$$\sum_{\ell\in\mathcal{S}_{x,N}}\|\widehat{g}_\ell - g_{0\ell}\|_2^2 = O_P(n^{-1}N_n) + O(N_n^{-2}) + O_P\left(n^{-2}\sum_{j=1}^{3}\lambda_{nj}^2\right).$$

3. Two-stage SBLL estimator and inference

After model selection, our next step is to conduct statistical inference for the nonparametric component functions of those important variables. Although the one-step penalized estimation in Section 2.3 can quickly identify the nonzero nonlinear components, the asymptotic distribution is not available for the resulting estimators.

To obtain estimators whose asymptotic distribution can be used for inference, we first refit the data using the selected model, viz.

$$Y_i = \sum_{k\in\widehat{\mathcal{S}}_z} Z_{ik}\alpha_k + \sum_{j\in\widehat{\mathcal{S}}_{x,PL}} X_{ij}\beta_j + \sum_{\ell\in\widehat{\mathcal{S}}_{x,N}} \phi_\ell(X_{i\ell}) + \epsilon_i. \tag{8}$$

We approximate the smooth functions $\{\phi_\ell : \ell \in \widehat{\mathcal{S}}_{x,N}\}$ in (8) by polynomial splines introduced in Section 2.2. Let $\mathcal{B}_\ell^{(d)}$ be the space of polynomial splines of order d, and

$$\mathcal{B}_\ell^0 = \{b \in \mathcal{B}_\ell^{(d)} : \mathrm{E}\{b(X_\ell)\} = 0,\ \mathrm{E}\{b^2(X_\ell)\} < \infty\}.$$

Working with $\mathcal{B}_\ell^0$ ensures that the spline functions are centered; see, e.g., [29–31].

Let $B_{1,\ell}^{(d)}, \ldots, B_{M_n,\ell}^{(d)}$ be a set of standardized spline basis functions for $\mathcal{B}_\ell^0$ with dimension $M_n = N_n + d - 1$, where

$$B_{J,\ell}^{(d)}(x_\ell) = b_{J,\ell}^{(d)}(x_\ell)/\|b_{J,\ell}^{(d)}\|_2$$

for all $J \in \{1, \ldots, M_n\}$, so that $\mathrm{E}\{B_{J,\ell}^{(d)}(X_\ell)\} \equiv 0$, $\mathrm{E}\{B_{J,\ell}^{(d)}(X_\ell)\}^2 \equiv 1$. Specifically, if $d = 1$, $M_n = N_n$ and $B_{J,\ell}^{(1)}$ is the standardized piecewise constant spline function defined in (4).

We propose a one-step backfitting using refitted pilot spline estimators in the first stage followed by local-linear estimators. The refitted coefficients are defined as

$$(\widehat{\boldsymbol{\alpha}}^*, \widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\gamma}}^*) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{k\in\widehat{\mathcal{S}}_z} Z_{ik}\alpha_k - \sum_{j\in\widehat{\mathcal{S}}_{x,PL}} X_{ij}\beta_j - \sum_{\ell\in\widehat{\mathcal{S}}_{x,N}} \mathbf{B}_{i\ell}^{(d)}\boldsymbol{\gamma}_\ell \right\}^2. \tag{9}$$

Then the refitted spline estimator for nonlinear functions ϕ_ℓ is, for $\ell \in \widehat{\mathcal{S}}_{x,N}$,

$$\widehat{\phi}_\ell^*(x_\ell) = \mathbf{B}_\ell^{(d)}(x_\ell)\,\widehat{\boldsymbol{\gamma}}_\ell^*. \tag{10}$$

Next we establish the asymptotic normal distribution for the parametric estimators. To make $\boldsymbol{\beta}_{0,\mathcal{S}_Z}$ estimable at the $\sqrt{n}$-rate, we need a condition to ensure $\boldsymbol{X}$ and $\boldsymbol{Z}$ are not functionally related. Define the Hilbert space of theoretically centered L_2 additive functions as

$$\mathcal{F}_+ = \left\{ f(\boldsymbol{x}) = \sum_{\ell\in\mathcal{S}_{x,N}} f_\ell(x_\ell),\ \mathrm{E}\{f_\ell(X_\ell)\} = 0,\ \ \|f_\ell\|_2 < \infty \right\}.$$

For any $k \in \mathcal{S}_z$, let z_k be the coordinate mapping that maps $\boldsymbol{Z}$ to its kth component so that $z_k(\boldsymbol{Z}) = Z_k$, and let

$$\psi_k^z = \mathrm{argmin}_{\psi\in\mathcal{F}_+}\|z_k - \psi\|_2^2 = \mathrm{argmin}_{\psi\in\mathcal{F}_+}\mathrm{E}\{Z_k - \psi(\boldsymbol{X})\}^2$$

be the orthogonal projection of z_k onto $\mathcal{F}_+$. Let $\widetilde{\boldsymbol{Z}}_{\mathcal{S}_z} = \{\psi_k^z(\boldsymbol{X}) : k \in \mathcal{S}_z\}^\top$. Similarly, for any $\ell \in \mathcal{S}_{x,PL}$, let x_ℓ be the coordinate mapping that maps $\boldsymbol{X}$ to its ℓth component so that $x_\ell(\boldsymbol{X}) = X_\ell$, and let

$$\psi_\ell^x = \mathrm{argmin}_{\psi\in\mathcal{F}_+}\|x_\ell - \psi\|_2^2 = \mathrm{argmin}_{\psi\in\mathcal{F}_+}\mathrm{E}\{X_\ell - \psi(\boldsymbol{X})\}^2 \tag{11}$$

be the orthogonal projection of x_ℓ onto $\mathcal{F}_+$.

Let $\widetilde{\boldsymbol{X}}_{\mathcal{S}_{x,PL}} = \{\psi_\ell^x(\boldsymbol{X}), \ell \in \mathcal{S}_{x,PL}\}^\top$. Define $\boldsymbol{Z}_{\mathcal{S}_z} = (\boldsymbol{Z}_k, k \in \mathcal{S}_z)^\top$ and $\boldsymbol{X}_{\mathcal{S}_{x,PL}} = \left(\boldsymbol{X}_\ell, \ell \in \mathcal{S}_{x,PL}\right)^\top$. Denote vector $\boldsymbol{T}$ and $\widetilde{\boldsymbol{T}}$ as $\boldsymbol{T} = (\boldsymbol{Z}_{\mathcal{S}_z}, \boldsymbol{X}_{\mathcal{S}_{x,PL}})^\top$, $\widetilde{\boldsymbol{T}} = (\widetilde{\boldsymbol{Z}}_{\mathcal{S}_z}, \widetilde{\boldsymbol{X}}_{\mathcal{S}_{x,PL}})$.

Theorem 3. *Under the Assumptions (A1), (A2)–(A6), (A3′) and (A6′) in Appendix B.1,*

$$(n\boldsymbol{\Sigma})^{1/2}\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^* - \boldsymbol{\alpha}_{0,\mathcal{S}_z} \\ \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^* - \boldsymbol{\beta}_{0,\mathcal{S}_{x,PL}} \end{pmatrix} \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\mathbf{I}$ *is an identity matrix and* $\boldsymbol{\Sigma} = \sigma^{-2}\mathrm{E}\{(\boldsymbol{T} - \widetilde{\boldsymbol{T}})(\boldsymbol{T} - \widetilde{\boldsymbol{T}})^\top\}$.

The proof of Theorem 3 is similar to the proof of [17] and [22], and thus omitted. Let $\mathbf{Z}_{\mathcal{S}_z} = (Z_{ik}, k \in \mathcal{S}_z)_{i=1}^n$ and $\mathbf{B}_{\mathcal{S}}^{(d)} = (B_{J,\ell}^{(d)}(X_{i\ell}), 1 \le \ell \le p_2, \ell \in \mathcal{S}_{x,N}, J = 1, \ldots, N_n)_{i=1}^n$. If $\mathcal{S}_z$ and $\mathcal{S}_x$ are given, $\boldsymbol{\Sigma}$ can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}}_n = (n\widehat{\sigma}^2)^{-1}(\mathbf{Z}_{\mathcal{S}_z} - \widehat{\mathbf{Z}}_{\mathcal{S}_z})^\top(\mathbf{Z}_{\mathcal{S}_z} - \widehat{\mathbf{Z}}_{\mathcal{S}_z}),$$

where $\widehat{\mathbf{Z}}_{\mathcal{S}_z}^\top = \mathbf{Z}_{\mathcal{S}_z}^\top\mathbf{B}_{\mathcal{S}}^{(d)}\mathbf{U}_{22}^{-1}\mathbf{B}_{\mathcal{S}}^{(d)\top}$ with $\mathbf{U}_{22}$ given in (B.4) in the Appendix B and $\widehat{\sigma}^2 = (n - |\mathcal{S}_z| - |\mathcal{S}_x|)^{-1}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$. In practice, we replace $\mathcal{S}_z$ and $\mathcal{S}_x$ with $\widehat{\mathcal{S}}_z$ and $\widehat{\mathcal{S}}_x$, respectively, to obtain the corresponding estimate.

Let $\Omega_n = \{\widehat{\mathcal{S}}_z = \mathcal{S}_z, \widehat{\mathcal{S}}_{x,PL} = \mathcal{S}_{x,PL}\}$. In the selection step, we estimate $\mathcal{S}_z$ and $\mathcal{S}_{x,PL}$ consistently, i.e., $\Pr(\Omega_n) \to 1$. Within the event Ω_n, i.e., $\widehat{\mathcal{S}}_z = \mathcal{S}_z$ and $\widehat{\mathcal{S}}_{x,PL} = \mathcal{S}_{x,PL}$, the estimator $(\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^{*\top}, \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^{*\top})^\top$ is root-n consistent according to Theorem 3.

Since Ω_n is shown to have probability tending to 1, we can conclude that $(\widehat{\boldsymbol{\alpha}}_{\widehat{\mathcal{S}}_z}^{*\top}, \widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}_{x,PL}}^{*\top})^\top$ is also root-n consistent.

These refitted pilot estimators defined in (9) and (10) are then used to define new pseudo-responses $\widehat{Y}_{i\ell}$, which are estimates of the unobservable "oracle" responses $Y_{i\ell}$. Specifically,

$$\begin{aligned} \widehat{Y}_{i\ell} &= Y_i - \left\{ \sum_{k\in\widehat{\mathcal{S}}_z} Z_{ik}\widehat{\alpha}_k^* + \sum_{\ell'\in\widehat{\mathcal{S}}_{x,PL}} X_{i\ell'}\widehat{\beta}_{\ell'}^* + \sum_{\ell''\in\widehat{\mathcal{S}}_{x,N}\setminus\{\ell\}} \widehat{\phi}_{\ell''}^*(X_{i\ell''}) \right\}, \\ Y_{i\ell} &= Y_i - \left\{ \sum_{k\in\mathcal{S}_z} Z_{ik}\alpha_{0k} + \sum_{\ell'\in\mathcal{S}_{x,PL}} X_{i\ell'}\beta_{0\ell'} + \sum_{\ell''\in\mathcal{S}_{x,N}\setminus\{\ell\}} \phi_{0\ell''}(X_{i\ell''}) \right\}. \end{aligned} \tag{12}$$

Denote K a continuous kernel function, and let $K_{h_\ell}(t) = K(t/h)/h$ be a rescaling of K, where h is usually called the bandwidth. Next, we define the spline-backfitted local-linear (SBLL) estimator of $\phi_\ell(x_\ell)$ as $\widehat{\phi}_\ell^{\mathrm{SBLL}}(x_\ell)$ based on $(X_{11}, \widehat{Y}_{1\ell}), \ldots, (X_{n1}, \widehat{Y}_{n\ell})$, which attempts to mimic the would-be SBLL estimator $\widehat{\phi}_\ell^o(x_\ell)$ of $\phi_\ell(x_\ell)$ based on $(X_{1\ell}, Y_{1\ell}), \ldots, (X_{n\ell}, Y_{n\ell})$ if the unobservable "oracle" responses $Y_{1\ell}, \ldots, Y_{n\ell}$ were available, viz.

$$(\widehat{\phi}_\ell^o(x_\ell), \widehat{\phi}_\ell^{\mathrm{SBLL}}(x_\ell)) = (1\ \ 0)(\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{X}_\ell^*)^{-1}\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell(\mathbf{Y}_\ell, \widehat{\mathbf{Y}}_\ell), \tag{13}$$

where $\mathbf{Y}_\ell = (Y_{1\ell}, \ldots, Y_{n\ell})^\top$ and $\widehat{\mathbf{Y}}_\ell = (\widehat{Y}_{1\ell}, \ldots, \widehat{Y}_{n\ell})^\top$, with $\widehat{Y}_{i\ell}$ and $Y_{i\ell}$ as defined in (12), respectively; and the weight and "design" matrices are

$$\mathbf{W}_\ell = n^{-1}\text{diag}\{K_{h_\ell}(X_{i\ell} - x_\ell)\}_{i=1}^n, \quad \mathbf{X}_\ell^{*\top} = \begin{pmatrix} 1 & , \ldots, & 1 \\ X_{1\ell} - x_\ell & , \ldots, & X_{n\ell} - x_\ell \end{pmatrix}.$$

Asymptotic properties of smoothers of $\widehat{\phi}_\ell^o(x_\ell)$ with $\ell \in \mathcal{S}_{x,N}$ can be easily established. Specifically, let $\mu_2(K) = \int u^2 K(u)\, du$, and let f_ℓ be the probability density function of X_ℓ, then under Assumptions (B1) and (B2) in Appendix B.1, for all $\ell \in \mathcal{S}_{x,N}$,

$$\sqrt{nh_\ell}\,\{\widehat{\phi}_\ell^o(x_\ell) - \phi_{0\ell}(x_\ell) - b_\ell(x_\ell)h_\ell^2\} \rightsquigarrow \mathcal{N}[0, v_\ell^2(x_\ell)], \tag{14}$$

where

$$b_\ell(x_\ell) = \mu_2(K)\phi_{0\ell}''(x_\ell)/2, \quad v_\ell^2(x_\ell) = \|K\|_2^2 f_\ell^{-1}(x_\ell)\sigma^2. \tag{15}$$

The following theorem states that the asymptotic uniform magnitude of the difference between $\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell)$ and $\widehat{\phi}_\ell^o(x_\ell)$ is of order $o_P\{(nh_\ell)^{-1/2}\}$, which is dominated by the asymptotic uniform size of $\widehat{\phi}_\ell^o(x_\ell) - \phi_{0\ell}(x_\ell)$.

As a result, $\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell)$ will have the same asymptotic distribution as $\widehat{\phi}_\ell^o(x_\ell)$. We say $x_\ell \in \chi_\ell$ is a boundary point if and only if $x_\ell = a + ch_\ell$ or $x_\ell = b - ch_\ell$ for some $c \in [0, 1]$ and an interior point otherwise. Let χ_{h_ℓ} be the interior of the support χ.

Theorem 4. *Suppose the assumptions in Theorem 3 hold. In addition, if Assumptions (B1) and (B2) in Appendix B.1 are satisfied, then the SBLL estimator $\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell)$ given in (13) satisfies, for all $\ell \in \mathcal{S}_{x,N}$,*

$$\sup_{x_\ell \in \chi_{h_\ell}} \left|\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell) - \widehat{\phi}_\ell^o(x_\ell)\right| = o_P\{(nh_\ell)^{-1/2}\}. \tag{16}$$

Hence with $b_\ell(x_\ell)$ and $v_\ell^2(x_\ell)$ as defined in (15), for any x_ℓ in its interior support $x_\ell \in \chi_{h_\ell}$, any $\ell \in \mathcal{S}_{x,N}$,

$$\sqrt{nh_\ell}\,\{\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell) - \phi_{0\ell}(x_\ell) - b_\ell(x_\ell)h_\ell^2\} \rightsquigarrow \mathcal{N}[0, v_\ell^2(x_\ell)]. \tag{17}$$

In addition, the estimator $\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell)$ satisfies, for any t and $\ell \in \mathcal{S}_{x,N}$,

$$\lim_{n\to\infty} \Pr\left[\sqrt{\ln(h_\ell^{-2})}\left\{\sup_{x_\ell \in \chi_{h_\ell}} \frac{\sqrt{nh_\ell}}{v_\ell(x_\ell)}\,|\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell) - \phi_{0\ell}(x_\ell)| - \tau_n\right\} < t\right] = \exp(-2e^{-t}), \tag{18}$$

where $\tau_n = \sqrt{\ln(h_\ell^{-2})} + \ln\{\|K'\|_2/(2\pi\|K\|_2)\}/\sqrt{\ln(h_\ell^{-2})}$.

