ESTIMATION AND INFERENCE IN GENERALIZED ADDITIVE COEFFICIENT MODELS FOR NONLINEAR INTERACTIONS WITH HIGH-DIMENSIONAL COVARIATES

BY SHUJIE MA[1,*], RAYMOND J. CARROLL[2,†,‡],
HUA LIANG[3,§] AND SHIZHONG XU[*]

University of California, Riverside[*], Texas A&M University[†], University of Technology Sydney[‡] and George Washington University[§]

In the low-dimensional case, the generalized additive coefficient model (GACM) proposed by Xue and Yang [*Statist. Sinica* **16** (2006) 1423–1446] has been demonstrated to be a powerful tool for studying nonlinear interaction effects of variables. In this paper, we propose estimation and inference procedures for the GACM when the dimension of the variables is high. Specifically, we propose a groupwise penalization based procedure to distinguish significant covariates for the "large p small n" setting. The procedure is shown to be consistent for model structure identification. Further, we construct simultaneous confidence bands for the coefficient functions in the selected model based on a refined two-step spline estimator. We also discuss how to choose the tuning parameters. To estimate the standard deviation of the functional estimator, we adopt the smoothed bootstrap method. We conduct simulation experiments to evaluate the numerical performance of the proposed methods and analyze an obesity data set from a genome-wide association study as an illustration.

1. Introduction. Regression analysis is a commonly used statistical tool for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables denoted as $\mathbf{T} = (T_1, T_2, \ldots, T_p)^{\mathrm{T}}$. To study the marginal effects of the predictors on the response, one may fit a generalized linear model (GLM),

$$(1) \qquad E(Y|\mathbf{T}) = \mu(\mathbf{T}) = g^{-1}\{\eta(\mathbf{T})\}, \qquad \eta(\mathbf{T}) = \sum_{\ell=1}^{p} \alpha_{\ell 0} T_{\ell},$$

where g is a known monotone link function, and $\alpha_{\ell 0}$, $1 \leq \ell \leq p$, are unknown parameters. Sometimes, the effect of one variable may change with other variables; that is, there is an interaction effect. By letting $T_1 = 1$, to incorporate the interaction effects of $\mathbf{T}$ and the other variables, denoted as $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, model (1) can be modified to $E(Y|\mathbf{X}, \mathbf{T}) = \mu(\mathbf{X}, \mathbf{T}) = g^{-1}\{\eta(\mathbf{X}, \mathbf{T})\}$ with

$$(2) \qquad \eta(\mathbf{X}, \mathbf{T}) = \alpha_{10} + \sum_{\ell=2}^{p} \alpha_{\ell 0} T_\ell + \sum_{k=1}^{d} \alpha_{1k} X_k + \sum_{\ell=2}^{p} \sum_{k=1}^{d} \alpha_{\ell k} X_k T_\ell,$$

where $\alpha_{\ell k}$ for $0 \leq k \leq d$ and $1 \leq \ell \leq p$ are parameters. After a direct reformulation, model (2) can be written as

$$(3) \qquad \eta(\mathbf{X}, \mathbf{T}) = \sum_{\ell=1}^{p} \left(\alpha_{\ell 0} + \sum_{k=1}^{d} \alpha_{\ell k} X_k \right) T_\ell.$$

Here the effect of each T_ℓ changes linearly with X_k. However, in practice, this simple linear relationship may not reflect the true changing patterns of the coefficient with other covariates. We here use an example of gene and environment $(G \times E)$ interactions for illustration. It has been noticed in the literature that obesity is linked to genetic factors. Their effects, however, can be altered under different environmental factors such as sleeping hours [Knutson (2012)] and physical activity [Wareham, van Sluijs and Ekelund (2005)]. To have a rough idea of how the effects of the genetic factors change with the environment, we explore data from the Framingham Heart Study [Dawber, Meadors and Moore (1951)]. In Figure 1 we plot the estimated mean body mass index (BMI) against sleeping hours per day and activity hours per day, respectively, for people with three possible genotype categories represented by AA, Aa and aa, and for one single nucleotide polymorphism (SNP). A detailed description and the analysis of this data set are given in Section 5. We define allele A as the minor (less frequent) allele. This figure clearly shows different nonlinear curves for the three groups in each of the two plots. By
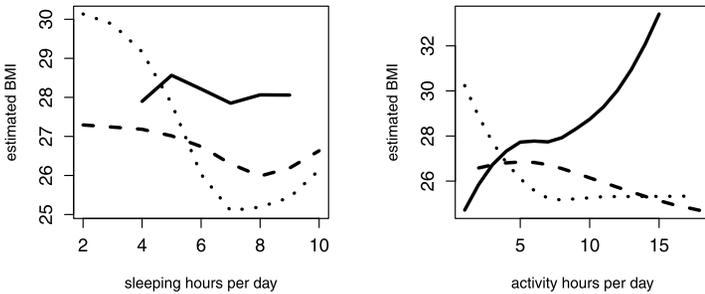

FIG. 1. *Plots of the estimated BMI against sleeping hours per day (left panel) and activity hours per day (right panel) for the three genotypes AA (solid line), Aa (dashed line) and aa (dotted line) of SNP rs242263 in the Framingham study, where A is the minor allele.*

letting T_ℓ be the indicator for the group ℓ, the linear function in model (3) is clearly misspecified.

To relax the linearity assumption, we allow each $\alpha_{\ell k} X_k$ term to be an unknown nonlinear function of X_k, and thus extend model (3) to the generalized additive coefficient model (GACM)

$$(4) \qquad \eta(\mathbf{X}, \mathbf{T}) = \sum_{\ell=1}^{p} \left\{ \alpha_{\ell 0} + \sum_{k=1}^{d} \alpha_{\ell k}(X_k) \right\} T_\ell = \sum_{\ell=1}^{p} \alpha_\ell(\mathbf{X}) T_\ell.$$

For identifiability, the functional components satisfy $E\{\alpha_{\ell k}(X_k)\} = 0$ for $1 \le k \le d$ and $1 \le \ell \le p$. The conditional variance of Y is modeled as a function of the mean, that is, $\mathrm{var}(Y|\mathbf{X}, \mathbf{T}) = V\{\mu(\mathbf{X}, \mathbf{T})\} = \sigma^2(\mathbf{X}, \mathbf{T})$. In each coefficient function of the GACM, covariates X_k are continuous variables. If some of them are discrete, they will enter linearly. For example, if X_k is binary, we let $\alpha_{\ell k}(X_k) = \alpha_{\ell k} X_k$. In such a case, model (4) turns out to be a partially linear additive coefficient model. The linearity of (4) in T_ℓ is particularly appropriate when those factors are discrete, for example, SNPs in a genome-wide association study (GWAS), as in the data example of Section 5.

For the low-dimensional case that the dimensions of $\mathbf{X}$ and $\mathbf{T}$ are fixed, estimation of model (4) has been studied; see Liu and Yang (2010), Xue and Liang (2010), Xue and Yang (2006) for a spline estimation procedure and Lee, Mammen and Park (2012) for a backfitting algorithm. In modern data applications, model (4), however, is particularly useful when p is large. For example, in GWAS, the number of SNPs, which is p, can be very large, but the dimension of $\mathbf{X}$ such as the environmental factors, which is d, is inevitably relatively small. Moreover, the number of variables in $\mathbf{T}$ which have nonzero effects is small. It therefore, poses new challenges to apply model (4) to the high-dimensional case including: (i) how to identify those important variables in $\mathbf{T}$, (ii) how to estimate the coefficient functions for the important covariates and (iii) how to conduct inferences for the nonzero coefficient functions. For example, it is of interest to know whether they are a function of a specific parametric form such as constant, linear or quadratic, etc.

In the high-dimensional data setting, studying nonlinear interaction effects has found much attention in recent years, and a few strategies have been proposed. For example, Jiang and Liu (2014) proposed to detect variables under the general index model, which enables the study of high-order interactions among components of continuous predictors, which are assumed to have a multivariate normal distribution. Moreover, Lian (2012) considered variable selection in varying coefficient models which allows the coefficient functions to depend on one index variable, such as a time-dependent variable.

When we would like to see how the effect of each genetic factor changes under the influence of multiple environmental variables, the proposed high-dimensional GACM (4) becomes a natural approach to consider, since both the index model

[Jiang and Liu (2014)] and the varying coefficient model [Lian (2012)] cannot address this question; the former is used to study interactions of components in a set of continuous predictors, and the latter only allows one index variable. For model selection and estimation, we apply a groupwise penalization method. Moreover, most existing high-dimensional nonparametric modeling papers [Lian (2012), Meier, van de Geer and Bühlmann (2009), Ravikumar et al. (2009), Wang et al. (2014), Huang, Horowitz and Wei (2010)] focus on variable selection and estimation. In this paper, after variable selection, we also propose a simultaneous inferential tool to further test the shape of the coefficient function for each selected variable, which has not been studied in the previous works.

To this end, we aim to address questions (i)–(iii). Specifically, for estimation and model selection, we apply a groupwise regularization method based on a penalized quasi-likelihood criterion. The penalty is imposed on the L_2 norm of the spline coefficients of the spline estimators for $\alpha_\ell(\cdot)$. We establish the asymptotic consistency of model selection and estimation for the proposed group penalized estimators with the quasi-likelihood criterion in the high-dimensional GACM (4). We allow p to grow with n at an almost exponential order. Importantly, establishment of these results is technically more difficult than other work based on least squares, since no closed-form of the estimators exists from the penalized quasi-likelihood method.

After selecting the important variables, the next question of interest is what shapes the nonzero coefficient functions may have. Then we need to provide an inferential tool to further check whether a coefficient function has some specific parametric form. For example, when it is a constant or a linear function, the corresponding covariate has no or linear interaction effects with another covariate, respectively. For global inference, we construct simultaneous confidence bands (SCBs) for the nonparametric additive functions based on a two-step estimation procedure. By using the selected variables, we first propose a refined two-step spline estimator for the function of interest, which is proved to have a pointwise asymptotic normal distribution and oracle efficiency. We then establish the bounds for the SCBs based on the absolute maxima distribution of a Gaussian process and on the strong approximation lemma [Csörgő and Révész (1981)]. Some other related works on SCBs for nonparametric functions include Claeskens and Van Keilegom (2003), Hall and Titterington (1988), Härdle and Marron (1991), among others. We provide an asymptotic formula for the standard deviation of the spline estimator for the coefficient function, which involves unknown population parameters to be estimated. The formula has somewhat complex expressions and contains many parameters. Direct estimation therefore may be not accurate, particularly with the small or moderate sample sizes. As an alternative, the bootstrap method provides us a reliable way to calculate the standard deviation by avoiding estimating those population parameters. We here apply the smoothed bootstrap method suggested by Efron (2014), which advocated that the method can improve coverage probability to calculate the pointwise estimated standard deviations for the

estimators of the coefficient functions. This method was originally proposed for calculating the estimated standard deviation of the estimate of a parameter of interest, such as the conditional mean. We extend this method to the case of functional estimation. We demonstrate by simulation studies in Section 4 that compared to the traditional resampling bootstrap method, the smoothed bootstrap method can successfully improve the empirical coverage rate.

The paper is organized as follows. Section 2 introduces the B-spline estimation procedure for the nonparametric functions, describes the adaptive group Lasso estimators and the initial Lasso estimators and presents asymptotic results. Section 3 describes the two-step spline estimators and introduces the simultaneous confidence bands and the bootstrap methods for calculating the estimated standard deviation. Section 4 describes simulation studies, and Section 5 illustrates the method through the analysis of an obesity data set from a genome-wide association study. Proofs are in the Appendix and additional supplementary material [Ma et al. (2015)].

2. Penalization based variable selection. Let $(Y_i, \mathbf{X}_i^{\mathrm{T}}, \mathbf{T}_i^{\mathrm{T}})$, $i = 1, \ldots, n$, be random vectors that are independently and identically distributed as $(Y, \mathbf{X}^{\mathrm{T}}, \mathbf{T}^{\mathrm{T}})$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$ and $\mathbf{T}_i = (T_{i1}, \ldots, T_{ip})^{\mathrm{T}}$. Write the negative quasi-likelihood function $Q(\mu, y) = \int_\mu^y \{(y - \zeta)/V(\zeta)\} \, d\zeta$. Estimation of the mean function can be achieved by minimizing the negative quasi-likelihood of the observed data

$$(5) \qquad \sum_{i=1}^n Q\{g^{-1}\{\eta(\mathbf{X}_i, \mathbf{T}_i)\}, Y_i\}.$$