Theorem 4 provides analytical expressions for constructing asymptotic confidence intervals and SCBs under certain conditions. Under Assumptions (A1)–(A6), (A3′), (A6′), (B1) and (B2) in Appendix B.1, for any $\alpha \in (0, 1)$, an asymptotic $100(1-\alpha)\%$ point-wise confidence interval for $\phi_{0\ell}(x_\ell)$ over the interval χ_{h_ℓ} is, for $\ell \in \mathcal{S}_{x,N}$,

$$\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell) - \widehat{b}_\ell(x_\ell)h_\ell^2 \pm \widehat{v}_\ell(x_\ell)(nh_\ell)^{-1/2}.$$

Under Assumptions (A1)–(A6), (A2′) (A3′), (A6′), (B1) and (B2) in the Appendix, for any $\alpha \in (0, 1)$, an asymptotic $100(1-\alpha)\%$ SCB for $\phi_{0\ell}(x_\ell)$ over the interval χ_{h_ℓ} is, for $\ell \in \mathcal{S}_{x,N}$,

$$\widehat{\phi}_\ell^{\text{SBLL}}(x_\ell) \pm \widehat{v}_\ell(x_\ell)(nh_\ell)^{-1/2}\left[\tau_n - \{\ln(h_\ell^{-2})\}^{-1/2}\ln\left\{-\frac{1}{2}\ln(1-\alpha)\right\}\right].$$

4. Implementation and simulation

In this section we discuss practical implementations for the SMILE procedure. To meet the zero mean requirement specified in Assumption (A4), we use the centralized $X_{i\ell}^*$ instead of $X_{i\ell}$ directly, for each $\ell \in \{1, \ldots, p_2\}$. At the risk of abusing the notation, we still use symbol X instead of X^* to avoid creating too many new symbols. To implement the proposed procedure, one needs to select the penalty parameters, the knots for a spline at the selection stage and refitting stage, and the bandwidth for a kernel at the backfitting stage.

Knot selection. For spline smoothing involved in both selection and refitting, we suggest placing knots on a grid of evenly spaced sample quantiles. Based on extensive simulation experiments in Appendix A, we find that the number of knots often has little effect on the model selection results. Therefore, we recommend using a small number of knots at the model selection stage to reduce the computing cost, especially when the sample size is small compared to the number of covariates. In practice, 2 to 5 interior knots is usually adequate to identify the model structure.

At the refitting stage, Assumption (A6′) in the Appendix suggests the number of interior knots M_n for a refitting spline needs to satisfy: $\{n^{1/(2d)} \vee n^{4/(10d-5)}\} \ll M_n \ll n^{1/3}$, where d is the degree of the polynomial spline basis functions used in

Algorithm 1 Iterative group coordinate descent algorithm

Input : Data $\{(Y_i, Z_{i1}, \ldots, Z_{ip_1}, X_{i1}, \ldots, X_{ip_2}, \mathbf{B}_{i1}^{(1)}, \ldots, \mathbf{B}_{ip_2}^{(1)})\}_{i=1}^n$
$\widehat{\boldsymbol{\alpha}}^{(0)}$, $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\gamma}}^{(0)}$: initial parameters of interest
δ_0: convergence criterion
Output: $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$: Estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

while $\|(\widehat{\boldsymbol{\alpha}}^{(m+1)\top}, \widehat{\boldsymbol{\beta}}^{(m+1)\top}, \widehat{\boldsymbol{\gamma}}^{(m+1)\top})^\top - (\widehat{\boldsymbol{\alpha}}^{(m)\top}, \widehat{\boldsymbol{\beta}}^{(m)\top}, \widehat{\boldsymbol{\gamma}}^{(m)\top})^\top\|^2 > \delta_0$ **do**
(i) Given $\widehat{\boldsymbol{\beta}}^{(m)}$ and $\widehat{\boldsymbol{\gamma}}^{(m)}$, obtain $w_1^{\alpha(m+1)}, \ldots, w_{p_1}^{\alpha(m+1)}$ by minimizing objective function (6) with $\widetilde{\lambda}_1$ selected via the modified BIC;
(ii) Given $\widehat{\boldsymbol{\beta}}^{(m)}$, $\widehat{\boldsymbol{\gamma}}^{(m)}$ and $w_1^{\alpha(m+1)}, \ldots, w_{p_1}^{\alpha(m+1)}$, obtain $\widehat{\boldsymbol{\alpha}}^{(m+1)}$ by minimizing objective function (7) with λ_1 selected via the modified BIC;
(iii) Given $\widehat{\boldsymbol{\alpha}}^{(m+1)}$ and $\widehat{\boldsymbol{\gamma}}^{(m)}$, obtain $w_1^{\beta(m+1)}, \ldots, w_{p_2}^{\beta(m+1)}$ by minimizing objective function (6) with $\widetilde{\lambda}_2$ selected via the modified BIC;
(iv) Given $\widehat{\boldsymbol{\alpha}}^{(m+1)}$, $\widehat{\boldsymbol{\gamma}}^{(m)}$ and $w_1^{\beta(m+1)}, \ldots, w_{p_2}^{\beta(m+1)}$, obtain $\widehat{\boldsymbol{\beta}}^{(m+1)}$ by minimizing objective function (7) with λ_2 selected via the modified BIC;
(v) Given $\widehat{\boldsymbol{\alpha}}^{(m+1)}$ and $\widehat{\boldsymbol{\beta}}^{(m+1)}$, obtain $w_1^{\gamma(m+1)}, \ldots, w_{p_2}^{\gamma(m+1)}$ by minimizing objective function (6) with $\widetilde{\lambda}_3$ selected via EBIC;
(vi) Given $\widehat{\boldsymbol{\alpha}}^{(m+1)}$, $\widehat{\boldsymbol{\beta}}^{(m+1)}$ and $w_1^{\gamma(m+1)}, \ldots, w_{p_2}^{\gamma(m+1)}$, obtain $\widehat{\boldsymbol{\gamma}}^{(m+1)}$ by minimizing objective function (7) with λ_3 selected via EBIC.
end
Set $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}^{(m+1)}$, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(m+1)}$ and $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}^{(m+1)}$.

the refitting. The widely used quadratic/cubic splines and any polynomial splines of degree $d \geq 2$ all satisfy this condition. Therefore, in practice we suggest taking the following rule-of-thumb number of interior knots

$$\min\{\lfloor n^{1/(2d)\vee 4/(10d-5)} \ln(n)\rfloor, \lfloor n/(4s)\rfloor\} + 1,$$

where s is the number of nonlinear components selected at the first stage, and the term $\lfloor n/(4s)\rfloor$ is to guarantee that we have at least four observations in each subinterval between two adjacent knots to avoid (near) singular design matrices in the spline refitting.

Bandwidth selection. Note that Condition (B2) in the Appendix requires that the bandwidths in the backfitting are of order $n^{-1/5}$. Thus, the bandwidth selection can be done using a standard routine in the literature. In our numerical studies, we find that the rule-of-thumb bandwidth selector [8] often works very well in both estimation and SCB construction.

Appendix A provides detailed investigations on how the smoothing parameters affect the proposed SMILE method and evaluates the practical performance in finite-sample simulation studies. Next we present our algorithm and discuss how to choose the penalty parameters.

4.1. Algorithm

The minimization of (7) can be solved by the group coordinate descent algorithm [11], implemented using `R` package `grpreg` [2]. As for the selection of penalty parameters, we consider two criteria widely used in high-dimensional settings, modified Bayesian information criteria (BIC, see [14]) and the extended BIC (EBIC, see [4,5]):

$$\begin{aligned}\mathrm{BIC}(\boldsymbol{\lambda}) &= \ln(RSS_{\boldsymbol{\lambda}}) + df_{\boldsymbol{\lambda}} \times \{\ln(p_1 + p_2 + p_2 N_n) \times \ln(n)\}/(2n),\\ \mathrm{EBIC}(\boldsymbol{\lambda}) &= \ln(RSS_{\boldsymbol{\lambda}}) + df_{\boldsymbol{\lambda}} \times \ln(n)/n + df_{\boldsymbol{\lambda}} \times \{\ln(p_1 + p_2 + p_2 N_n)\}/n,\end{aligned}$$

where $RSS_{\boldsymbol{\lambda}}$ is the residual sum of squares associated with penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^\top$ and $df_{\boldsymbol{\lambda}}$ is the number of estimated nonzero coefficients for the given $\boldsymbol{\lambda}$. The simulation results are similar based on these two criteria, so in the following, we choose λ_1 and λ_2 by modified BIC and λ_3 by EBIC for illustration using an approach described below.

The classical coordinate descent algorithm deals with the optimization problem with one tuning parameter, and there are several ways to address the triple-penalization or multiple-penalization issue. A natural idea is to solve the optimization problem by searching over a three-dimensional grid for tuning parameters, which can be computationally expensive. To pose a balance between computational efficiency and precision, we propose to solve the triple-penalization problem in two steps. In the first step, BIC is minimized with a common smoothing parameter λ, i.e., we set $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$, and we choose λ by minimizing $\mathrm{BIC}(\lambda)$ over a grid of λ values. Using the selected common smoothing parameter, we obtain the initial estimators $\widehat{\boldsymbol{\alpha}}^{(0)}$, $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\gamma}}^{(0)}$. In Step 2, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ estimates are obtained one at a time by minimizing (7). More precisely, an $\boldsymbol{\alpha}$ estimate is obtained with $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ fixed at current estimates, where λ_1 is set equal to its minimum BIC value and $\lambda_2 = \lambda_3 = 0$. One cycles in this way through $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ estimation steps for a fixed number of iterations. Three iterations generally works well in practice. Algorithm 1 outlines the iterative group coordinate descent algorithm.

4.2. Simulation studies

In this section, we investigate the performance of the proposed sparse model identification and learning estimator, abbreviated as SMILE, in terms of model selection, estimation accuracy and inference performance in a simulation study. Additional simulation results can be found in the longer full version of the paper [16], which show that our proposed SMILE procedure performs well relative to competing methods under a wider range of conditions; see Appendices B–D of [16].

We compare SMILE with the sparse APLM estimator with adaptive group LASSO penalty (SAPLM) proposed in [17], the ordinary linear least squares estimator with the adaptive LASSO penalty (SLM), and the oracle estimator (ORACLE), which uses the same estimation techniques as SMILE except that no penalization or data-driven variable selection is used because all active and inactive index sets are treated as known. Note that SAPLM ignores the potential linear structure in covariate $\boldsymbol{X}$, and estimates the effects of each component of $\boldsymbol{X}$ with all nonparametric forms; in contrast, SLM ignores the potential nonlinear structure in covariate $\boldsymbol{X}$ and requires selected components of covariates $\boldsymbol{Z}$ and $\boldsymbol{X}$ to enter the model in a linear form. In terms of the performances of SCBs, we compare SMILE with SAPLM and ORACLE. In our simulation, ORACLE works as a benchmark for estimation comparison. It is worth pointing out that the ORACLE estimator is only computable in simulations, not real examples.

We generate simulated datasets using the APLM structure

$$Y_i = \sum_{k=1}^{p_1} Z_{ik}\alpha_k + \sum_{\ell=1}^{p_2} \phi_\ell(X_{i\ell}) + \varepsilon_i,$$

where $\alpha_1 = 3$, $\alpha_2 = 4$, $\alpha_3 = -2$, $\alpha_4 = \cdots = \alpha_{p_1} = 0$,

$$\phi_1(x) = 9x, \quad \phi_2(x) = -1.5\cos^2(\pi x) + 3\sin^2(\pi x) - \mathrm{E}\{-1.5\cos^2(\pi X_2) + 3\sin^2(\pi X_2)\},$$

and

$$\phi_3(x) = 6x + 18x^2 - \mathrm{E}(6X_3 + 18X_3^2), \quad \phi_4(x) = \cdots = \phi_{p_2}(x) = 0.$$

Notice that ϕ_1 is actually a linear function. So there are three variables in the active index set for $\boldsymbol{Z}$, one variable in the active pure linear index set for $\boldsymbol{X}$, one variable in the active pure nonlinear index set for $\boldsymbol{X}$, and one variable in the active linear and nonlinear index set for $\boldsymbol{X}$.

We simulate Z_{ik}^* independently from the $\mathcal{U}[0, 1]$ and $X_{i\ell}$ independently from the $\mathcal{U}[-.5, .5]$, and set $Z_{ik} = \mathbf{1}(Z_{ik}^* > 0.75)$, for $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, p_1\}$, $\ell \in \{1, \ldots, p_2\}$. To make an ultra-high-dimensional scenario, we let the sample size $n = 300$ and $n = 500$, and consider three different dimensions: $p_1 = p_2 = p$, where p is taken to be 1000, 2000 and 5000. The error term ε_i is simulated from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$ and 1.0.

To approximate the nonlinear functions, we use the constant B-spline ($d = 1$) with four interior knots for selection and use the cubic B-spline ($d = 4$) with four interior knots in the refitting step. For both selection and refitting, the knots are on a grid of evenly spaced sample quantiles. To construct the SCBs, in our simulation studies below, we choose the Epanechnikov kernel function with the rule-of-thumb bandwidth described in Section 4.2 in [8], which usually works well in our experimental investigation. More simulation studies have been conducted with different choices for spline knots and kernel bandwidth selectors; see Section A of the Appendix.

We evaluate the methods on the accuracy of variable selection, prediction and inference. In detail, we adopt the following criteria for evaluation:

(B-i) percent of covariates in $\boldsymbol{Z}$ with nonzero linear coefficients that are correctly identified (“CorrZ”);
(B-ii) percent of covariates in $\boldsymbol{Z}$ with zero linear coefficients that are correctly identified (“CorrZ0”);
(B-iii) percent of covariates in $\boldsymbol{X}$ with nonzero purely linear functions that are correctly identified (“CorrL”);
(B-iv) percent of covariates in $\boldsymbol{X}$ with nonzero purely nonlinear functions that are correctly identified (“CorrN”);
(B-v) percent of covariates in $\boldsymbol{X}$ with nonzero linear and nonlinear functions that are correctly identified (“CorrLN’);
(B-vi) percent of covariates in $\boldsymbol{X}$ with zero functions that are correctly identified (“CorrX0”);

(C-i) percent of covariates in $\boldsymbol{Z}$ with nonzero linear coefficients incorrectly identified as having zero linear coefficients (“Zto0”);
(C-ii) percent of covariates in $\boldsymbol{X}$ with nonzero purely linear functions incorrectly identified as having nonlinear functions (“LtoN”);
(C-iii) percent of covariates in $\boldsymbol{X}$ with nonzero purely nonlinear functions incorrectly identified as having linear functions (“NtoL”);
(C-iv) percent of covariates in $\boldsymbol{X}$ with nonzero linear or nonzero nonlinear functions incorrectly identified as having both zero linear and zero nonlinear functions (“Xto0”);

(D-i) mean squared errors (MSE) for linear coefficients α_1, α_2, α_3 and β_1;
(D-ii) average MSE (AMSE) for ϕ_1, ϕ_2 and ϕ_3, defined as $\sum_{i=1}^{n}\{\widehat{\phi}_\ell^{\mathrm{SBLL}}(x_{i\ell}) - \phi_\ell(x_{i\ell})\}^2/n$;

Table 1
Statistics (B-i)–(B-vi) comparing the SMILE, SAPLM and SLM.

Size	Noise	p	Method	Z Part		X Part			
n	sig			corrZ	corrZ0	corrL	corrN	corrLN	corrX0
300	0.5	1000	SMILE	100	99.99960	100	100	100	99.99940
			SAPLM	100	100	0	100	0	100
			SLM	98.6	99.99920	100	0	0	99.99850
		2000	SMILE	100	99.99995	100	100	100	99.99985
			SAPLM	100	100	0	100	0	100
			SLM	97.3	99.99950	100	0	0	99.99915
		5000	SMILE	100	99.99996	100	100	100	100
			SAPLM	100	100	0	100	0	100
			SLM	96.63333	99.99988	100	0	0	99.99974
	1.0	1000	SMILE	100	99.99920	100	100	100	99.99990
			SAPLM	100	99.99920	0	100	0	100
			SLM	96.56667	99.99799	100	0	0	99.99719
		2000	SMILE	99.93333	99.99995	100	99.8	99.8	99.99975
			SAPLM	100	99.99970	0	100	0	100
			SLM	95.7	99.99975	100	0	0	99.99905
		5000	SMILE	99.86667	99.99996	100	99.5	99.5	99.99996
			SAPLM	100	99.99990	0	100	0	100
			SLM	93.73333	99.99982	100	0	0	99.99978
500	0.5	1000	SMILE	100	99.99990	100	100	100	99.99980
			SAPLM	100	100	0	100	0	100
			SLM	100	99.99990	100	0	0	99.99960
		2000	SMILE	100	99.99995	100	100	100	100
			SAPLM	100	100	0	100	0	100
			SLM	100	99.99985	100	0	0	99.99985
		5000	SMILE	100	99.99996	100	100	100	100
			SAPLM	100	100	0	100	0	100
			SLM	99.96667	99.99994	100	0	0	99.99994
	1.0	1000	SMILE	100	99.99950	100	100	100	99.99970
			SAPLM	100	100	0	100	0	100
			SLM	99.96667	99.99940	100	0	0	99.99930
		2000	SMILE	100	99.99980	100	100	100	99.99990
			SAPLM	100	99.99990	0	100	0	100
			SLM	99.93333	99.99990	100	0	0	99.99960
		5000	SMILE	100	99.99994	100	100	100	100
			SAPLM	100	100	0	100	0	100
			SLM	99.76667	100	100	0	0	99.99994

(D-iii) 10-fold cross-validation mean squared prediction error (CV-MSPE) for the response variable, defined as $\sum_{m=1}^{10}\sum_{i\in\kappa_m}(\widehat{Y}_i - Y_i)^2/(10|\kappa_m|)$, where $\kappa_1,\ldots,\kappa_{10}$ comprise a random partition of the dataset into 10 disjoint subsets of approximately equal size, and $\widehat{Y}_i$ is the prediction obtained from all data aside from the subset containing the ith observation;

(D-iv) the coverage rates of the proposed 95% SCB for functions ϕ_2 and ϕ_3 (Coverage).

All these performance measures are computed based on 1000 replicates. Note that Criteria (B-i)–(B-vi) measure the frequency of getting the correct model structure; Criteria (C-i)–(C-iv) measure the frequency of getting an incorrect model structure; Criteria (D-i)–(D-iii) focus on the estimation and prediction accuracy for the model components; and Criterion (D-iv) measures the inferential performance.

The model selection results are provided in Tables 1 and 2, respectively. SMILE can effectively identify informative linear and nonlinear components as well as correctly discover the linear and nonlinear structure in covariate $\boldsymbol{X}$, while SAPLM neglects linear structure in $\boldsymbol{X}$ and SLM fails in representing the nonlinear part of covariate $\boldsymbol{X}$. For SMILE, the numbers of correctly selected nonzero covariates in $\boldsymbol{Z}$, linear, nonlinear, linear-and-nonlinear components in $\boldsymbol{X}$, nonzero covariates are very close to ORACLE (100% for corrZ, corrL, corrN, corrLN, corrZ0 and corrX0, respectively); and the numbers of incorrectly identified components approach to 0 as the sample size n increases, as shown in Table 2. SMILE is close in the selection of covariates $\boldsymbol{Z}$ to the SAPLM estimator, and it far outperforms SAPLM in identifying the linear-and-nonlinear structure of covariate $\boldsymbol{X}$. From the results in Tables 1 and 2, it is also evident that model misspecification leads to poor variable selection performance for SLM. Especially for the selection of covariates in $\boldsymbol{X}$, which is our main focus for real data analysis, SLM fails to select the right nonlinear components in each simulation.

The estimation and prediction results are displayed in Table 3. Specifically, we present the MSEs for linear coefficients α_1, α_2, α_3 and β_1 and AMSEs for functions ϕ_1, ϕ_2 and ϕ_3 and the CV-MSPEs for predicting Y. The case with known active covariates (ORACLE) is also reported in each setting and serves as a gold standard. SMILE performs the best in predicting Y and estimating the coefficients of covariates $\boldsymbol{Z}$, as indicated by CV-MSPE and MSEs for α_1, α_2 and α_3 that are closest to ORACLE in most simulation settings, while SLM is much higher (around $2\sim 18$ times higher). As for the linear structure

Table 2
Statistics (C-i)–(C-iv) comparing the SMILE, SAPLM and SLM.

Size	Noise	p	Method	Z Part	X Part		
n	sig			Zto0	LtoN	NtoL	Xto0
300	0.5	1000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	1.4	0	100	33.33333
		2000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	2.7	0	100	33.33333
		5000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	3.36667	0	100	33.33333
	1.0	1000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	3.43333	0	100	33.33333
		2000	SMILE	0.06667	0	0	0.06667
			SAPLM	0	100	0	0
			SLM	4.3	0	100	33.33333
		5000	SMILE	0.13333	0	0	0.16667
			SAPLM	0	100	0	0
			SLM	6.26667	0	100	33.33333
500	0.5	1000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0	0	100	33.33333
		2000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0	0	100	33.33333
		5000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0.03333	0	100	33.33333
	1.0	1000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0.03333	0	100	33.33333
		2000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0.06667	0	100	33.33333
		5000	SMILE	0	0	0	0
			SAPLM	0	100	0	0
			SLM	0.23333	0	100	33.33333

in $\boldsymbol{X}$, as shown in MSE for β_1 and AMSE for ϕ_1, the performance of SMILE is comparable to SAPLM and SLM, even though restricted to the selection bias; as the sample size n increases, the performance of SMILE is perfect and matches with ORACLE. Note that the SAPLM estimator is incapable in estimating β_1 in this case. The estimation of nonlinear functions ϕ_2 and ϕ_3 is also good for SMILE, and matches with ORACLE as sample size n increases. The inferior performance of SAPLM and the poor performance of SLM, in both estimation and prediction, illustrates the importance and necessity of identifying correct model structure.

Next we investigate the coverage rates of the proposed SCB. For each replication, we test whether the true functions are covered by the SCB at the simulated values of the covariate in the interval $[-0.5 + h, 0.5 - h]$, where h is the bandwidth. Table 4 shows the empirical coverage probabilities for a nominal 95% confidence level out of 500 replications. For comparison, we also provide the SCBs from the SAPLM and ORACLE estimators. From Table 4, we observe that coverage probabilities for the SMILE, SAPLM and ORACLE SCBs all approach the nominal levels as n increases, which provides positive confirmation of Theorem 4. In most cases, SMILE performs as well as or better than SAPLM, and arrives at about the nominal coverage when $n = 500$ and $\sigma = 1.0$. Fig. 1 depicts the true function ϕ_ℓ, the corresponding SMILE $\widehat{\phi}_\ell^{\mathrm{SBLL}}$ and the 95% SCB for ϕ_ℓ based on $\widehat{\phi}_\ell^{\mathrm{SBLL}}$, for $\ell = 2, 3$, which are based on a typical run with $n = 500$, $p = 1000$ and $\sigma = 1.0$.

5. Application

We illustrate the application of our proposed method in the ultra-high-dimensional setting by using the SAM data generated by [15] and further analyzed by [21]. The maize SAM is a small pool of stem cells located in the plant shoot that generate all the above-ground tissues of maize plants. Leiboff et al. [15] showed that SAM volume is correlated with a variety of agronomically important traits in adult plants. The goal of our analysis is to model and predict SAM volume as a function of single nucleotide polymorphism (SNP) genotypes and messenger RNA transcript abundance levels using data from maize inbred lines. Following the preprocessing steps described in Section B.5 in the Appendix in [17], linear sure independent screening [10] for SNP genotypes, and nonlinear independent screening [7] for RNA transcripts, the dataset

Table 3
Estimation results comparing the ORACLE, SMILE, SAPLM and SLM.

n	σ	p	Method	MSE ($\times 10^{-2}$)				AMSE ($\times 10^{-2}$)			CV-
				α_1	α_2	α_3	β_1	ϕ_1	ϕ_2	ϕ_3	MSPE
300	0.5	1000	ORACLE	0.47	0.48	0.50	1.06	0.09	0.98	0.83	0.28
			SMILE	0.47	0.48	0.50	1.06	0.11	0.94	0.77	0.28
			SAPLM	0.49	0.48	0.54	–	0.41	0.99	0.83	0.28
			SLM	9.13	8.86	25.99	18.65	1.63	253.82	178.85	4.79
		2000	ORACLE	0.47	0.46	0.47	1.08	0.09	0.99	0.85	0.27
			SMILE	0.47	0.45	0.47	1.08	0.19	0.95	0.79	0.27
			SAPLM	0.49	0.47	0.53	–	0.43	1.01	0.86	0.28
			SLM	10.51	8.70	41.05	21.29	1.84	252.75	180.40	4.80
		5000	ORACLE	0.45	0.44	0.53	1.06	0.09	0.97	0.81	0.27
			SMILE	0.45	0.44	0.53	1.06	0.16	0.94	0.75	0.27
			SAPLM	0.47	0.47	0.57	–	0.42	0.98	0.81	0.28
			SLM	9.29	8.66	49.64	19.90	1.73	252.44	179.41	4.83
	1.0	1000	ORACLE	1.94	1.98	1.82	4.48	0.37	2.97	2.63	1.08
			SMILE	1.94	1.98	1.82	4.48	0.56	2.80	2.29	1.09
			SAPLM	1.98	2.01	1.96	–	1.44	2.98	2.53	1.09
			SLM	11.33	10.55	51.08	22.51	1.95	253.34	180.31	5.59
		2000	ORACLE	1.90	1.77	1.82	4.16	0.35	3.04	2.57	1.08
			SMILE	1.91	1.84	2.62	4.23	0.73	3.30	3.20	1.11
			SAPLM	1.98	1.85	1.98	–	1.40	3.07	2.49	1.09
			SLM	11.04	10.19	60.67	23.02	1.99	252.71	179.98	5.61
		5000	ORACLE	1.71	1.89	1.93	4.03	0.33	2.93	2.54	1.08
			SMILE	1.82	1.92	3.52	4.05	0.39	3.97	4.67	1.28
			SAPLM	1.77	1.97	2.09	–	1.43	2.96	2.44	1.25
			SLM	16.78	10.55	80.30	23.28	2.01	252.41	180.60	5.72
500	0.5	1000	ORACLE	0.27	0.28	0.28	0.67	0.06	0.67	0.58	0.27
			SMILE	0.27	0.28	0.28	0.67	0.07	0.65	0.55	0.27
			SAPLM	0.29	0.29	0.31	–	0.27	0.67	0.58	0.27
			SLM	5.08	4.91	5.88	10.70	0.97	253.20	180.13	4.66
		2000	ORACLE	0.27	0.27	0.31	0.65	0.05	0.65	0.55	0.26
			SMILE	0.27	0.27	0.31	0.65	0.06	0.63	0.52	0.26
			SAPLM	0.28	0.28	0.34	–	0.27	0.66	0.55	0.27
			SLM	5.25	4.99	5.90	11.93	1.07	252.96	179.22	4.66
		5000	ORACLE	0.29	0.25	0.29	0.62	0.05	0.67	0.57	0.26
			SMILE	0.29	0.25	0.29	0.62	0.17	0.64	0.54	0.26
			SAPLM	0.30	0.26	0.32	–	0.28	0.67	0.57	0.27
			SLM	5.30	4.87	6.35	11.96	1.07	252.99	179.99	4.66
	1.0	1000	ORACLE	1.18	1.08	1.09	2.43	0.20	1.90	1.62	1.05
			SMILE	1.18	1.08	1.09	2.43	0.56	1.83	1.47	1.05
			SAPLM	1.21	1.12	1.15	–	0.87	1.92	1.60	1.06
			SLM	6.45	5.26	7.42	12.11	1.09	253.05	180.33	5.41
		2000	ORACLE	1.12	1.02	1.12	2.45	0.20	1.94	1.66	1.04
			SMILE	1.12	1.02	1.12	2.45	0.22	1.84	1.49	1.04
			SAPLM	1.15	1.05	1.21	–	0.85	1.94	1.63	1.05
			SLM	6.12	5.99	7.62	13.76	1.22	252.81	180.10	5.43
		5000	ORACLE	1.12	1.05	1.16	2.46	0.20	1.96	1.67	1.05
			SMILE	1.12	1.05	1.16	2.46	0.22	1.87	1.48	1.05
			SAPLM	1.14	1.08	1.22	–	0.87	1.97	1.64	1.06
			SLM	6.16	5.64	9.37	12.28	1.10	252.69	180.26	5.43

we analyze consists of log-scale SAM volume measurements, binary SNP genotypes at $p_1 = 5203$ markers, and log-scale measures of abundance for $p_2 = 1020$ transcripts for each of $n = 368$ maize inbred lines.

Li et al. [17] used the APLM to model the relationship between the log SAM volume response and predictors determined by SNP genotypes and RNA transcript abundance levels. Because the SNP genotypes are binary, they naturally entered the linear part of the APLM, and for convenience all the RNA transcripts were included in the nonlinear part of the APLM in [17]. As discussed before, failing to account for exactly linear features makes the APLM less efficient statistically and computationally. In the following we apply our proposed SMILE method to distinguish among RNA transcripts entering the nonparametric and parametric parts of the APLM and to identify significant SNP genotypes and RNA transcripts simultaneously.

To compare the results of SMILE to the sparse APLM and the sparse linear regression model, we also analyze the data using the SAPLM and SLM estimators presented in [17]. Parallel to the settings in Section 4, we use constant B-splines with four quantile knots for model structure identification, and use cubic B-splines with one quantile knot for nonlinear function approximation. We use the iterative algorithm proposed in Section 4.1 for penalty parameter selection and estimation.

Table 4
Coverage rates comparing the ORACLE, SMILE and SAPLM.

Size	Noise	p	ϕ_2 coverage (%)			ϕ_3 coverage (%)		
n	σ		ORACLE	SMILE	SAPLM	ORACLE	SMILE	SAPLM
300	0.5	1000	93.7	94.5	93.9	92.4	92.6	91.7
		2000	92.6	93.3	92.6	92.3	93.8	92.5
		5000	92.3	93.0	92.7	93.3	92.3	91.7
	1	1000	96.0	95.6	94.7	96.1	96.4	95.3
		2000	95.4	95.7	94.9	96.1	96.2	95.5
		5000	95.1	95.6	94.2	95.9	96.4	94.8
500	0.5	1000	92.9	93.8	93.5	92.7	90.6	92.0
		2000	92.5	92.7	92.3	92.0	92.0	92.3
		5000	92.5	92.6	91.8	91.5	89.9	90.4
	1	1000	97.1	96.7	96.3	96.0	96.0	95.2
		2000	95.2	95.0	94.5	95.2	94.6	94.3
		5000	94.7	95.1	95.0	96.2	96.0	95.5

Table 5
Selected SNPs and transcripts by SMILE, SAPLM and SLM.

RNA transcripts selected	SMILE	SAPLM	SLM
X_{725}	✓	✓	✓
X_{127}, X_{136}, X_{141}, X_{208}, X_{289}, X_{312}, X_{493}, X_{749}, X_{855}	✓		✓
X_{153}[a], X_{677}[a]	✓		
X_{157}, X_{701}		✓	
X_{209}, X_{314}, X_{320}, X_{321}, X_{342}, X_{419}, X_{472}, X_{489}, X_{553},			✓
X_{589}, X_{601}, X_{615}, X_{783}, X_{785}, X_{793}, X_{846}, X_{863}, X_{940},			✓
X_{946}, X_{978}, X_{1002}, X_{1018}			✓
Number of SNP genotypes	169	177	167
Number of linear RNA transcripts	10	0	32
Number of functional RNA transcripts	2	3	0
CV MSPE	0.060	0.102	0.132
CV mean number of SNPs	153.9	175.9	83.1
CV mean number of linear transcripts	8.7	0	17.7
CV mean number of nonlinear transcripts	1.9	3.8	0

[a]Nonlinear association identified by SMILE for X_{153} and X_{677}.

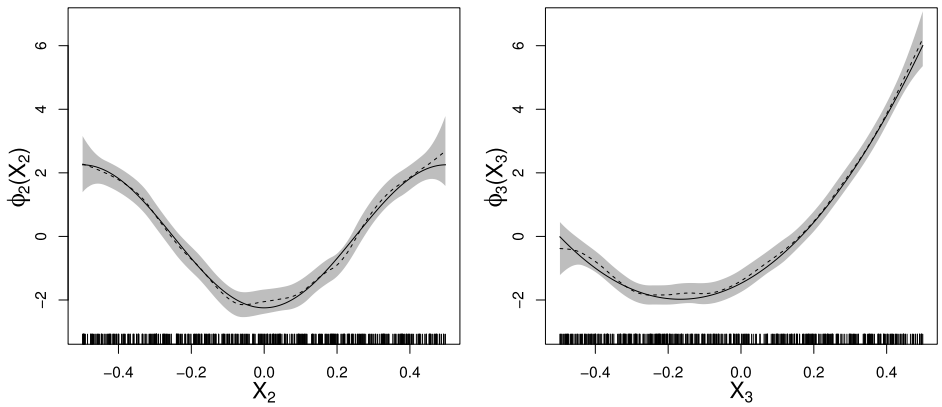

Fig. 1. Plots of the SMILE (dashed curve) and the 95% SCB (shaded area) of the nonparametric component $\phi_\ell(x_\ell)$, $\ell = 2, 3$ (solid curve).