2.1. *Spline approximation.* We approximate the smooth functions $\alpha_{\ell k}(\cdot)$, $1 \le k \le d$ and $1 \le \ell \le p$ in (4) by B-splines. As in most work on nonparametric smoothing, estimation of the functions $\alpha_{\ell k}(\cdot)$ is conducted on compact sets. Without loss of generality, let the compact set be $\mathcal{X} = [0, 1]$. Let G_n^0 be the space of polynomial splines of order $q \ge 2$. We introduce a sequence of spline knots

$$t_{-q-1} = \cdots = t_{-1} = t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1} = \cdots = t_{N+q},$$

where $N \equiv N_n$ is the number of interior knots. In the following, let $J_n = N_n + q$. For $0 \le j \le N$, let $H_j = t_{j+1} - t_j$ be the distance between neighboring knots and let $H = \max_{0 \le s \le N} H_j$. Following Zhou, Shen and Wolfe (1998), to study asymptotic properties of the spline estimators for $\alpha_{\ell k}(\cdot)$, we assume that $\max_{0 \le j \le N-1} |H_{j+1} - H_j| = o(N^{-1})$ and $H/\min_{0 \le j \le N} H_j \le M$, where $M > 0$ is a predetermined constant. Such an assumption is necessary for numerical implementation. In practice, we can use the quantiles as the locations of the knots. Let $\{b_{j,k}(x_k) : 1 \le j \le J_n\}^{\mathrm{T}}$ be the qth order B spline basis functions given on page 87 of de Boor (2001). For positive numbers a_n and b_n, $a_n \asymp b_n$ means that $\lim_{n \to \infty} a_n/b_n = c$, where c is some nonzero finite constant. For $1 \le j \le J_n$, we

adopt the centered B-spline functions given in Xue and Yang (2006) such that $B_{j,k}(x_k) = \sqrt{N}[b_{j,k}(x_k) - \{E(b_{j,k})/E(b_{1,k})\}b_{1,k}(x_k)]$, so that $E\{B_{j,k}(X_k)\} = 0$ and var$\{B_{j,k}(X_k)\} \asymp 1$. Define the space G_n of additive spline functions as the linear space spanned by $B(\mathbf{x}) = \{1, B_{j,k}(x_k), 1 \le j \le J_n, 1 \le k \le d\}^{\mathrm{T}}$, where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$. According to the result on page 149 of de Boor (2001), for $\alpha_{\ell k}(\cdot)$ satisfying condition (C3) in Appendix A.2 such that $\alpha_{\ell k}^{(r-1)}(x_k) \in C^{0,1}[0, 1]$ for given integer $r \ge 1$, where $C^{0,1}[0, 1]$ is the space of Lipschitz continuous functions on $[0, 1]$ defined in Appendix A.2, there is a function

$$
(6) \qquad \alpha_{\ell k}^0(x_k) = \sum_{j=1}^{J_n} \gamma_{j,\ell k} B_j(x_k) \in G_n^0,
$$

such that $\sup_{x_k \in [0,1]} |\alpha_{\ell k}^0(x_k) - \alpha_{\ell k}(x_k)| = O(J_n^{-r})$. Then for every $1 \le \ell \le p$, $\alpha_\ell(\mathbf{x})$ can be approximated well by a linear combination of spline functions in G_n^0, so that

$$
(7) \qquad \alpha_\ell(\mathbf{x}) \approx \alpha_\ell^0(\mathbf{x}) = \gamma_{\ell 0} + \sum_{k=1}^{d} \sum_{j=1}^{J_n} \gamma_{j,\ell k} B_{j,k}(x_k) = B(\mathbf{x})^{\mathrm{T}} \boldsymbol{\gamma}_\ell,
$$

where $\boldsymbol{\gamma}_\ell = (\gamma_{\ell 0}, \boldsymbol{\gamma}_{\ell 1}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{\ell d}^{\mathrm{T}})^{\mathrm{T}}$, in which $\boldsymbol{\gamma}_{\ell k} = (\gamma_{j,\ell k} : 1 \le j \le J_n)^{\mathrm{T}}$. Thus the minimization problem in (5) is equivalent to finding $\widetilde{\boldsymbol{\gamma}}^0 = (\widetilde{\boldsymbol{\gamma}}_\ell^{0\mathrm{T}}, 1 \le \ell \le p)^{\mathrm{T}}$ with $\widetilde{\boldsymbol{\gamma}}_\ell^0 = (\widetilde{\gamma}_{\ell 0}^0, \widetilde{\boldsymbol{\gamma}}_{\ell 1}^{0\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\gamma}}_{\ell d}^{0\mathrm{T}})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\gamma}}_{\ell k}^0 = (\widetilde{\gamma}_{j,\ell k}^0 : 1 \le j \le J_n)^{\mathrm{T}}$ to minimize $\sum_{i=1}^{n} Q[g^{-1}\{\sum_{\ell=1}^{p} B(\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{\gamma}_\ell T_\ell\}, Y_i]$. The components of the additive coefficients are estimated by $\widetilde{\alpha}_{\ell k}^0(x_k) = \sum_{j=1}^{J_n} \widetilde{\gamma}_{j,\ell k}^0 B_j(x_k) = B(\mathbf{x})^{\mathrm{T}} \widetilde{\boldsymbol{\gamma}}_{\ell k}^0$ and $\widetilde{\alpha}_{\ell 0}^0 = \widetilde{\gamma}_{\ell 0}^0$.

2.2. *Adaptive group Lasso estimator.* We now describe the procedure for estimating and selecting the additive coefficient functions by using the adaptive group Lasso. The estimators are obtained by minimizing a penalized negative quasi-likelihood criterion. We establish asymptotic selection consistency as well as the convergence rate of the estimators to the true nonzero functions. For any vector $\mathbf{a} = (a_1, \ldots, a_s)^{\mathrm{T}}$, let its L_2 norm be $\|\mathbf{a}\|_2 = \sqrt{a_1^2 + \cdots + a_s^2}$. For any measurable L_2-integrable function ϕ on $[0, 1]^d$, define the L_2 norm as $\|\phi\|^2 = E\{\phi^2(\mathbf{X})\}$.

We are interested in identifying the significant components of the vector $\mathbf{T} = (T_1, \ldots, T_p)^{\mathrm{T}}$. Let s, a fixed number, be the total number of nonzero α_ℓ's and $I_1 = \{\ell : \|\alpha_\ell\| \ne 0, 1 \le \ell \le p\}$. Let I_2 be the complementary set of I_1; that is, $I_2 = \{\ell : \alpha_\ell(\cdot) \equiv 0, 1 \le \ell \le p\}$. Recalling the approximation given in (7), $\boldsymbol{\gamma}_\ell$ is zero if and only if each element of $\boldsymbol{\gamma}_\ell$ is zero; that is, $\|\boldsymbol{\gamma}_\ell\|_2 = 0$. We apply the adaptive group Lasso approach in Huang, Horowitz and Wei (2010) for variable selection in model (4). In order to identify zero additive coefficients, we penalize the L_2 norm of the coefficients $\boldsymbol{\gamma}_\ell$ for $1 \le \ell \le p$. Let $w_n = (w_{n1}, \ldots, w_{np})^{\mathrm{T}}$ be a given vector of weights, which needs to be chosen appropriately to achieve selection consistency.

Their choice will be discussed in Section 2.3. We consider the penalized negative quasi-likelihood

$$(8) \qquad L_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{\ell=1}^{p} B^{\mathrm{T}}(\mathbf{X}_i)\boldsymbol{\gamma}_\ell T_\ell\right\}, Y_i\right] + n\lambda_n \sum_{\ell=1}^{p} w_{n\ell}\|\boldsymbol{\gamma}_\ell\|_2,$$

where λ_n is a regularization parameter controlling the amount of shrinkage. The estimator $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\gamma}}_p^{\mathrm{T}})^{\mathrm{T}}$ is obtained by minimizing (8). Minimization of (8) is solved by local quadratic approximation as adopted by Fan and Li (2001).

For $\ell = 1, \ldots, p$, the ℓth additive coefficient function is estimated by

$$\widehat{\alpha}_\ell(\mathbf{x}) = \widehat{\gamma}_{\ell 0} + \sum_{k=1}^{d} \sum_{j=1}^{J_n} \widehat{\gamma}_{j,\ell k} B_{j,k}(x_k) = B^{\mathrm{T}}(\mathbf{x})\widehat{\boldsymbol{\gamma}}_\ell.$$

We will make the following two assumptions on the order requirements of the tuning parameters. Write $w_{n,I_1} = (w_{n\ell} : \ell \in I_1)$.

ASSUMPTION 1. $J_n^2\{n\log(n)\}^{-1} \to 0$ and $\lambda_n\|w_{n,I_1}\|_2 \to 0$, as $n \to \infty$.

ASSUMPTION 2. $n\lambda_n\|w_{n,I_1}\|_2 + n^{1/2}J_n^{1/2}\sqrt{\log(pJ_n)} + nJ_n^{-r} = o(n\lambda_n w_{n\ell})$, for all $\ell \in I_2$.

The following theorem presents the selection consistency and estimation properties of the adaptive group Lasso estimators.

THEOREM 1. *Under conditions* (C1)–(C5) *in the Appendix and Assumptions* 1 *and* 2: (i) *as* $n \to \infty$, $P(\|\widehat{\alpha}_\ell\| > 0, \ell \in I_1$ *and* $\|\widehat{\alpha}_\ell\| = 0, \ell \in I_2) \to 1$, *and* (ii) $\|\widehat{\alpha}_\ell - \alpha_\ell\| = O_p(\lambda_n\|w_{n,I_1}\|_2 + n^{-1/2}J_n^{1/2} + J_n^{-r}), \ell \in I_1$.

2.3. *Choice of the weights.* We now discuss how to choose the weights used in (8) based on the initial estimates. For low-dimensional data settings with $p < n$, an unpenalized estimator such as least squares estimator [Zou (2006)] can be used as an initial estimate. For high-dimensional settings with $p \gg n$, it has been discussed [Meier and Bühlmann (2007)] that the Lasso estimator is a more appropriate choice. Following Huang, Horowitz and Wei (2010), we obtain an initial estimate with the group Lasso by minimizing

$$L_{n1}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{\ell=1}^{p} B(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\gamma}_\ell T_\ell\right\}, Y_i\right] + n\lambda_{n1} \sum_{\ell=1}^{p} \|\boldsymbol{\gamma}_\ell\|_2,$$

with respect to $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_p^{\mathrm{T}})^{\mathrm{T}}$. Denote the resulting estimators by $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\gamma}}_p^{\mathrm{T}})^{\mathrm{T}}$. Let $\widetilde{I}_1 = \{\ell : \|\widetilde{\boldsymbol{\gamma}}_\ell\|_2 \neq 0, 1 \leq \ell \leq p\}$, and let $\widetilde{s}$ be the number of elements in $\widetilde{I}_1$.

Under conditions (C1)–(C5) in the Appendix, and when $\lambda_{n1} \geq C n^{-1/2} J_n^{1/2} \times \sqrt{\log(p J_n)}$ for a sufficiently large constant C, we have: (i) the number of estimated nonzero functions are bounded; that is, as $n \to \infty$, there exists a constant $1 < C_1 < \infty$ such that $P(\tilde{s} \leq C_1 s) \to 1$; (ii) if $\lambda_{n1} \to 0$, then $P(\|\tilde{\boldsymbol{\gamma}}_\ell\|_2 > 0$ for all $l \in I_1) \to 1$; (iii) $\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = O_p(\lambda_{n1} + n^{-1/2} J_n^{1/2} + J_n^{-r})$. We refer to Theorems 1(i) and (ii) of Huang, Horowitz and Wei (2010) for the proofs of (i) and (ii), and Theorem 1 in our paper for the proof of (iii).

The weights we use are $w_{n\ell} = \|\tilde{\boldsymbol{\gamma}}_\ell\|_2^{-1}$, if $\|\tilde{\boldsymbol{\gamma}}_\ell\|_2 > 0$; $w_{n\ell} = \infty$, if $\|\tilde{\boldsymbol{\gamma}}_\ell\|_2 = 0$.

REMARK 1. Assumptions 1 and 2 give the order requirements of J_n and λ_n. Based on the condition that $J_n^2 \{n \log(n)\}^{-1} \to 0$ given in Assumption 1, we need $J_n \ll \{n \log(n)\}^{1/2}$, where $a_n \ll b_n$ denotes that $a_n / b_n = o(1)$ for any positive numbers a_n and b_n, and λ_n needs to satisfy $n^{-1/2} J_n^{1/2} \sqrt{\log(p J_n)} \times \{\min_{\ell \in I_2}(w_{n\ell})\}^{-1} \ll \lambda_n \ll 1$. From the above theoretical properties of the group Lasso estimators, we know that, with probability approaching 1, $\|\tilde{\boldsymbol{\gamma}}_\ell\|_2 > 0$ for nonzero components, and then the corresponding weights $w_{n\ell}$ are bounded away from 0 and infinity for $\ell \in I_1$. By defining $0 \cdot \infty = 0$, the components not selected by the group Lasso are not included in the adaptive group Lasso procedure. Let $J_n \asymp n^{1/(2r+1)}$, so that J_n has the optimal order for spline regression. If $p = \exp[o\{n^{2r/(2r+1)}\}]$, then $n^{-1/2} J_n^{1/2} \sqrt{\log(p J_n)} \to 0$. This means the dimension p can diverge with the sample size at an almost exponential rate.