As shown in Table 5, SMILE identified 169 SNPs, 10 RNA transcripts linearly associated with log SAM size and two RNA transcripts that have nonlinear association with log SAM size. In contrast, SAPLM selected 177 SNPs and three RNA transcripts, and SLM selected 167 SNPs and 32 RNA transcripts. To evaluate the predictive performance of the two methods, we computed 10-fold cross-validation mean squared prediction error (CV-MSPE) for each method. The SMILE-estimated nonlinear function for the selected nonlinear RNA transcript is plotted, along with 95% SCBs, in Fig. 2.

6. Discussion

This paper focuses on the simultaneous sparse model identification and learning for ultra-high-dimensional APLMs which strikes a delicate balance between the simplicity of the standard linear regression models and the flexibility of

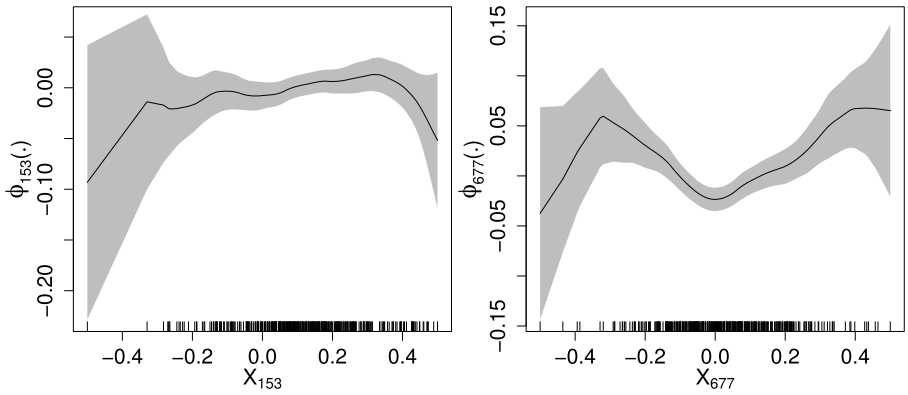

Fig. 2. Plot of the SMILE (solid curve) and the 95% confidence band (shaded area) for the selected RNA transcript.

the additive regression models. We proposed a two-stage penalization method, called SMILE, which can efficiently select nonzero components and identify the linear-and-nonlinear structure in the functional terms, as well as simultaneously estimate and make inference for both linear coefficients and nonlinear functions. First, we have devised a groupwise penalization method in the APLM for simultaneous variable selection and structure identification. After identifying important covariates and the functional forms for the selected covariates, we have further constructed SCBs for the nonzero nonparametric functions based on refined spline-backfitted local-linear estimators. Our simulation studies and applications demonstrate the proposed SMILE procedure can be more efficient than penalized linear regression and the penalized APLM without model identification, and can improve predictions.

Our work differs from previous works in practical, theoretical and computational aspects:

(i) We perform variable selection and model structure identification simultaneously, for both the linear components in $\boldsymbol{Z}$, and the linear and nonlinear forms for the components of $\boldsymbol{X}$. In contrast, existing works either perform only model structure identification or perform variable selection only for components in $\boldsymbol{X}$.

(ii) Besides the consistency of model structure identification, we also provide inference tools for both the regression coefficients and the component functions.

(iii) Compared to the local quadratic approximation approach used in [18], which cannot provide exactly zero solutions and is inefficient for fitting large regression problems, our proposed iterative group coordinate descent algorithm takes advantage of sparsity in computation and is able to deal with the triple penalization problem very efficiently. See [3] for a detailed comparison of these two algorithms.

Our algorithm is easy to implement and can provide analysis results for large data sets with thousands of dimensions within seconds.

Our work deals with independent observations but can be extended to longitudinal data settings through marginal models or mixed-effects models. In addition, although we consider continuous response variables in our work, or approach can be readily extended to generalized additive partially linear models, to deal with different types of responses. Currently, the APLM assumes that the effects of all covariates are additive, which may overlook the potential interaction between covariates. Our method can be extended to models that can accommodate interactions between covariates, for example, APLMs with interaction terms. We leave such extensions to future work. Another limitation of our work is a reliance on the assumption of constant error variance. However, heteroscedasticity may be encountered in the analysis of genomic data sets. It is of interest to develop a new methodology that allows non-constant error variance for high-dimensional estimation and model selection, and this is another challenge we leave for future work.

Acknowledgments

This work was supported by the Iowa State University Plant Sciences Institute Scholars Program. In addition, Wang's research was supported by NSF, USA grant DMS-1542332, and Nettleton's research was supported by NSF, USA grant IOS-1238142. We sincerely thank the Editor-in-Chief, Christian Genest, the Associate Editor and the anonymous reviewers for their insightful comments that have led to significant improvements on the paper.

Appendix A. Effect of smoothing parameters on performance of SMILE

To implement the proposed SMILE procedure, one needs to select the knots for a spline at the selection stage and refitting stage, and the bandwidth for a kernel at the backfitting stage. In this section, we study how these smoothing

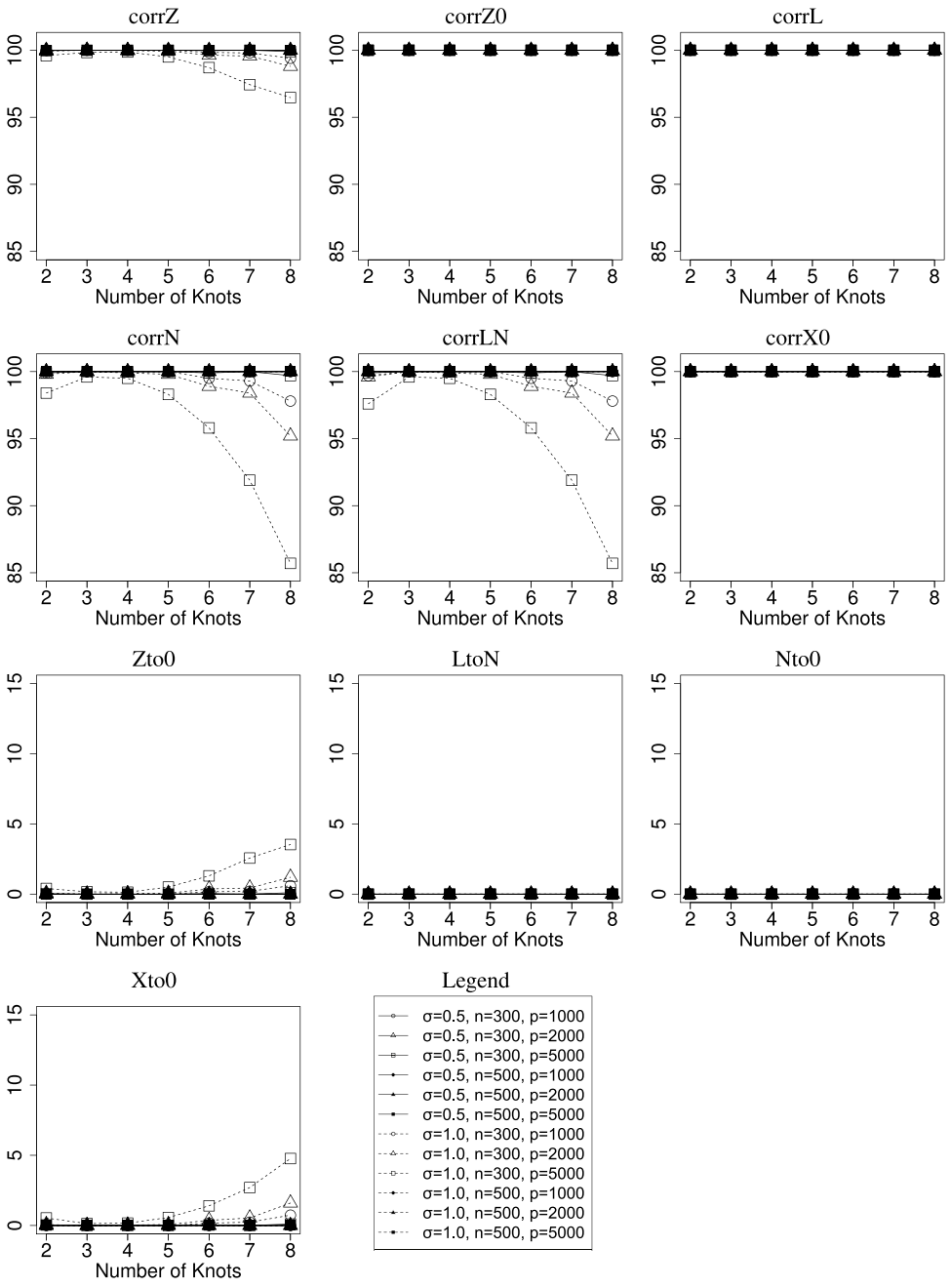

Fig. A.1. First stage selection results using different number of knots.

parameters affect the proposed SMILE method and evaluate the practical performance in the finite-sample simulation studies described in Section 4.2 of the main paper. In the literature of polynomial spline smoothing, the knots for a spline are generally put on a grid of equally spaced sample quantiles [25]. Therefore, we only need to investigate the effect of the number of knots on the performance of SMILE.

At the first stage (model selection), we use piecewise constant splines with the number of interior knots $N \in \{2, \ldots, 8\}$ in the simulation. Fig. A.1 shows the effect of N on the accuracy of model selection based on the criteria defined in the main paper: (B-i)–(B-vi) and (C-i)–(C-iv). From Fig. A.1, it appears that the value N has little effect on the selection results. For all combinations of n, p and σ, no matter which N is used, the "corrZ0", "corrL", "corrX0" are all 100%, and the "LtoN" and "Nto0" are all 0%. The values of "corrZ", "corrN", "corrLN" and "Zto0" and "Xto0" are not exactly the same when using different values of N, but they are almost constant for $N = 2, 3, \ldots, 8$. Especially when the sample size $n = 500$,

the proposed SMILE is able to identify the true model structure regardless of $p \in \{1000, 2000, 5000\}$. When $n = 300$ and $p = 5000$, the selection results become slightly worse when we increase to $N \geq 6$.

In summary, the values of N often have little effect on the model selection results. Choosing small values of N can also help to reduce computational burden. So we recommend using fewer knots at the model selection stage, especially when the sample size is small compared to the number of predictors. In practice, $N = 2 \sim 5$ usually would be adequate to identify the model structure.

Next, we study the effect of the smoothing parameters at the refitting stage. For the selected model, we approximate the nonlinear functional components using higher order polynomial splines to obtain more accurate pilot estimators. Then we apply spline backfitted local-linear smoothing to obtain the final SBLL estimators and the corresponding SCBs. According to Assumption (A6′), to obtain the SCB with the desired confidence level, the number of interior knots M_n for a refitting spline needs to satisfy: $\{n^{1/(2d)} \vee n^{4/(10d-5)}\} \ll M_n \ll n^{1/3}$, where d is the degree of the polynomial spline basis functions used in the refitting. The widely used quadratic/ cubic splines and any polynomial splines of degree $d \geq 2$ all satisfy this condition. Therefore, in practice we suggest choosing

$$M_n = \min\{\lfloor n^{1/(2d)\vee 4/(10d-5)} \ln(n) \rfloor, \lfloor n/(4s) \rfloor\} + 1,$$

where s is the number of nonlinear components selected at the first stage and the term $\lfloor n/(4s) \rfloor$ is to guarantee that we have at least four observations in each subinterval between two adjacent knots to avoid getting (near) singular design matrices in the spline smoothing. A researcher with some knowledge of the shape of the nonlinear component may be able to select a more suitable number of knots. In our simulation studies, we try 4, 6 and 8 interior knots to test the sensitivity of the SBLL estimators and the corresponding SCBs.

For the local-linear smoothing in the backfitting, Condition (B2) requires that the bandwidths are of order $n^{-1/5}$. Any bandwidths with this rate lead to the same limiting distribution for $\widehat{\phi}_\ell^{\text{SBLL}}$, so the user can consider any standard routine for bandwidth selection. There have been many proposals for bandwidth selection in the literature. In our simulation, we consider three popular bandwidth selectors described in [8] and [26]: rule-of-thumb bandwidth ("thumbBw"), plug-in bandwidth selector ("pluginBw") and leave-one-out cross-validation bandwidth selector ("regCVBwSelC"). Below we present simulation results to compare the performance of three bandwidth selectors. The kernel that we use here is the Epanechnikov kernel: $K(u) = 3/4(1 - u^2)\mathbf{1}(|u| \leq 1)$.

To see how the refitting smoothing parameters affect estimation accuracy, we report the average mean squared errors (AMSEs) of the SBLL estimators based on 4, 6 and 8 interior knots in the spline refitting and three different bandwidth selectors in the kernel backfitting. Fig. A.2 presents the AMSEs of the resulting SBLL estimators based on different combinations of the refitting smoothing parameters. For both ϕ_1 and ϕ_2, the AMSEs are very similar across the different combinations of knots and bandwidth selectors.

Fig. A.3 shows the coverage rates of the SCBs based on different combinations of knots and bandwidth selectors. From Fig. A.3, it is clear that the number of knots for a spline in the refitting has very little effect on the coverage of the SCBs. One also observes that the performances of the SCBs based on different smoothing parameters become more similar with increasing sample size, whereas the coverage rates of the SCBs using the "thumbBw" selector are the closest to the nominal level in all the simulation settings. Thus we recommend the "thumbBw" selector, especially when the sample size is small.

Appendix B. Technical details

This section contains some technical assumptions, lemmas and proofs. For any real numbers a and b, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b, respectively. For any two sequences a_n, b_n, $n = 1, 2, \ldots$, we use $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 < a_n/b_n < c_2$ for all n sufficiently large. On any fixed interval $[a, b]$, we denote the space of the second order smooth functions as $C^{(d)}[a, b] = \{f | f^{(d)} \in C[a, b]\}$ and the class of Lipschitz continuous functions for any fixed constant $C > 0$ as $\text{Lip}([a, b], C) = \{f | |f(x) - f(x')| \leq C |x - x'| \text{ for all } x, x' \in [a, b]\}$.

Furthermore, let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ be an n-dimensional vector, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{p_1})$ be an $n \times p_1$ matrix, where $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{nk})^\top$ with $k \in \{1, \ldots, p_1\}$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{p_2})$ be an $n \times p_2$ matrix, where $\mathbf{X}_\ell = (X_{1\ell}, \ldots, X_{n\ell})^\top$, $\ell = 1, \ldots, p_2$. Let $\mathbf{B}^{(d)} = (\mathbf{B}_1^{(d)}, \ldots, \mathbf{B}_{p_2}^{(d)})$ be a dimension $n \times (p_2 M_n)$ matrix, where $\mathbf{B}_\ell^{(d)} = (\mathbf{B}_\ell^{(d)}(X_{1\ell}), \ldots, \mathbf{B}_\ell^{(d)}(X_{n\ell}))^\top$ is a dimension $n \times M_n$ matrix of spline basis functions of order d, for $\ell \in \{1, \ldots, p_2\}$. Let $\mathcal{A} \subseteq \{1, \ldots, p_1 + 2p_2\}$ be an index set, and let $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$.

B.1. Technical assumptions

In addition to the sparsity condition (A1) stated in Section 2, we need the following additional regularity conditions to establish the theoretical results in this paper.

(A2) (*Conditions on errors*) The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with $\text{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2$, $\text{E}|\varepsilon_i|^{2+\delta} \leq M_\delta$ for some positive constant M_δ ($\delta > 0.5$), and have b-sub-Gaussian tails, i.e., $\text{E}\{\exp(t\varepsilon_i)\} \leq \exp(b^2 t^2/2)$, for any $t \geq 0$ and some $b > 0$.

(A3) (*Conditions on nonlinear functions*) The additive component function $g_\ell \in C^{(2)}[a, b]$ for all $\ell \in \{1, \ldots, p_2\}$.