2.4. *Selection of tuning parameters.* Tuning parameter selection always plays an important role in model and variable selection. An underfitted model can lead to severely biased estimation, and an overfitted model can seriously degrade the estimation efficiency. Among different data-driven methods, the Bayesian information criterion (BIC) tuning parameter selector has been shown to be able to identify the true model consistently in the fixed dimensional setting [Wang, Li and Tsai (2007)]. In the high-dimensional setting, an extend BIC (EBIC) and a generalized information criterion have been proposed by Chen and Chen (2008) and Fan and Tang (2013), respectively. In this paper, we adopt the EBIC method [Chen and Chen (2008)] to select the tuning parameter λ_n in (8). Specifically, the EBIC(λ_n) is defined as

$$2 \sum_{i=1}^n \left(Q\left[g^{-1} \left\{ \sum_{\ell=1}^p B(\mathbf{X}_i)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_\ell T_{i\ell} \right\}, Y_i \right] \right) + s^*(1 + d J_n) \log(n) + 2\nu \log\binom{p}{s^*},$$

where $(\widehat{\boldsymbol{\gamma}}_\ell)_{\ell=1}^p$ is the minimizer of (8) for a given λ_n, s^* is the number of nonzero estimated functions $(\widehat{\alpha}_\ell)_{\ell=1}^p$ and $0 \leq \nu \leq 1$ is a constant. Here we use $\nu = 0.5$. When $\nu = 0$, the EBIC is ordinary BIC.

We use cubic B-splines for the nonparametric function estimation, so that $q = 4$. In the penalized estimation procedure, we let the number of interior knots $N = \lfloor c n^{1/(2q+1)} \rfloor$ satisfy the optimal order, where $\lfloor a \rfloor$ denotes the largest integer no greater than a and c is a constant. In the simulations, we take $c = 2$.

3. Inference and the bootstrap smoothing procedure.

3.1. *Background.* After model selection, our next step is to conduct statistical inference for the coefficient functions of those important variables. We will establish a simultaneous confidence band (SCB) based on a two-step estimator for global inference. An asymptotic formula of the SCB will be provided based on the distribution of the maximum value of the normalized deviation of the spline functional estimate. To improve accuracy, we calculate the estimated standard deviation in the SCB by using the nonparametric bootstrap smoothing method as discussed in Efron (2014). For specificity, we focus on the construction of $\alpha_{\ell 1}(x_1)$, with $\alpha_{\ell k}(x_k)$ for $k \geq 2$ defined similarly, for $\ell \in \widehat{I}_1$, where $\widehat{I}_1 = \{\ell : \|\widehat{\alpha}_\ell\| \neq 0, 1 \leq \ell \leq p\}$.

Although the one-step penalized estimation in Section 2 can quickly identify nonzero coefficient functions, no asymptotic distribution is available for the resulting estimators. Thus we construct the SCB based on a refined two-step spline estimator for $\alpha_{\ell 1}(x_1)$, which will be shown to have the oracle property that the estimator of $\alpha_{\ell 1}(x_1)$ has the same asymptotic distribution as the univariate oracle estimator obtained by pretending that $\alpha_{\ell 0}$ and $\alpha_{\ell k}(X_k)$ for $\ell \in \widehat{I}_1$, $k \geq 2$ and $\alpha_\ell(\mathbf{X})$ for $\ell \notin \widehat{I}_1$ are known. See Horowitz, Klemelä and Mammen (2006), Horowitz and Mammen (2004), Liu, Yang and Härdle (2013) for kernel-based two-step estimators in generalized additive models, which also have the oracle property but are not as computationally efficient as the two-step spline method. We next introduce the oracle estimator and the proposed two-step estimator before we present the SCB.

3.2. *Oracle estimator.* In the following, we describe the oracle estimator of $\alpha_{\ell 1}(x_1)$. We rewrite model (4) as

(9)
$$\mu(\mathbf{X}, \mathbf{T}) = g^{-1}\{\eta(\mathbf{X}, \mathbf{T})\}$$
$$= \sum_{\ell \in \widehat{I}_1} \alpha_{\ell 1}(X_1)T_\ell + \sum_{\ell \in \widehat{I}_1}\left\{\alpha_{\ell 0} + \sum_{k \geq 2}\alpha_{\ell k}(X_k)\right\}T_\ell + \sum_{\ell \notin \widehat{I}_1}\alpha_\ell(\mathbf{X})T_\ell.$$

By assuming that $\alpha_{\ell 0}$ and $\alpha_{\ell k}(X_k)$ for $\ell \in \widehat{I}_1$, $k \geq 2$ and $\alpha_\ell(\mathbf{X})$ for $\ell \notin \widehat{I}_1$ are known, estimation in (9) involves only the nonparametric functions $\alpha_{\ell 1}(X_1)$ of a scalar covariate X_1. It will be shown in Theorem 2 that the estimator achieves the univariate optimal convergence rate when the optimal order for the number of knots is applied. We estimate $\alpha_1(x_1) = \{\alpha_{\ell 1}(x_1), \ell \in \widehat{I}_1\}^{\mathrm{T}}$ by minimizing the negative quasi-likelihood function as follows. Denote the oracle estimator by $\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) = B_1^{\mathcal{S}}(x_1)^{\mathrm{T}}\widehat{\gamma}_{\ell 1}^{\mathrm{OR}}$, where $\widehat{\gamma}_{\ell 1}^{\mathrm{OR}}$ is defined directly below, $B_1^{\mathcal{S}}(x_1) = \{B_{j,1}^{\mathcal{S}}(x_1), 1 \leq j \leq J_n^{\mathcal{S}}\}$ where $B_{j,1}^{\mathcal{S}}(x_1)$ is the centered B-spline function defined in the same way as $B_{j,1}(x_1)$ in Section 2, but with $N^{\mathcal{S}} = N_n^{\mathcal{S}}$ interior knots and $J_n^{\mathcal{S}} = N_n^{\mathcal{S}} + q$. Rates of increase for $J_n^{\mathcal{S}}$ are described in Assumptions 3 and 4 below. Let $\alpha_{\ell,-1}(\mathbf{X}_i) = \alpha_{\ell 0} + \sum_{k \geq 2}\alpha_{\ell k}(X_{ik})$. Then $\widehat{\gamma}_{,1}^{\mathrm{OR}} = \{(\widehat{\gamma}_{\ell 1}^{\mathrm{OR}})^{\mathrm{T}}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$ is