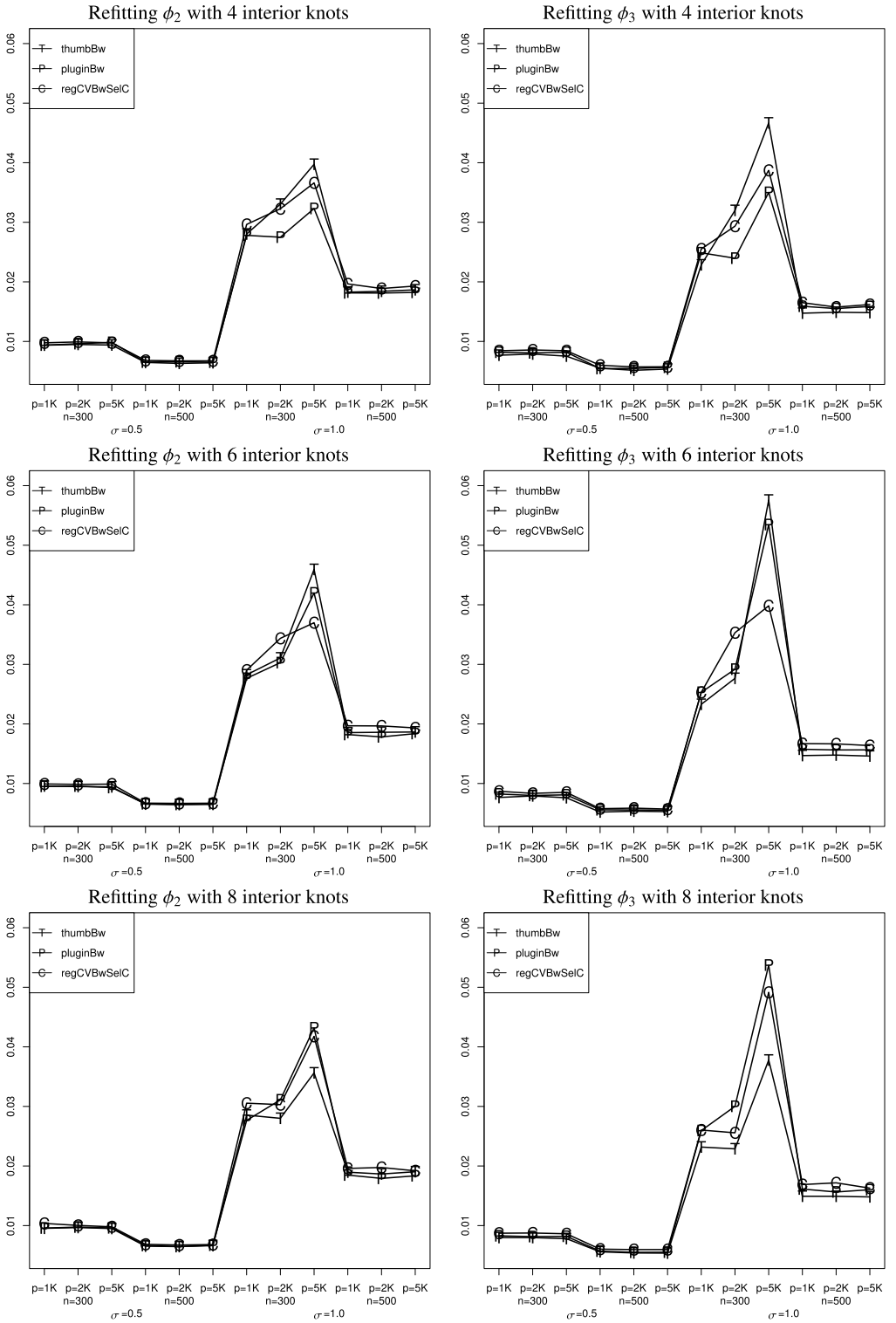

Fig. A.2. Average mean squared errors (AMSEs) of the SBLL estimators of ϕ_2 and ϕ_3.

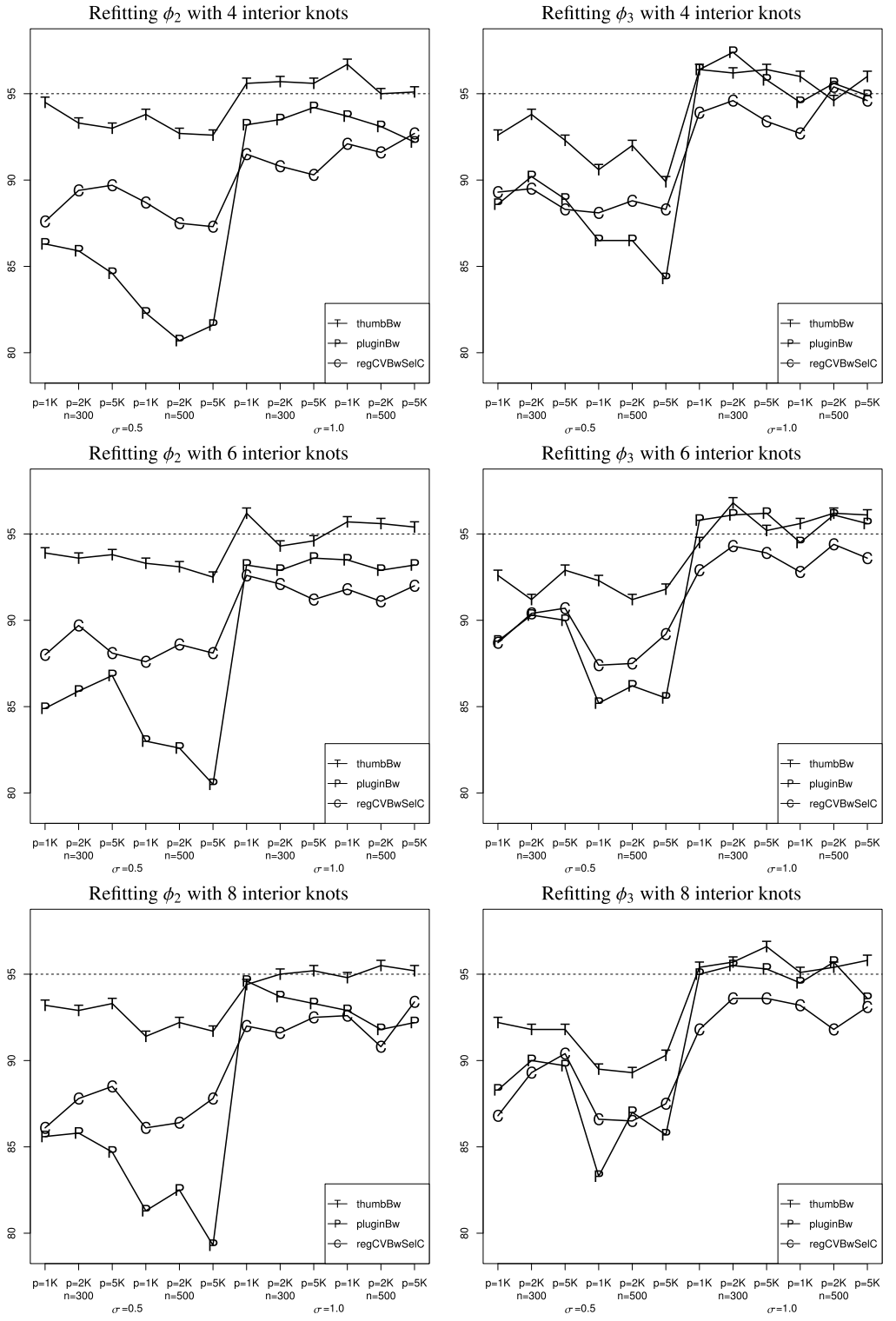

Fig. A.3. Coverage rates of the SCBs for ϕ_2 and ϕ_3.

(A4) (*Conditions on covariates*) Each covariate in the parametric part of the model is bounded, i.e., there is a positive constant C_3 such that $|Z_k| \le C_3$ for all $k \in \{1, \ldots, p_1\}$; also, $\mathrm{E}(X_\ell) = 0$, and there is a positive constant C_4 such that $|X_\ell| \le C_4$, $1 \le \ell \le p_2$. The joint density function of active pure linear $\boldsymbol{X}$ is continuous and bounded below and above. Each covariate in the nonparametric part of the model has a continuous density and there exist constants C_1 and C_2 such that the marginal density function f_ℓ of X_ℓ has continuous derivatives on its support, and satisfies $0 < C_1 \le f_\ell(x_\ell) \le C_2 < \infty$ on its support for every $\ell \in \{1, \ldots, p_2\}$. In addition, the eigenvalues of $\mathrm{E}\{(\boldsymbol{Z}\boldsymbol{Z}^\top)|\boldsymbol{X}\}$ are bounded away from 0.

(A5) (*Conditions on the initial estimators*) The initial estimators satisfy

$$r_{n1} \max_{k \in \mathcal{N}_z} |\widetilde{\alpha}_k| = O_P(1), \quad r_{n2} \max_{\ell \in \mathcal{N}_x} |\widetilde{\beta}_\ell| = O_P(1), \quad r_{n3} \max_{\ell \in \mathcal{N}_x} \|\widetilde{\boldsymbol{\gamma}}_\ell\|_2 = O_P(1)$$

as well as $r_{n1}, r_{n2}, r_{n3} \to \infty$, and there exist positive constants c_{b1}, c_{b2} and c_{b3} such that

$$\Pr\left(\min_{k \in \mathcal{S}_z} |\widetilde{\alpha}_k| \ge c_{b1} b_{n1}\right) \to 1, \quad \Pr\left(\min_{\ell \in \mathcal{S}_{x,L}} |\widetilde{\beta}_\ell| \ge c_{b2} b_{n2}\right) \to 1, \quad \Pr\left(\min_{\ell \in \mathcal{S}_{x,N}} \|\widetilde{\boldsymbol{\gamma}}_\ell\|_2 \ge c_{b3} b_{n3}\right) \to 1,$$

where $b_{n1} = \min_{k \in \mathcal{S}_z} |\alpha_{0k}|$, $b_{n2} = \min_{\ell \in \mathcal{S}_{x,L}} |\beta_{0\ell}|$, and $b_{n3} = \min_{\ell \in \mathcal{S}_{x,N}} \|g_{0\ell}\|_2$.

(A6) (*Conditions on parameters and spline basis functions*) Let p_1 and p_2 be the number of linear and nonlinear components, respectively. Suppose that $N_n/n + \sum_{j=1}^3 \lambda_{nj}^2/n^2 = o(1)$, and

$$\frac{\sqrt{n \ln(p_1)}}{\lambda_{n1} r_{n1}} + \frac{\sqrt{n \ln(p_2)}}{\lambda_{n2} r_{n2}} + \frac{\sqrt{n N_n \ln(p_2 N_n)}}{\lambda_{n3} r_{n3}} + \sum_{j=1}^3 \frac{n}{\lambda_{nj} r_{nj} N_n} = o(1).$$

Assumptions (A1)–(A4) are regularity conditions that are commonly used in the APLM literature. To obtain the selection consistency of the SBLL-AGLASSO, we need an order requirement for a general initial estimator; see Assumption (A5). Theorem B.1 in [16] demonstrates that the group LASSO estimator defined in (6) satisfies Assumption (A5) under some weak conditions, specifically if $\sum_{j=1}^3 \widetilde{\lambda}_{nj}^2 \asymp n\{\ln(p_1) \vee N_n \ln(p_2 N_n)\}$ and $N_n \asymp n^{1/3}$, then the consistent rates for the group LASSO estimator in (A5) have order

$$r_{n1} \asymp r_{n2} \asymp r_{n3} = O\{n^{1/2}/\sqrt{\ln(p_1) \vee N_n \ln(p_2 N_n)}\}.$$

Consequently, Assumption (A6) is equivalent to

$$\frac{\sum_{j=1}^3 \lambda_{nj}^2}{n^2} + \frac{\ln(p_1) \vee N_n \ln(p_2 N_n)}{(\lambda_{n1} \wedge \lambda_{n2} \wedge \lambda_{n3})} + \frac{n^{1/6}\sqrt{\ln(p_1) \vee N_n \ln(p_2 N_n)}}{(\lambda_{n1} \wedge \lambda_{n2} \wedge \lambda_{n3})} = o(1), \tag{B.1}$$

If we take $\lambda_{n1} \asymp \lambda_{n2} \asymp \lambda_{n3} = O(n^{1/2})$, then (B.1) indicates $p_1 = \exp\{o(n^{1/2})\}$ and $p_2 = \exp\{o(n^{1/6})\}$.

We need the following additional assumptions in order to develop the asymptotic SCBs for the nonparametric components.

(A3′) (*Conditions on nonlinear functions*) For any $\ell \in \mathcal{S}_{x,N}$, $\phi_{0\ell} \in C^{(d)}[a, b]$, for some integer $d \ge 2$. In addition, ψ_ℓ^x defined in (11) satisfies $\psi_\ell^x \in C^{(d)}[a, b]$.

(A6′) (*Conditions on spline basis functions*) The order of the spline basis functions is at least d, and the number of interior knots M_n satisfies: $\{n^{1/(2d)} \vee n^{4/(10d-5)}\} \ll M_n \ll n^{1/3}$.

(B1) (*Conditions on the kernel function*) The kernel function $K \in \mathrm{Lip}\,([-1, 1], C_K)$ for some constant $C_K > 0$, and is bounded, nonnegative, symmetric, and supported on $[-1, 1]$ with the second moment $\mu_2(K) = \int u^2 K(u)\, du$.

(B2) (*Conditions on bandwidth*) For each $\ell \in \mathcal{S}_{x,N}$, the bandwidth of the kernel K is $h_\ell^{-1} = O(n^{1/5} \ln^\delta n)$ for some constant $\delta > 1/5$.

Assumptions (A3′), (B1) and (B2) are typical in the local polynomial smoothing literature; see, for instance, [34]. Assumption (A6′) imposes the condition of the number of knots for spline smoothing. For example, if $d = 2$, we can take $M_n \sim n^{4/15} \ln n$.

B.2. Selection and estimation properties of the adaptive group LASSO estimators

In this section, we establish the selection and estimation properties of the adaptive group LASSO estimators as stated in Theorems 1 and 2.

In the following, denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_1})^\top$ with length p_1, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_2})^\top$ with length p_2, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_{p_2}^\top)^\top$ with length $(p_2 N_n)$. Let

$$\boldsymbol{\theta}^\top = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top) = (\alpha_1, \ldots, \alpha_{p_1}, \beta_1, \ldots, \beta_{p_2}, \boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_{p_2}^\top) = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_m^\top, \ldots, \boldsymbol{\theta}_{p_1+2p_2}^\top),$$

where $\boldsymbol{\theta}_m = \alpha_m \mathbf{1}(1 \le m \le p_1) + \beta_{m-p_1} \mathbf{1}(p_1 + 1 \le m \le p_1 + p_2) + \boldsymbol{\gamma}_{m-p_1-p_2} \mathbf{1}(p_1 + p_2 + 1 \le m \le p_1 + 2p_2)$, with $\mathbf{1}$ being an indicator function. Let

$$\mathbf{D} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{p_1}, \mathbf{X}_1, \ldots, \mathbf{X}_{p_2}, \mathbf{B}_1^{(1)}, \ldots, \mathbf{B}_{p_2}^{(1)}) \equiv (\mathbf{D}_1, \ldots, \mathbf{D}_m, \ldots, \mathbf{D}_{p_1+2p_2})$$

be an $n \times (p_1 + p_2 + p_2 N_n)$ matrix, where

$$\mathbf{D}_m = \mathbf{Z}_m \mathbf{1}(1 \le m \le p_1) + \mathbf{X}_m \mathbf{1}(1 \le m \le p_1) + \mathbf{B}^{(1)}_{m-p_1-p_2} \mathbf{1}(p_1 + p_2 + 1 \le m \le p_1 + 2p_2),$$

an $n \times d_m$ submatrix of $\mathbf{D}$ with $d_m = \mathbf{1}(1 \le m \le p_1) + \mathbf{1}(p_1 + 1 \le m \le p_1 + p_2) + N_n \mathbf{1}(p_1 + p_2 + 1 \le m \le p_1 + 2p_2)$.

Next we define the active linear index set for $\boldsymbol{X}$ as $\mathcal{S}_{x,L} = \mathcal{S}_{x,PL} \cup \mathcal{S}_{x,LN}$, the inactive linear index set for $\boldsymbol{X}$ as $\mathcal{N}_{x,L}$, and the inactive nonlinear index set for $\boldsymbol{X}$ as $\mathcal{N}_{x,N}$. Note that $\mathcal{N}_x = \mathcal{N}_{x,L} \cap \mathcal{N}_{x,N}$. Further, let

$$\mathcal{S} = \mathcal{S}_z \cup \{\ell + p_1 : \ell \in \mathcal{S}_{x,L}\} \cup \{\ell + p_1 + p_2 : \ell \in \mathcal{S}_{x,N}\},$$
$$\mathcal{N} = \mathcal{N}_z \cup \{\ell + p_1 : \ell \in \mathcal{N}_{x,L}\} \cup \{\ell + p_1 + p_2 : \ell \in \mathcal{N}_{x,N}\}.$$

For any index set $\mathcal{A} \subseteq \{1, \ldots, p_1 + 2p_2\}$, define $\mathbf{D}_\mathcal{A} = \{\mathbf{D}_m : m \in \mathcal{A}\}$. Next denote $\mathbf{C}_\mathcal{A} = n^{-1}\mathbf{D}_\mathcal{A}^\top \mathbf{D}_\mathcal{A}$, and let $\pi_{\min}(\mathbf{C}_\mathcal{A})$ and $\pi_{\max}(\mathbf{C}_\mathcal{A})$ represent the minimum and maximum eigenvalues of $\mathbf{C}_\mathcal{A}$, respectively.

Lemma B.1. *Let $N_n = O(n^\gamma)$, where $0 < \gamma < 0.5$. Suppose that $|\mathcal{A}|$ is bounded by a fixed constant independent of n, p_1 and p_2. Then under Assumption (A4), with probability approaching 1 as $n \to \infty$, $c_1 \le \pi_{\min}(\mathbf{C}_\mathcal{A}) \le \pi_{\max}(\mathbf{C}_\mathcal{A}) \le c_2$, where c_1 and c_2 are two positive constants.*

Proof. Similar to the proof of Lemma A.1 in [17]. □

Lemma B.2. *Under Assumption (A3), there exists a vector $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{01}^\top, \ldots, \boldsymbol{\gamma}_{0p_2}^\top)^\top$, such that $\|\boldsymbol{\gamma}_{0\ell}\| \ne 0$ for $\ell \in \mathcal{S}_{x,N}$, $\|\boldsymbol{\gamma}_{0\ell}\| = 0$ for $\ell \in \mathcal{N}_{x,N}$ and $\|g_{0\ell} - \mathbf{B}_\ell^{(d)\top} \boldsymbol{\gamma}_{0\ell}\|_2 = O(M_n^{-d})$.*

Proof. Similar to the proof of Lemma A.2 in [17]. □

In the following, we denote $g_{n\ell}(\cdot) = \sum_{J=1}^{N_n} \gamma_{0\ell J} B^{(1)}_{J,\ell}(\cdot)$ the best constant spline approximation of $g_{0\ell}(\cdot)$ such that

$$\|g_{0\ell} - g_{n\ell}\|_\infty = \sup_{x \in [a,b]} |g_{0\ell}(x) - g_{n\ell}(x)| = O(N_n^{-1}).$$

Let $\boldsymbol{\gamma}_{0\ell} = (\gamma_{0\ell J}, J = 1, \ldots, N_n)^\top$ be the vector of the coefficients of the best spline approximation in Lemma B.2. Denote $\boldsymbol{\theta}_0^\top = (\boldsymbol{\theta}_{01}^\top, \ldots, \boldsymbol{\theta}_{0m}^\top, \ldots, \boldsymbol{\theta}_{0,p_1+2p_2}^\top) = (\boldsymbol{\alpha}_0^\top, \boldsymbol{\beta}_0^\top, \boldsymbol{\gamma}_0^\top) = (\alpha_{01}, \ldots, \alpha_{0p_1}, \beta_{01}, \ldots, \beta_{0p_2}, \boldsymbol{\gamma}_{01}^\top, \ldots, \boldsymbol{\gamma}_{0p_2}^\top)$. Define $\boldsymbol{\theta}_\mathcal{A} = (\boldsymbol{\theta}_m^\top : m \in \mathcal{A})^\top$, $\boldsymbol{\theta}_{0,\mathcal{A}} = (\boldsymbol{\theta}_{0m}^\top : m \in \mathcal{A})^\top$.

Proof of Theorem 1. By the Karush–Kuhn–Tucker (KKT) condition [1], if $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ is the unique minimizer of $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}; \lambda_1, \lambda_2, \lambda_3)$, it is equivalent to satisfy

(C1-1) $\mathbf{Z}_k^\top(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'}) = \lambda_{n1} w_k^\alpha \alpha_k / |\alpha_k|$ for any $k \in \mathcal{S}_z$;
(C1-2) $\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'}) = \lambda_{n2} w_\ell^\beta \beta_\ell / |\beta_\ell|$ for any $\ell \in \mathcal{S}_{x,L}$;
(C1-3) $\mathbf{B}_\ell^{(1)\top}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'}) = \lambda_{n3} w_\ell^\gamma \boldsymbol{\gamma}_\ell / \|\boldsymbol{\gamma}_\ell\|$ for any $\ell \in \mathcal{S}_{x,N}$;
(C2) $|\mathbf{Z}_k^\top(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'})| \le \lambda_{n1} w_k^\alpha$ for any $k \in \mathcal{N}_z$;
(C3) $|\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'})| \le \lambda_{n2} w_\ell^\beta$ for any $\ell \in \mathcal{N}_x \cup \mathcal{S}_{x,PN}$;
(C4) $\|\mathbf{B}_\ell^{(1)\top}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} - \sum_{\ell'=1}^{p_2} \mathbf{B}_{\ell'}^{(1)} \boldsymbol{\gamma}_{\ell'})\| \le \lambda_{n3} w_\ell^\gamma$ for any $\ell \in \mathcal{N}_x \cup \mathcal{S}_{x,PL}$.

Define $\overline{\boldsymbol{\theta}}^o = (\mathbf{D}_\mathcal{S}^\top \mathbf{D}_\mathcal{S})^{-1} \mathbf{D}_\mathcal{S}^\top \mathbf{Y}$, a vector with length $|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}| N_n$. Denote three vectors, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$, whose elements are in the form

$$\boldsymbol{v}_{1m} = \frac{\omega_m^\alpha \overline{\theta}_{0m}}{|\overline{\theta}_{0m}|} \mathbf{1}(m \in \mathcal{S}_z) + \mathbf{0}_N \mathbf{1}(m - |\mathcal{S}_z| - |\mathcal{S}_{x,LN}| \in \mathcal{S}_{x,N}),$$
$$\boldsymbol{v}_{2m} = \frac{\omega_{m-|\mathcal{S}_z|}^\beta \overline{\theta}_{0m}}{|\overline{\theta}_{0m}|} \mathbf{1}(m - |\mathcal{S}_z| \in \mathcal{S}_{x,L}) + \mathbf{0}_N \mathbf{1}(m - |\mathcal{S}_z| - |\mathcal{S}_{x,LN}| \in \mathcal{S}_{x,N}),$$
$$\boldsymbol{v}_{3m} = \frac{\omega_{m-|\mathcal{S}_z|-|\mathcal{S}_{x,L}|}^\gamma \overline{\boldsymbol{\theta}}_{0m}}{\|\overline{\boldsymbol{\theta}}_{0m}\|} \mathbf{1}(m - |\mathcal{S}_z| - |\mathcal{S}_{x,LN}| \in \mathcal{S}_{x,N}),$$

for all $m \in \mathcal{S}$.

Next define $\widehat{\boldsymbol{\theta}}^o = (\widehat{\boldsymbol{\theta}}_m^o, 1 \le m \le p_1 + 2p_2)^\top$, where

$$\widehat{\boldsymbol{\theta}}_\mathcal{S}^o \equiv (\widehat{\boldsymbol{\theta}}_m^o, m \in \mathcal{S})^\top = (\mathbf{D}_\mathcal{S}^\top \mathbf{D}_\mathcal{S})^{-1} \left(\mathbf{D}_\mathcal{S}^\top \mathbf{Y} - \sum_{j=1}^{3} \lambda_{nj} \boldsymbol{v}_j \right),$$

$\widehat{\theta}_m^o = 0$ for $m \in \mathcal{N}_z$ and $m - p_1 \in \mathcal{N}_{x,L}$, and $\widehat{\boldsymbol{\theta}}_m^o = \mathbf{0}_N$ for $m - p_1 - p_2 \in \mathcal{N}_{x,N}$. So we can represent

$$\widehat{\boldsymbol{\theta}}^o \equiv (\widehat{\boldsymbol{\theta}}_{\mathcal{S}_z}^{o\top}, \widehat{\boldsymbol{\theta}}_{\mathcal{N}_z}^{o\top}, \widehat{\boldsymbol{\theta}}_{\mathcal{S}_{x,L}}^{o\top}, \widehat{\boldsymbol{\theta}}_{\mathcal{N}_{x,L}}^{o\top}, \widehat{\boldsymbol{\theta}}_{\mathcal{S}_{x,N}}^{o\top}, \widehat{\boldsymbol{\theta}}_{\mathcal{N}_{x,N}}^{o\top})^\top,$$

and $\widehat{\boldsymbol{\theta}}^o_{\mathcal{S}} \equiv (\widehat{\boldsymbol{\theta}}^{o\top}_{\mathcal{S}_z}, \widehat{\boldsymbol{\theta}}^{o\top}_{\mathcal{S}_{x,L}}, \widehat{\boldsymbol{\theta}}^{o\top}_{\mathcal{S}_{x,N}})^\top$. Denote

$$\widehat{\mathcal{S}}^o = \{1 \le m \le p_1 + 2p_2 : \|\widehat{\boldsymbol{\theta}}^o_m\| > 0\}.$$

Apparently, $\widehat{\mathcal{S}}^o \subseteq \mathcal{S}$. Notice that $\mathbf{D}\widehat{\boldsymbol{\theta}}^o = \mathbf{D}_{\mathcal{S}}\widehat{\boldsymbol{\theta}}^o_{\mathcal{S}}$ and $\{\mathbf{D}_m,\ m \in \mathcal{S}\}$ are linearly independent, so by the definition of $\widehat{\boldsymbol{\theta}}^o$, (C1-1), (C1-2) and (C1-3) hold for $\widehat{\boldsymbol{\theta}}^o$ if $\widehat{\mathcal{S}}^o \supseteq \mathcal{S}$. Therefore, if $\widehat{\boldsymbol{\theta}}^o$ satisfies

(C1′) $\widehat{\mathcal{S}}^o \supseteq \mathcal{S}$,
(C2′) $|\mathbf{Z}_k^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)| \le \lambda_{n1}\omega_k^\alpha$, for any $k \in \mathcal{N}_z$,
(C3′) $\|\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| \le \lambda_{n2}\omega_\ell^\beta$, for any $\ell \in \mathcal{N}_{x,L}$,
(C4′) $\|\mathbf{B}_\ell^{(1)\top}(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| \le \lambda_{n3}\omega_\ell^\gamma$, for any $\ell \in \mathcal{N}_{x,N}$,

then $\widehat{\boldsymbol{\theta}}^o$ is the unique minimizer of $L_n(\boldsymbol{\theta}; \lambda_{n1}, \lambda_{n2}, \lambda_{n3})$, in other words, $\widehat{\boldsymbol{\theta}}^o = \widehat{\boldsymbol{\theta}}$ with probability approaching 1. Therefore, in order to show $\Pr(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$, it is equivalent to show $\widehat{\boldsymbol{\theta}}^o$ satisfies (C1′)–(C3′) with probability approaching 1, as $n \to \infty$. Further notice that

(C1″) $\forall_{m\in\mathcal{S}}\ \ \|\boldsymbol{\theta}_{0m}\| - \|\widehat{\boldsymbol{\theta}}^o_m\| < \|\boldsymbol{\theta}_{0m}\|$,

implies Condition (C1′). Therefore, to show $\widehat{\boldsymbol{\theta}}^o$ is the unique minimizer of $L_n(\boldsymbol{\theta}; \lambda_{n1}, \lambda_{n2}, \lambda_{n3})$, and consequently, $\Pr(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$, it suffices to show that $\widehat{\boldsymbol{\theta}}^o$ satisfies Conditions (C1″), (C2′) and (C3′) with probability approaching 1, as $n \to \infty$.

According to Lemmas B.3 and B.4, we obtain that, as $n \to \infty$,

$$\begin{aligned}\Pr(\widehat{\mathcal{S}} \ne \mathcal{S}) \le \Pr(\exists_{m\in\mathcal{S}}\ \ \|\boldsymbol{\theta}_{0m} - \widehat{\boldsymbol{\theta}}^o_m\| \ge \|\boldsymbol{\theta}_{0m}\|,) + \Pr(\exists_{k\in\mathcal{N}_z};\ |\mathbf{Z}_k^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)| > \lambda_{n1}\omega_k^\alpha)\\ + \Pr(\exists_{\ell\in\mathcal{N}_{x,L}}\ \|\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| > \lambda_{n2}\omega_\ell^\beta) + \Pr(\exists_{\ell\in\mathcal{N}_{x,N}}\ \|\mathbf{B}_\ell^{(1)\top}(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| > \lambda_{n3}\omega_\ell^\gamma) \to 0.\end{aligned}$$

This completes the proof of Theorem 1. □

The following Lemmas B.3 and B.4 are used in the proof of Theorem 1. The proofs are given in [16].

Lemma B.3. *Under Assumptions (A3)–(A6), as $n \to \infty$, $\Pr(\exists_{m\in\mathcal{S}}\ \|\boldsymbol{\theta}_{0m} - \widehat{\boldsymbol{\theta}}^o_m\| \ge \|\boldsymbol{\theta}_{0m}\|) \to 0$.*

Lemma B.4. *Under Assumptions (A3)–(A6), as $n \to \infty$,*

$$\Pr\{\exists_{k\in\mathcal{N}_z}\ |\mathbf{Z}_k^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)| > \lambda_{n1}\omega_k^\alpha\} \to 0, \quad \Pr\{\exists_{\ell\in\mathcal{N}_{x,L}}\ \|\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| > \lambda_{n2}\omega_\ell^\beta\} \to 0,$$
$$\Pr\{\exists_{\ell\in\mathcal{N}_{x,N}}\ \|\mathbf{B}_\ell^{(1)\top}(\mathbf{Y} - \mathbf{D}\widehat{\boldsymbol{\theta}}^o)\| > \lambda_{n3}\omega_\ell^\gamma\} \to 0.$$

Proof of Theorem 2. Let π_1 and π_2 be the minimum and maximum eigenvalues of $\mathbf{C}_{\mathcal{S}}$, respectively, and let $\pi_3 = \max_{m\notin\mathcal{S}} \|n^{-1}\mathbf{D}_m^\top\mathbf{D}_m\|$. By Lemma B.1, $\pi_1 \asymp 1$, $\pi_2 \asymp 1$ and $\pi_3 \asymp 1$. For any $\ell \in \mathcal{S}_{x,N}$, let $g_{0\ell}(\mathbf{X}_\ell) = (g_{0\ell}(X_{1\ell}), \ldots, g_{0\ell}(X_{n\ell}))^\top$, $\boldsymbol{\delta}_\ell = g_{0\ell}(\mathbf{X}_\ell) - \mathbf{B}_\ell^{(1)}\boldsymbol{\gamma}_\ell$ and $\boldsymbol{\delta} = \sum_{\ell\in\mathcal{S}_{x,N}}\boldsymbol{\delta}_\ell$. According to the proof of Theorem 1, with probability approaching 1, we have

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\mathcal{S}} = \widehat{\boldsymbol{\theta}}^o_{\mathcal{S}} &= (\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1}\left(\mathbf{D}_{\mathcal{S}}^\top\mathbf{Y} - \sum_{j=1}^3 \lambda_{nj}\boldsymbol{v}_j\right) = (\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1} \times\\ &\left[\mathbf{D}_{\mathcal{S}}^\top\left\{\mathbf{Z}_{\mathcal{S}_z}\boldsymbol{\alpha}_{0,\mathcal{S}_z} + \mathbf{X}_{\mathcal{S}_{x,L}}\boldsymbol{\beta}_{0,\mathcal{S}_{x,L}} + \sum_{\ell\in\mathcal{S}_{x,N}}\left(\mathbf{B}_\ell^{(1)}\boldsymbol{\gamma}_{0\ell} + \boldsymbol{\delta}_\ell\right) + \boldsymbol{\varepsilon}\right\} - \sum_{j=1}^3\lambda_{nj}\boldsymbol{v}_j\right]\\ &= \boldsymbol{\theta}_{0,\mathcal{S}} + (\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1}\left\{\mathbf{D}_{\mathcal{S}}^\top(\boldsymbol{\delta} + \boldsymbol{\varepsilon}) - \sum_{j=1}^3\lambda_{nj}\boldsymbol{v}_j\right\}.\end{aligned}$$

Let $\mathbf{C}_{\mathcal{S}} = n^{-1}\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}}$ be an $(|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}|N_n) \times (|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}|N_n)$ matrix, and let $\mathbf{H} = \mathbf{I} - \mathbf{D}_{\mathcal{S}}(\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1}\mathbf{D}_{\mathcal{S}}^\top$ be an $n \times n$ matrix. Then

$$\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}^o_{\mathcal{S}} = n^{-1}\mathbf{C}_{\mathcal{S}}^{-1}\left\{\mathbf{D}_{\mathcal{S}}^\top(\boldsymbol{\delta} + \boldsymbol{\varepsilon}) - \sum_{j=1}^3\lambda_{nj}\boldsymbol{v}_j\right\}. \tag{B.2}$$