obtained by minimizing the negative quasi-likelihood

(10)
$$L_n^{\mathrm{OR}}(\boldsymbol{\gamma}_{,1}) = \sum_{i=1}^{n} Q\bigg[g^{-1}\bigg\{ \sum_{\ell \in \widehat{I}_1} B_1^{\mathcal{S}}(X_{i1})^{\mathrm{T}} \boldsymbol{\gamma}_{\ell 1} T_{i\ell} + \sum_{\ell \in \widehat{I}_1} \alpha_{\ell, -1}(\mathbf{X}_i) T_{i\ell} + \sum_{\ell \notin \widehat{I}_1} \alpha_{\ell}(\mathbf{X}_i) T_{i\ell} \bigg\}, Y_i \bigg],$$

where $\boldsymbol{\gamma}_{,1} = \{(\boldsymbol{\gamma}_{\ell 1})^{\mathrm{T}}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$. Similarly, the oracle estimator of $\boldsymbol{\alpha}_0 = \{\alpha_{\ell 0}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$, which is denoted as $\widehat{\boldsymbol{\alpha}}_0^{\mathrm{OR}} = \{\widehat{\alpha}_{\ell 0}^{\mathrm{OR}}, \ell \in \widehat{I}_1\}^{\mathrm{T}} = \{\widehat{\gamma}_{\ell 0}^{\mathrm{OR}}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$, is obtained by minimizing $L_n^{\mathrm{OR}}(\boldsymbol{\gamma}_{,0}) = \sum_{i=1}^{n} Q[g^{-1}\{\sum_{\ell \in \widehat{I}_1} \gamma_{\ell 0} T_{i\ell} + \sum_{\ell \in \widehat{I}_1} \alpha_{\ell, -0}(\mathbf{X}_i) T_{i\ell} + \sum_{\ell \notin \widehat{I}_1} \alpha_{\ell}(\mathbf{X}_i) T_{i\ell}\}, Y_i]$, where $\boldsymbol{\gamma}_{,0} = (\boldsymbol{\gamma}_{\ell 0}, \ell \in \widehat{I}_1)$ and $\alpha_{\ell, -0}(\mathbf{X}_i) = \sum_{k=1}^{d} \alpha_{\ell k}(X_{ik})$.

3.3. *Initial estimator.* The oracle estimator is infeasible because it assumes knowledge of the other functions. In order to obtain the two-step estimators of $\alpha_{\ell 1}(x_1)$ for $\ell \in \widehat{I}_1$, we first need initial estimators for $\alpha_{\ell 0}$ and $\alpha_{\ell k}(x_k)$ for $k \geq 2$ and $\ell \in \widehat{I}_1$, denoted as $\widehat{\alpha}_{\ell 0}^{\mathrm{ini}} = \widehat{\gamma}_{\ell 0}^{\mathrm{ini}}$ and $\widehat{\alpha}_{\ell k}^{\mathrm{ini}}(x_k) = B_k^{\mathrm{ini}}(x_k)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{\ell k}^{\mathrm{ini}}$, where $B_k^{\mathrm{ini}}(x_k) = \{B_{j,k}^{\mathrm{ini}}(x_k) : 1 \leq j \leq J_n^{\mathrm{ini}}\}^{\mathrm{T}}$ and $B_{j,k}^{\mathrm{ini}}(x_k)$ are B-spline functions with the number of interior knots N_n^{ini} and $J_n^{\mathrm{ini}} = N_n^{\mathrm{ini}} + q$. Rates of increase for J_n^{ini} are described in Assumptions 3 and 4 below. We need an undersmoothed procedure in the first step, so that the approximation bias can be reduced, and the difference between the two-step and oracle estimators is asymptotically negligible. We obtain $\widehat{\boldsymbol{\gamma}}_{\widehat{I}_1}^{\mathrm{ini}} = \{(\widehat{\boldsymbol{\gamma}}_{\ell}^{\mathrm{ini}})^{\mathrm{T}} : \ell \in \widehat{I}_1\}^{\mathrm{T}}$, where $\widehat{\boldsymbol{\gamma}}_{\ell}^{\mathrm{ini}} = \{\widehat{\gamma}_{\ell 0}^{\mathrm{ini}}, (\widehat{\boldsymbol{\gamma}}_{\ell k}^{\mathrm{ini}})^{\mathrm{T}}\}^{\mathrm{T}}$, by minimizing the negative quasi-likelihood $\sum_{i=1}^{n} Q[g^{-1}\{\sum_{\ell \in \widehat{I}_1} B(\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{\gamma}_{\ell} T_{\ell}\}, Y_i]$. The adaptive group Lasso penalized estimator $\widehat{\boldsymbol{\gamma}}_{\widehat{I}_1} = \{(\widehat{\boldsymbol{\gamma}}_{\ell})^{\mathrm{T}} : \ell \in \widehat{I}_1\}^{\mathrm{T}}$ obtained in Section 2 can also be used as the initial estimator. We, however, refit the model with the selected variables and obtain the initial estimator $\widehat{\boldsymbol{\gamma}}_{\widehat{I}_1}^{\mathrm{ini}}$ in order to improve estimation accuracy in high-dimensional data settings.

3.4. *Final estimator.* In the second step, we construct the two-step estimator of $\alpha_{\ell 1}$ for $\ell \in \widehat{I}_1$. We replace $\alpha_{\ell 0}$ and $\alpha_{\ell k}(X_k)$ by the initial estimators $\widehat{\alpha}_{\ell 0}^{\mathrm{ini}}$ and $\widehat{\alpha}_{\ell k}^{\mathrm{ini}}(X_k)$ for $\ell \in \widehat{I}_1$ and $k \geq 2$ and replace $\alpha_{\ell}(\mathbf{X})$ for $\ell \notin \widehat{I}_1$ by $\widehat{\alpha}_{\ell}(\mathbf{X}) = 0$. Let $\widehat{\alpha}_{\ell, -1}^{\mathrm{ini}}(\mathbf{X}_i) = \widehat{\alpha}_{\ell 0}^{\mathrm{ini}} + \sum_{k \geq 2} \widehat{\alpha}_{\ell k}^{\mathrm{ini}}(X_{ik})$. Denote the two-step spline estimator of $\alpha_{\ell 1}(x_1)$ as $\widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) = B_1^{\mathcal{S}}(x_1)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{\ell 1}^{\mathcal{S}}$ with $\widehat{\boldsymbol{\gamma}}_{,1}^{\mathcal{S}} = \{(\widehat{\boldsymbol{\gamma}}_{\ell 1}^{\mathcal{S}})^{\mathrm{T}}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$ minimizing