For $\boldsymbol{\eta} = \mathbf{Y} - \mathbf{D}\boldsymbol{\theta}$, define $\boldsymbol{\eta}_*$ as the projection of $\boldsymbol{\eta}$ to the column space of $\mathbf{D}_{\mathcal{S}}$, i.e., $\boldsymbol{\eta}_* \equiv \mathbf{P}_{\mathbf{D}_{\mathcal{S}}}\boldsymbol{\eta} = \mathbf{D}_{\mathcal{S}}(\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1}\mathbf{D}_{\mathcal{S}}^\top\boldsymbol{\eta}$. Then for $\boldsymbol{\varepsilon}_* \equiv \mathbf{P}_{\mathbf{D}_{\mathcal{S}}}\boldsymbol{\varepsilon}$, and by Lemma B.1,

$$\|\boldsymbol{\varepsilon}_*\|^2 = \|(\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}})^{-1/2}\mathbf{D}_{\mathcal{S}}^\top\boldsymbol{\varepsilon}\|^2 \le (n\pi_1)^{-1}\|\mathbf{D}_{\mathcal{S}}^\top\boldsymbol{\varepsilon}\|^2 = O_P\{\pi_1^{-1}(|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}|N_n)\},$$

$$\|\boldsymbol{\eta}_*\|^2 \le 2\|\boldsymbol{\varepsilon}_*\|^2 + O_P(n|\mathcal{S}_{x,N}|N_n^{-2}) = O_P\{\pi_1^{-1}(|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}|N_n)\} + O_P(nN_n^{-2}),$$

and

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}\|^2 &\le 8\|\boldsymbol{\eta}_*\|^2/(n\pi_1) + 4\{\lambda_{n1}^2|\mathcal{S}_z| + \lambda_{n2}^2|\mathcal{S}_{x,L}| + \lambda_{n3}^2|\mathcal{S}_{x,N}|\}/(n^2\pi_1^2)\\ &= O_P\left\{\frac{|\mathcal{S}_z| + |\mathcal{S}_{x,L}| + |\mathcal{S}_{x,N}|N_n}{n\pi_1^2}\right\} + O\left(\frac{|\mathcal{S}_{x,N}|}{\pi_1 N_n^2}\right) + O_P\left\{\frac{\lambda_{n1}^2|\mathcal{S}_z| + \lambda_{n2}^2|\mathcal{S}_{x,L}| + \lambda_{n3}^2|\mathcal{S}_{x,N}|}{n^2\pi_1^2}\right\}.\end{aligned}$$

Therefore, the results follow by the facts that

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z} - \boldsymbol{\alpha}_{0,\mathcal{S}_z} &= \left(\mathbf{I}_{|\mathcal{S}_z|}\ \ \mathbf{0}_{|\mathcal{S}_z|\times|\mathcal{S}_{x,L}|}\ \ \mathbf{0}_{|\mathcal{S}_z|\times(|\mathcal{S}_{x,N}|N_n)}\right)(\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}),\\ \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,L}} - \boldsymbol{\beta}_{0,\mathcal{S}_{x,L}} &= \left(\mathbf{0}_{|\mathcal{S}_{x,L}|\times|\mathcal{S}_z|}\ \ \mathbf{I}_{|\mathcal{S}_{x,L}|}\ \ \mathbf{0}_{|\mathcal{S}_{x,L}|\times(|\mathcal{S}_{x,N}|N_n)}\right)(\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}),\\ \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}} - \boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}} &= \left(\mathbf{0}_{(|\mathcal{S}_{x,N}|N_n)\times|\mathcal{S}_z|}\ \ \mathbf{0}_{(|\mathcal{S}_{x,N}|N_n)\times|\mathcal{S}_{x,L}|}\ \ \mathbf{I}_{|\mathcal{S}_{x,N}|N_n}\right)(\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}),\end{aligned}$$

and $\|\widehat{g}_\ell - g_{n\ell}\|_2^2 \asymp \|\widehat{\boldsymbol{\gamma}}_\ell - \boldsymbol{\gamma}_{0\ell}\|^2$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_z} = (\widehat{\beta}_k, k \in \mathcal{S}_z)^\top$, $\boldsymbol{\beta}_{0,\mathcal{S}_z} = (\beta_{0k}, k \in \mathcal{S}_z)^\top$, $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}} = (\widehat{\boldsymbol{\gamma}}_\ell, \ell \in \mathcal{S}_{x,N})^\top$ and $\boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}} = (\boldsymbol{\gamma}_\ell, \ell \in \mathcal{S}_{x,N})^\top$. This completes the proof of Theorem 2. □

B.3. Proof of Theorem 4

In this section, the spline basis functions considered are of order d. For any index set $\mathcal{A} \subseteq \{1, \ldots, p_1 + p_2\}$, denote

$$\boldsymbol{\beta}_{\mathcal{A}} = (\beta_k, 1 \le k \le p_1, k \in \mathcal{A})^\top, \quad \widehat{\boldsymbol{\beta}}_{\mathcal{A}} = (\widehat{\beta}_k, 1 \le k \le p_1, k \in \mathcal{A})^\top, \quad \boldsymbol{\gamma}_{\mathcal{A}} = (\boldsymbol{\gamma}_\ell, 1 \le \ell \le p_2, \ell + p_1 \in \mathcal{A})^\top$$

and $\widehat{\boldsymbol{\gamma}}_{\mathcal{A}} = (\widehat{\boldsymbol{\gamma}}_\ell, 1 \le \ell \le p_2, \ell + p_1 \in \mathcal{A})^\top$. Next, denote $\mathbf{Z}_{\mathcal{A}} = (\mathbf{Z}_{i,\mathcal{A}}^\top, i = 1, \ldots, n)^\top$, where

$$\mathbf{Z}_{i,\mathcal{A}} = (Z_{ik}, 1 \le k \le p_1, k \in \mathcal{A})^\top, \quad \mathbf{X}_{i,\mathcal{A}} = (X_{i\ell}, 1 \le \ell \le p_2, \ell + p_1 \in \mathcal{A})^\top.$$

Similarly, denote $\mathbf{B}_{\mathcal{A}}^{(d)} = (\mathbf{B}_{i,\mathcal{A}}^{(d)\top}, i = 1, \ldots, n)^\top$, where $\mathbf{B}_{i,\mathcal{A}}^{(d)} = (B_{J,\ell}^{(d)}(X_{i\ell}), 1 \le \ell \le p_2, \ell + p_1 + p_2 \in \mathcal{A}, J = 1, \ldots, N_n)^\top$. Define $\mathbf{T}_{\mathcal{S}} = (\mathbf{Z}_{\mathcal{S}_z}, \mathbf{X}_{\mathcal{S}_{x,PL}})$. By an abuse of notation, let $\mathbf{D}_{\mathcal{S}} = (\mathbf{T}_{\mathcal{S}}, \mathbf{B}_{\mathcal{S}}^{(d)})$, and we define

$$\mathbf{C}_{\mathcal{S}} = n^{-1}\mathbf{D}_{\mathcal{S}}^\top\mathbf{D}_{\mathcal{S}} = \begin{pmatrix} n^{-1}\mathbf{T}_{\mathcal{S}}^\top\mathbf{T}_{\mathcal{S}} & n^{-1}\mathbf{T}_{\mathcal{S}}^\top\mathbf{B}_{\mathcal{S}}^{(d)} \\ n^{-1}\mathbf{B}_{\mathcal{S}}^{(d)\top}\mathbf{T}_{\mathcal{S}} & n^{-1}\mathbf{B}_{\mathcal{S}}^{(d)\top}\mathbf{B}_{\mathcal{S}}^{(d)} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}, \tag{B.3}$$

$$\mathbf{U}_{\mathcal{S}} = \mathbf{C}_{\mathcal{S}}^{-1} = \begin{pmatrix} \mathbf{U}_{11} & -\mathbf{U}_{11}\mathbf{C}_{12}\mathbf{C}_{22}^{-1} \\ -\mathbf{U}_{22}\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{U}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix}, \tag{B.4}$$

where $\mathbf{U}_{11}^{-1} = \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21} = n^{-1}\mathbf{T}_{\mathcal{S}}^\top(\mathbf{I}_n - \mathbf{P}_{\mathbf{B}_{\mathcal{S}}^{(d)}})\mathbf{T}_{\mathcal{S}}$ and $\mathbf{U}_{22}^{-1} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} = n^{-1}\mathbf{B}_{\mathcal{S}}^{(d)\top}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{T}_{\mathcal{S}}}\right)\mathbf{B}_{\mathcal{S}}^{(d)}$, with $\mathbf{P}_{\mathbf{B}_{\mathcal{S}}^{(d)}}$ and $\mathbf{P}_{\mathbf{T}_{\mathcal{S}}}$ being projection matrices for $\mathbf{B}_{\mathcal{S}}^{(d)}$ and $\mathbf{T}_{\mathcal{S}}$, respectively.

In the following, we give the proof of Theorem 4.

Proof of Theorem 4. The structure of the proof is consisted of two parts: (i) we show the oracle efficiency of $\widehat{\phi}_\ell^{\mathrm{SBLL}}$; (ii) we show the uniform asymptotic normality for the "oracle" estimator $\widehat{\phi}_\ell^o$.

For the first part, note that, for $\ell \in \mathcal{S}_{x,N}$,

$$\widehat{\phi}_\ell^{\mathrm{SBLL}}(x_\ell) - \widehat{\phi}_\ell^o(x_\ell) = (1,\ 0)(\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{X}_\ell^*)^{-1}\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell(\widehat{\mathbf{Y}}_\ell - \mathbf{Y}_\ell), \quad \text{where}$$

$$\begin{aligned}\widehat{\mathbf{Y}}_\ell - \mathbf{Y}_\ell &= \mathbf{Z}_{\mathcal{S}_z}(\boldsymbol{\alpha}_{0,\mathcal{S}_z} - \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^*) + \mathbf{X}_{\mathcal{S}_{x,PL}}(\boldsymbol{\beta}_{0,\mathcal{S}_{x,PL}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^*) + \sum_{\ell' \in \mathcal{S}_{x,N}\setminus\{\ell\}}\{\phi_{0\ell'}(\mathbf{X}_{\ell'}) - \widehat{\phi}_{\ell'}^*(\mathbf{X}_{\ell'})\}\\ &= \mathbf{Z}_{\mathcal{S}_z}(\boldsymbol{\alpha}_{0,\mathcal{S}_z} - \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^*) + \mathbf{X}_{\mathcal{S}_{x,PL}}(\boldsymbol{\beta}_{0,\mathcal{S}_{x,PL}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^*)\\ &\quad + \mathbf{B}_{\mathcal{S}_{x,N}\setminus\{\ell\}}^{(d)}(\boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}\setminus\{\ell\}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}\setminus\{\ell\}}^*) + \sum_{\ell' \in \mathcal{S}_{x,N}\setminus\{\ell\}}\{\phi_{0\ell'}(\mathbf{X}_{\ell'}) - \phi_{n\ell'}(\mathbf{X}_{\ell'})\},\end{aligned}$$

and

$$\begin{aligned}&\mathrm{diag}(1, h_\ell^{-1})\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{X}_\ell^*\,\mathrm{diag}(1, h_\ell^{-1})\\ &= n^{-1}\begin{pmatrix} \sum_{i=1}^n K_{h_\ell}(X_{i\ell} - x_\ell) & \sum_{i=1}^n\left(\frac{X_{i\ell} - x_\ell}{h_\ell}\right)K_{h_\ell}(X_{i\ell} - x_\ell) \\ \sum_{i=1}^n\left(\frac{X_{i\ell} - x_\ell}{h_\ell}\right)K_{h_\ell}(X_{i\ell} - x_\ell) & \sum_{i=1}^n\left(\frac{X_{i\ell} - x_\ell}{h_\ell}\right)^2 K_{h_\ell}(X_{i\ell} - x_\ell) \end{pmatrix} = f_\ell(x_\ell)\begin{pmatrix} 1 & 0 \\ 0 & \mu_2(K) \end{pmatrix} + u_P(1),\end{aligned}$$

with $u_P(\cdot) = o_P(\cdot)$ uniformly for all $x_\ell \in [a, b]$. So

$$\left(\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{X}_\ell^*\right)^{-1} = \mathrm{diag}(1, h_\ell^{-1}) f_\ell^{-1}(x_\ell)\left\{\begin{pmatrix}1 & 0\\ 0 & \mu_2(K)\end{pmatrix} + u_P(1)\right\}\mathrm{diag}(1, h_\ell^{-1}),$$

$$\mathrm{diag}(1, h_\ell^{-1})\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell = \frac{1}{n}\times\begin{pmatrix} K_{h_\ell}(X_{1\ell}-x_\ell) & \dots, & K_{h_\ell}(X_{n\ell}-x_\ell)\\ \left(\frac{X_{1\ell}-x_\ell}{h_\ell}\right)K_{h_\ell}(X_{1\ell}-x_\ell) & \dots, & \left(\frac{X_{n\ell}-x_\ell}{h_\ell}\right)K_{h_\ell}(X_{n\ell}-x_\ell)\end{pmatrix}.$$

Thus,

$$\begin{aligned}\widehat{\phi}_\ell^{\mathrm{SBLL}}(x_\ell) - \widehat{\phi}_\ell^{o}(x_\ell) = f_\ell^{-1}(x_\ell)\Bigg[&\frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{Z}_{i,\mathcal{S}_z}^\top(\boldsymbol{\alpha}_{0,\mathcal{S}_z} - \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^*) + \frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{X}_{i,\mathcal{S}_{x,PL}}^\top(\boldsymbol{\beta}_{0,\mathcal{S}_{x,PL}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^*)\\ &+\frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}_{x,N}\setminus\{\ell\}}^{(d)\top}(\boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}\setminus\{\ell\}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}\setminus\{\ell\}}^*)\\ &+\frac{1}{n}\sum_{i=1}^n\sum_{\ell'\in\mathcal{S}_{x,N}\setminus\{\ell\}} K_{h_\ell}(X_{i\ell}-x_\ell)\{\phi_{0\ell'}(X_{i\ell'}) - \phi_{n\ell'}(X_{i\ell'})\} + u_P(1)\Bigg].\end{aligned} \tag{B.5}$$

For the first and second summation terms on the right-hand side of (B.5), by Theorem 3, we have

$$n^{-1}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{Z}_{i,\mathcal{S}_z}^\top(\boldsymbol{\alpha}_{0,\mathcal{S}_z} - \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_z}^*) = u_P(n^{-1/2}),\quad n^{-1}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{X}_{i,\mathcal{S}_{x,PL}}^\top(\boldsymbol{\beta}_{0,\mathcal{S}_{x,PL}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}_{x,PL}}^*) = u_P(n^{-1/2}).$$

By Lemma B.2,

$$\frac{1}{n}\sum_{i=1}^n\sum_{\ell'\in\mathcal{S}_{x,N}\setminus\{\ell\}} K_{h_\ell}(X_{i\ell}-x_\ell)\{\phi_{0\ell'}(X_{i\ell'}) - \phi_{n\ell'}(X_{i\ell'})\} = u_P\{|\mathcal{S}_{x,N}|M_n^{-d}\}.$$

As for the third terms, define $\zeta_{i\ell} = \phi_{0\ell}(X_{i\ell}) - \sum_{J=1}^{M_n}\gamma_{0J,l}^* B_{J,\ell}^{(d)}(X_{i\ell})$, $\zeta_i = \sum_{\ell\in\mathcal{S}_{x,N}}\zeta_{i\ell}$, and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)^\top$, similar to the induction with (B.2), we have

$$\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^* - \boldsymbol{\theta}_{\mathcal{S}}^o = n^{-1}\mathbf{C}_{\mathcal{S}}^{-1}\{\mathbf{D}_{\mathcal{S}}^\top(\boldsymbol{\zeta}+\boldsymbol{\epsilon})\},$$

then $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}}^* - \boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}} = (\mathbf{0}_{\{|\mathcal{S}_{x,N}|M_n\}\times(|\mathcal{S}_z|+|\mathcal{S}_{x,L}|)}\mathbf{I}_{(|\mathcal{S}_{x,N}|M_n)})\mathbf{C}_{\mathcal{S}}^{-1}n^{-1}\{\mathbf{D}_{\mathcal{S}}^\top(\boldsymbol{\zeta}+\boldsymbol{\epsilon})\}$. Define a diagonal matrix $\mathbf{I}_\ell^0 = \mathrm{diag}\{\mathbf{1}_{(\ell-1)M_n}, \mathbf{0}_{M_n}, \mathbf{1}_{(|\mathcal{S}_{x,N}|-\ell)M_n}\}$ for $\ell \in \mathcal{S}_{x,N}$. Then

$$\mathbf{B}_{i,\mathcal{S}_{x,N}\setminus\{\ell\}}^{(d)\top}(\boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}\setminus\{\ell\}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}\setminus\{\ell\}}^*) = \mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}}^* - \boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}}).$$