(11)
$$L_n^{\mathcal{S}}(\boldsymbol{\gamma}_{,1}) = \sum_{i=1}^{n} Q\bigg[g^{-1}\bigg\{ \sum_{\ell \in \widehat{I}_1} B_1^{\mathcal{S}}(X_{i1})^{\mathrm{T}} \boldsymbol{\gamma}_{\ell 1} T_{i\ell} + \sum_{\ell \in \widehat{I}_1} \widehat{\alpha}_{\ell, -1}^{\mathrm{ini}}(\mathbf{X}_i) T_{i\ell} + \sum_{\ell \notin \widehat{I}_1} \widehat{\alpha}_{\ell}(\mathbf{X}_i) T_{i\ell} \bigg\}, Y_i \bigg].$$

Then the two-step of $\alpha_{\ell 0}$, denoted as $\widehat{\alpha}_{\ell 0}^{\mathcal{S}} = \widehat{\gamma}_{\ell 0}^{\mathcal{S}}$, is obtained in the same way as $\widehat{\alpha}_{\ell 0}^{\mathrm{OR}}$ by replacing $\alpha_{\ell,0}(\mathbf{X}_i)$ with $\widehat{\alpha}_{\ell,0}^{\mathrm{ini}}(\mathbf{X}_i) = \sum_{k=1}^{d} \widehat{\alpha}_{\ell k}^{\mathrm{ini}}(X_{ik})$ for $\ell \in \widehat{I}_1$ and replacing $\alpha_\ell(\mathbf{X}_i)$ with $\widehat{\alpha}_\ell(\mathbf{X}_i) = 0$ for $\ell \notin \widehat{I}_1$. Let $\widehat{\boldsymbol{\alpha}}_0^{\mathcal{S}} = \{\widehat{\alpha}_{\ell 0}^{\mathcal{S}}, \ell \in \widehat{I}_1\}^{\mathrm{T}}$.

3.5. *Asymptotic normality and uniform oracle efficiency.* We now establish the asymptotic normality and uniform oracle efficiency for the oracle and final estimators. Let $Z_{ij\ell,1} = B_{j,1}^{\mathcal{S}}(X_{i1})T_{i\ell}$ and $Z_{i,1} = (Z_{ij\ell,1}, 1 \leq j \leq J_n^{\mathcal{S}}, \ell \in \widehat{I}_1)^{\mathrm{T}}$. Let s^* be the number of elements in $\widehat{I}_1$. By Theorem 1, $P(s^* = s) \to 1$. For simplicity of notation, denote $\sigma_i^2 = \sigma^2(\mathbf{X}_i, \mathbf{T}_i)$ and $\eta_i = \eta(\mathbf{X}_i, \mathbf{T}_i)$. Define $s^* \times s^* J_n^{\mathcal{S}}$ matrix $B^{\mathcal{S}}(x_1)$ as

$$\begin{bmatrix} B_{1,1}^{\mathcal{S}}(x_1) & \cdots & B_{J_n^{\mathcal{S}},1}^{\mathcal{S}}(x_1) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{1,1}^{\mathcal{S}}(x_1) & \cdots & B_{J_n^{\mathcal{S}},1}^{\mathcal{S}}(x_1) \end{bmatrix}.$$

To establish the asymptotic distribution of the two-step estimator, in addition to Assumptions 1 and 2 given in Section 2, we make the following two assumptions on the number of basis functions $J_n^{\mathcal{S}}$ and J_n^{ini}:

ASSUMPTION 3. (i) $s^*(J_n^{\mathcal{S}})^2 \{n \log(n)\}^{-1} = o(1)$ and $s^*(J_n^{\mathcal{S}})^{-r} = o(1)$, and (ii) $n(\log n)^{-1}(J_n^{\mathcal{S}} J_n^{\mathrm{ini}})^{-1} \to \infty$, as $n \to \infty$.

ASSUMPTION 4. $(n/J_n^{\mathcal{S}})^{1/2}(J_n^{\mathrm{ini}})^{-r} \to 0$, as $n \to \infty$.

First we describe the asymptotic normality of the oracle estimator $\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1)$ of $\alpha_{\ell 1}(x_1)$. Let $\widehat{\alpha}_1^{\mathrm{OR}}(x_1) = \{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1), \ell \in \widehat{I}_1\}^{\mathrm{T}}$. Let $b_1(x_1) = E\{\widehat{\alpha}_1^{\mathrm{OR}}(x_1)|\mathbb{X}, \mathbb{T}\}$ and $b_{\ell 1}(x_1) = E\{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1)|\mathbb{X}, \mathbb{T}\}$, for $\ell \in \widehat{I}_1$, where $(\mathbb{X}, \mathbb{T}) = (\mathbf{X}_i, \mathbf{T}_i)_{i=1}^{n}$.

THEOREM 2. *Under conditions* (C1)–(C5) *and Assumption* 3(i), *for any vector* $\mathbf{a} \in R^{s^*}$ *with* $\|\mathbf{a}\|_2 = 1$, *for any* $x_1 \in [0, 1]$, $\mathbf{a}^{\mathrm{T}}\sigma_n^{-1}(x_1)\{\widehat{\alpha}_1^{\mathrm{OR}}(x_1) - b_1(x_1)\} \to N(0, 1)$, *where*

$$(12) \qquad \sigma_n^2(x_1) = B^{\mathcal{S}}(x_1)\left[\sum_{i=1}^{n} Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2 \Big/ \sigma_i^2\right]^{-1} B^{\mathcal{S}}(x_1)^{\mathrm{T}},$$

where $\dot{g}^{-1}(\eta_i)$ *is the first-order derivative of* $g^{-1}(\eta_i)$ *with respect to* η_i, *and*

$$\sum_{\ell \in \widehat{I}_1} \|\widehat{\alpha}_{\ell 1}^{\mathrm{OR}} - b_{\ell 1}\|^2 = O_p(s^* J_n^{\mathcal{S}} n^{-1}), \qquad \sum_{\ell \in I_1} \|b_{\ell 1} - \alpha_{\ell 1}\|^2 = O_p\{(s^*)^2 (J_n^{\mathcal{S}})^{-2r}\}.$$

Thus for $\ell \in \widehat{I}_1$, $\sigma_{n1}^{-1}(x_1)\{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - b_{\ell 1}(x_1)\} \to N(0, 1)$, *where*

$$(13) \qquad \sigma_{n1}^2(x_1) = \mathbf{e}_\ell^{\mathrm{T}} \sigma_n^2(x_1) \mathbf{e}_\ell,$$

and $\mathbf{e}_\ell$ *is the* s^*-*dimensional vector with the* ℓth *element* 1 *and other elements* 0, *and* $\|\widehat{\boldsymbol{\alpha}}_0^{\mathrm{OR}} - \boldsymbol{\alpha}_0\|_2 = O_p(\sqrt{s^*/n})$.