Next by Lemma B.2, (B.3) and (B.4), for any $\ell \in \mathcal{S}_{x,N}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}_{x,N}}^* - \boldsymbol{\gamma}_{0,\mathcal{S}_{x,N}}) &= \frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0\mathbf{U}_{22}(-\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\ \ \mathbf{I}_{(|\mathcal{S}_{x,N}|M_n)})\frac{1}{n}\mathbf{B}_{\mathcal{S}}^{(d)\top}(\boldsymbol{\zeta}+\boldsymbol{\epsilon})\\ &= \frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0\mathbf{U}_{22}\frac{1}{n}\mathbf{B}_{\mathcal{S}}^{(d)\top}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{T}_{\mathcal{S}}}\right)(\boldsymbol{\zeta}+\boldsymbol{\epsilon}).\end{aligned}$$

Following the same idea as in the proof of Lemma B.1, we have that there exist constants $0 < c_{U_2} < C_{U_2} < \infty$, such that with probability approaching 1, $c_{U_2}\mathbf{I}_{|\mathcal{S}_{x,N}|M_n} \le \mathbf{U}_{22} \le C_{U_2}\mathbf{I}_{|\mathcal{S}_{x,N}|M_n}$. Similar to Lemma A.4 in [30], for any $\ell, \ell' \in \mathcal{S}_{x,N}$ and $\ell \ne \ell'$, we have

$$\sup_{x_\ell\in\chi_{h_\ell}}\max_{1\le J\le M_n}\left|\frac{1}{n}\sum_{i=1}^n\left[K_{h_\ell}(X_{i\ell}-x_\ell)B_{J\ell'}^{(d)}(X_{i\ell'}) - E\{K_{h_\ell}(X_{i\ell}-x_\ell)B_{J\ell'}^{(d)}(X_{i\ell'})\}\right]\right| = O_P\{\sqrt{\ln n/(nh_\ell)}\},$$

$$\sup_{x_\ell\in\chi_{h_\ell}}\max_{1\le J\le M_n}\left|\frac{1}{n}\sum_{i=1}^n E\{K_{h_\ell}(X_{i\ell}-x_\ell)B_{J\ell}^{(d)}(X_{i\ell})\}\right| = O_P(M_n^{-1/2}).$$

We can show that

$$\sup_{x_\ell\in\chi_{h_\ell}}\frac{1}{n}\sum_{i=1}^n K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0\mathbf{U}_{22}\frac{1}{n}\mathbf{B}_{\mathcal{S}}^{(d)\top}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{T}_{\mathcal{S}}}\right)\boldsymbol{\delta} = O_P\{M_n^{-d+1}(\sqrt{\ln n/(nh_\ell)} + M_n^{-1/2})\},$$

and by Proposition 2 in [27],

$$\sup_{x_\ell\in\chi_{h_\ell}}\frac{1}{n}\sum_{i=1}^{n}K_{h_\ell}(X_{i\ell}-x_\ell)\mathbf{B}_{i,\mathcal{S}}^{(d)\top}\mathbf{I}_\ell^0\mathbf{U}_{22}\frac{1}{n}\mathbf{B}_{\mathcal{S}}^{(d)\top}\left(\mathbf{I}_n-\mathbf{P}_{\mathbf{T}_\mathcal{S}}\right)\boldsymbol{\epsilon}=O_P\{\sqrt{\ln(n)/n}\}.$$

Therefore,

$$\sup_{x_\ell\in\chi_{h_\ell}}|\widehat{\phi}_\ell^{\mathrm{SBLL}}(x_\ell)-\widehat{\phi}_\ell^{o}(x_\ell)|=O_P[\sqrt{\ln n/n}+M_n^{-d+1}\{\sqrt{\ln n/(nh_\ell)}+\sqrt{1/M_n}\}]$$

For part (ii), below we show that for any t and $\ell\in\mathcal{S}_{\chi,N}$,

$$\lim_{n\to\infty}\Pr\left[\sqrt{\ln(h_\ell^{-2})}\left\{\sup_{x_\ell\in\chi_{h_\ell}}\frac{\sqrt{nh_\ell}}{v_\ell(x_\ell)}|\widehat{\phi}_\ell^{o}(x_\ell)-\phi_{0\ell}(x_\ell)|-\tau_n\right\}<t\right]=\exp(-2e^{-t}),$$

where $v_\ell^2(x_\ell)=\|K\|_2^2f_\ell^{-1}(x_\ell)\sigma^2$, $\tau_n=\sqrt{\ln(h_\ell^{-2})}+\ln\{\|K'\|_2/(2\pi\|K\|_2)\}/\sqrt{\ln(h_\ell^{-2})}$.

Define $M_h(x)=h_\ell^{-1/2}\int K\{(x'-x)/h_\ell\}dW(x')$, where $W(x)$ is a Wiener process defined on $(0,\infty)$. By Lemma 1 in [34], one has

$$\lim_{n\to\infty}\Pr\left[\sqrt{\ln(h_\ell^{-2})}\left\{\sup_{x\in\chi_{h_\ell}}|M_{h_\ell}(x)|/\|K\|_{L_2}^2-\tau_n\right\}<t\right]=\exp(-2e^{-t}).\tag{B.6}$$

Recall the definition of $\widehat{\phi}_\ell^{o}(x_\ell)$ in (15), we have

$$\widehat{\phi}_\ell^{o}(x_\ell)-\phi_{0\ell}(x_\ell)=(1\ \ 0)(\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{X}_\ell^{*})^{-1}\mathbf{X}_\ell^{*\top}\mathbf{W}_\ell\mathbf{Y}_\ell-\phi_{0\ell}(x_\ell)=f_\ell^{-1}(x_\ell)\frac{1}{n}\sum_{i=1}^{n}K_{h_\ell}(X_{i\ell}-x_\ell)\varepsilon_i+O_P(h_\ell^2).$$

According to the proof of Theorem 1 in [34], we have

$$\sup_{x_\ell\in\chi_{h_\ell}}\left|\frac{\sqrt{nh_\ell}}{v_\ell(x_\ell)}\{\widehat{\phi}_\ell^{o}(x_\ell)-\phi_{0\ell}(x_\ell)\}-M_{h_\ell}(x_\ell)/\|K\|_{L_2}^2\right|=o_P(\ln^{-1/2}n).$$

Consequently, we have

$$\sup_{x_\ell\in\chi_{h_\ell}}\sqrt{\ln(h_\ell^{-2})}\left|\frac{\sqrt{nh_\ell}}{v_\ell(x_\ell)}\{\widehat{\phi}_\ell^{o}(x_\ell)-\phi_{0\ell}(x_\ell)\}-M_{h_\ell}(x_\ell)/\|K\|_{L_2}^2\right|=o_P(1),$$

as $\sqrt{\ln(h_\ell^{-2})}/\sqrt{\ln(n)}=O(1)$. The uniformly asymptotic normality of the "oracle" estimator $\widehat{\phi}_\ell^{o}(x_\ell)$ follows from (B.6) and Slutsky's Theorem.

Hence, the result in (16) is established. Consequently, the result in (17) follows from (14), and the result in (18) follows from [6]. This completes the proof of Theorem 4. □

B.4. Technical lemmas

The following lemmas are used in the proof of Theorem 1. The proofs are given in [16].

For any random variable X, denote $\|X\|_p=(\mathrm{E}|X|^p)^{1/p}$ as the L_p norm for random variable X; and denote $\|X\|_\varphi=\inf\{C>0:\mathrm{E}\{\varphi(|X|/C)\}\le 1\}$ as the Orlicz norm for random variable X, where φ is required as a non-decreasing, convex function with $\varphi(0)=0$.

Lemma B.5. *Suppose that Assumptions (A2) and (A4) hold. Let, for all $k\in\{1,\dots,p_1\}$, $\ell\in\{1,\dots,p_2\}$, and $J\in\{1,\dots,N_n\}$,*

$$T_{1k}=n^{-1/2}\sum_{i=1}^{n}Z_{ik}\varepsilon_i,\quad T_{2\ell}=n^{-1/2}\sum_{i=1}^{n}X_{i\ell}\varepsilon_i,\quad T_{3J\ell}=n^{-1/2}\sum_{i=1}^{n}B_{J,\ell}^{(d)}(X_{i\ell})\varepsilon_i,$$

and $T_1=\max_{1\le k\le p_1}|T_{1k}|$, $T_2=\max_{1\le\ell\le p_2}|T_{2\ell}|$ and $T_3=\max_{1\le\ell\le p_2,1\le J\le N_n}|T_{3J\ell}|$. Then we have

$$\mathrm{E}(T_1)\le C_1\sqrt{\ln(p_1)},\quad \mathrm{E}(T_2)\le C_2\sqrt{\ln(p_2)},$$
$$\mathrm{E}(T_3)\le C_3n^{-1/2}\sqrt{\ln(p_2N_n)}\left\{\sqrt{2C_4\,nN_n\ln(2p_2N_n)}+C_5N_n^{1/2}\ln(2p_2N_n)+n\right\}^{1/2},$$

where C_1, C_2, C_3, C_4 and C_5 are positive constants. In particular, when $N_n\ln(p_2N_n)/n\to 0$, we have

$$\mathrm{E}(T_1)=O\{\sqrt{\ln(p_1)}\},\quad \mathrm{E}(T_2)=O\{\sqrt{\ln(p_2)}\},\quad \mathrm{E}(T_3)=O\{\sqrt{\ln(p_2N_n)}\}.$$

Lemma B.6. *For*

$$\boldsymbol{v}_1 = \left(\left(\frac{\omega_k^{\alpha} \overline{\theta}_{0,k}}{|\overline{\theta}_{0,k}|}, k \in \mathcal{S}_z \right)^{\top}, \mathbf{0}_{|\mathcal{S}_{x,L}|}^{\top}, \mathbf{0}_{|\mathcal{S}_{x,N}|N_n}^{\top} \right)^{\top}, \quad \boldsymbol{v}_2 = \left(\mathbf{0}_{|\mathcal{S}_z|}^{\top}, \left(\frac{\omega_{\ell}^{\beta} \overline{\theta}_{0,|\mathcal{S}_z|+\ell}}{|\overline{\theta}_{0,|\mathcal{S}_z|+\ell}|}, \ell \in \mathcal{S}_{x,L} \right)^{\top}, \mathbf{0}_{|\mathcal{S}_{x,N}|N_n}^{\top} \right)^{\top},$$

$$\boldsymbol{v}_3 = \left(\mathbf{0}_{|\mathcal{S}_z|}^{\top}, \mathbf{0}_{|\mathcal{S}_{x,L}|}^{\top}, \left(\frac{\omega_{\ell}^{\gamma} \overline{\boldsymbol{\theta}}_{0,|\mathcal{S}_z|+|\mathcal{S}_{x,L}|+\ell}^{\top}}{\|\overline{\boldsymbol{\theta}}_{0,|\mathcal{S}_z|+|\mathcal{S}_{x,L}|+\ell}\|_2}, \ell \in \mathcal{S}_{x,N} \right)^{\top} \right)^{\top},$$

under Assumption (A5),

$$\|\boldsymbol{v}_1\|_2^2 = O_P(h_{n1}^2) = O_P(b_{n1}^{-4} c_{b1}^{-2} r_{n1}^{-2} + |\mathcal{S}_z| b_{n1}^{-2}), \quad \|\boldsymbol{v}_2\|_2^2 = O_P(h_{n2}^2) = O_P(b_{n2}^{-4} c_{b2}^{-2} r_{n2}^{-2} + |\mathcal{S}_{x,L}| b_{n2}^{-2}),$$
$$\|\boldsymbol{v}_3\|_2^2 = O_P(h_{n3}^2) = O_P(b_{n3}^{-4} c_{b3}^{-2} r_{n3}^{-2} + |\mathcal{S}_{x,N}| b_{n3}^{-2}).$$

References

[1] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, 2004.
[2] P. Breheny, `Grpreg`: Regularization paths for regression models with grouped covariates, 2016, `R` package version 3.0-2. Available at "https://cran.r-project.org/web/packages/grpreg/index.html".
[3] P. Breheny, J. Huang, Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, Stat. Comput. 25 (2015) 173–187.
[4] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.
[5] Z. Chen, J. Chen, Tournament screening cum EBIC for feature selection with high-dimensional feature spaces, Sci. China Ser. A: Math. 52 (2009) 1327–1341.
[6] G. Claeskens, I. Van Keilegom, Bootstrap confidence bands for regression curves and their derivatives, Ann. Statist. 31 (2003) 1852–1884.
[7] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, J. Amer. Statist. Assoc. 106 (2011) 544–557.
[8] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, CRC Press, Boca Raton, FL, 1996.
[9] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[10] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B. Stat. Methodol. 70 (2008) 849–911.
[11] J. Huang, P. Breheny, S. Ma, A selective review of group selection in high-dimensional models, Statist. Sci. 27 (2012) 481–499.
[12] J. Huang, J.L. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Statist. 38 (2010) 2282–2313.
[13] J. Huang, F. Wei, S. Ma, Semiparametric regression pursuit, Statist. Sinica 22 (2012) 1403–1426.
[14] E.R. Lee, H. Noh, B.U. Park, Model selection via bayesian information criterion for quantile regression models, J. Amer. Statist. Assoc. 109 (2014) 216–229.
[15] S. Leiboff, X. Li, H.-C. Hu, N. Todt, J. Yang, X. Li, X. Yu, G.J. Muehlbauer, M.C. Timmermans, J. Yu, P.S. Schnable, M.J. Scanlon, Genetic control of morphometric diversity in the maize shoot apical meristem, Nature Comm. 6 (2015) 8974–9974.
[16] X. Li, L. Wang, D. Nettleton, Sparse model identification and learning for ultra-high-dimensional additive partially linear models, 2018, Full version available at: arXiv.org/abs/1811.00488.
[17] X. Li, L. Wang, D. Nettleton, Ultra-high-dimensional additive partial linear models, Stat (2019).
[18] H. Lian, H. Liang, D. Ruppert, Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models, Statist. Sinica 25 (2015) 591–607.
[19] H. Lian, H. Liang, L. Wang, Generalized additive partial linear models for clustered data with diverging number of covariates using GEE, Statist. Sinica 24 (2014) 173–196.
[20] H. Liang, S.W. Thurston, D. Ruppert, T. Apanasovich, R. Hauser, Additive partial linear models with measurement errors, Biometrika 95 (2008) 667–678.
[21] H.Y. Lin, Q. Liu, X. Li, J. Yang, S. Liu, Y. Huang, M.J. Scanlon, D. Nettleton, P.S. Schnable, Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS, Genome Biol. 18 (2017) 192.
[22] X. Liu, L. Wang, H. Liang, Estimation and variable selection for semiparametric additive partial linear models, Statist. Sinica 21 (2011) 1225–1248.
[23] S. Ma, Q. Song, L. Wang, Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data, Bernoulli 19 (2013) 252–274.
[24] S. Ma, L. Yang, Spline-backfitted kernel smoothing of partially linear additive model, J. Statist. Plann. Inference 141 (2011) 204–219.
[25] D. Ruppert, Selecting the number of knots for penalized splines, J. Comput. Graph. Statist. 11 (2002) 735–757.
[26] M.P. Wand, M.C. Jones, Kernel Smoothing, CRC Press, Boca Raton, FL, 1995.
[27] J. Wang, L. Yang, Efficient and fast spline-backfitted kernel smoothing of additive models, Ann. Inst. Statist. Math. 61 (2009) 663–690.
[28] L. Wang, X. Liu, H. Liang, R.J. Carroll, Estimation and variable selection for generalized additive partial linear models, Ann. Statist. 39 (2011) 1827–1851.
[29] L. Wang, L. Xue, A. Qu, H. Liang, Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates, Ann. Statist. 42 (2014) 592–624.
[30] L. Wang, L. Yang, Spline-backfitted kernel smoothing of nonlinear additive autoregression model, Ann. Statist. 35 (2007) 2474–2503.
[31] L. Xue, L. Yang, Additive coefficient modeling via polynomial spline, Statist. Sinica 16 (2006) 1423–1446.
[32] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.
[33] H.H. Zhang, G. Cheng, Y. Liu, Linear or nonlinear? Automatic structure discovery for partially linear models, J. Amer. Statist. Assoc. 106 (2011) 1099–1112.
[34] S. Zheng, R. Liu, L. Yang, W.K. Härdle, Statistical inference for generalized additive models: Simultaneous confidence corridors and variable selection, Test 25 (2016) 607–626.
[35] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.