The next result shows the uniform oracle efficiency of the two-step estimator that the difference between the two-step estimator $\widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1)$ and oracle estimator $\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1)$ is uniformly asymptotically negligible, and thus the two-step estimator is oracle in the sense that it has the same asymptotic distribution as the oracle estimator. Let $\widehat{\alpha}_1^{\mathcal{S}}(x_1) = \{\widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1), \ell \in \widehat{I}_1\}^{\mathrm{T}}$.

THEOREM 3. *Under conditions* (C1)–(C5) *in the Appendix and Assumptions* 1–3,

$$\sup_{x_1 \in [0,1]} \|\widehat{\alpha}_1^{\mathcal{S}}(x_1) - \widehat{\alpha}_1^{\mathrm{OR}}(x_1)\|_\infty = O_p\{(n^{-1}\log n)^{1/2} + (J_n^{\mathrm{ini}})^{-r}\},$$

$\|\widehat{\boldsymbol{\alpha}}_0^{\mathcal{S}} - \widehat{\boldsymbol{\alpha}}_0^{\mathrm{OR}}\|_2 = o_p(n^{-1/2})$, *and furthermore under Assumption* 4,

$$\sup_{x_1 \in [0,1]} |\mathbf{a}^{\mathrm{T}}\sigma_n^{-1}(x_1)\{\widehat{\alpha}_1^{\mathcal{S}}(x_1) - \widehat{\alpha}_1^{\mathrm{OR}}(x_1)\}| = o_p(1),$$

for any vector $\mathbf{a} \in R^{s^*}$ *with* $\|\mathbf{a}\|_2 = 1$ *and* $\sigma_n^2(x_1)$ *given in* (12). *Hence, for any* $x_1 \in [0, 1]$, $\mathbf{a}^{\mathrm{T}}\sigma_n^{-1}(x_1)\{\widehat{\alpha}_1^{\mathcal{S}}(x_1) - b_1(x_1)\} \to N(0, 1)$.

REMARK 2. Under Assumptions 1 and 2, by Theorem 1, with probability approaching 1, $s^* = s$, which is a fixed number. In the second step, by letting $J_n^{\mathcal{S}} \asymp n^{1/(2r+1)}$, the nonparametric functions $\alpha_{\ell 1}$ for $\ell \in \widehat{I}_1$ are approximated by spline functions with the optimal number of knots. By the conditions that $(n/J_n^{\mathcal{S}})(J_n^{\mathrm{ini}})^{-1} \to 0$ and $n(\log n)^{-1}(J_n^{\mathcal{S}} J_n^{\mathrm{ini}})^{-1} \to \infty$ given in Assumptions 3 and 4, J_n^{ini} needs to satisfy $n^{1/(2r+1)} \ll J_n^{\mathrm{ini}} \ll n^{2r/(2r+1)}(\log n)^{-1}$ where $r \geq 1$. By using the adaptive group lasso estimator as the initial estimator, Assumption 1 requires that $J_n^{\mathrm{ini}} \ll \{n \log(n)\}^{1/2}$. Hence $n^{1/(2r+1)} \ll J_n^{\mathrm{ini}} \ll \{n \log(n)\}^{1/2}$. We therefore can let $J_n^{\mathrm{ini}} \asymp n^{(1+\vartheta)/(2r+1)}$, where ϑ is any small positive number close to 0. This increase in the number of basis functions ensures undersmoothing in the first step in order that the uniform difference between the two-step and the oracle estimators become asymptotically negligible. Based on Assumptions 1 and 2, the tuning parameter λ_n needs to satisfy $n^{-1/2}(J_n^{\mathrm{ini}})^{1/2}\sqrt{\log(p J_n^{\mathrm{ini}})}\{\min_{\ell \in I_2}(w_{n\ell})\}^{-1} \ll \lambda_n \ll 1$.

REMARK 3. The number of interior knots has the same order requirement as the number of basis functions. In the first step, with the undersmoothing requirement as discussed in Remark 2, we let the number of interior knots $N^{\mathrm{ini}} = \lfloor cn^{(1+0.01)/(2q+1)} \rfloor$, where c is a constant, by assuming that $r = q$. In the simulations, we let $c = 2$. In the second-step estimation, we use BIC to select the number of knots N^S, so the optimal N^S ranges in $[\lfloor n^{1/(2q+1)} \rfloor, \lfloor 2n^{1/(2q+1)} \rfloor]$ by minimizing BIC: $\mathrm{BIC}(N^S) = 2L_n^{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_{\cdot 1}^{\mathcal{S}}) + d(N^S + q)\log(n)$.

3.6. *Simultaneous confidence bands.* In this section, we propose a SCB for $\alpha_{\ell 1}(x_1)$ by studying the asymptotic behavior of the maximum of the normalized deviation of the spline functional estimate. To construct asymptotic SCBs for $\alpha_{\ell 1}(x_1)$ over the interval $x_1 \in [0, 1]$ with confidence level $100(1 - \alpha)\%$, $\alpha \in (0, 1)$, we need to find two functions $l_{\ell n}(x_1)$ and $u_{\ell n}(x_1)$ such that

(14) $\lim\limits_{n \to \infty} P\big(l_{\ell n}(x_1) \leq \alpha_{\ell 1}(x_1) \leq u_{\ell n}(x_1) \text{ for all } x_1 \in [0, 1]\big) = 1 - \alpha.$

In practice, we consider a variant of (14) and construct SCBs over a subset $S_{n,1}$ of $[0, 1]$ with $S_{n,1}$ becoming denser as $n \to \infty$. We, therefore, partition $[0, 1]$ according to L_n equally spaced intervals based on $0 < \xi_0 < \xi_1 < \cdots < \xi_{L_n} < \xi_{L_n+1} = 1$ where $L_n \to \infty$ as $n \to \infty$. Let $S_{n,1} = (\xi_0, \ldots, \xi_{L_n})$. Define $d_{L_n}(\alpha) = 1 - \{2\log(L_n + 1)\}^{-1}[\log\{-(1/2)\log(1 - \alpha)\} + (1/2)\{\log\log(L_n + 1) + \log(4\pi)\}]$, and $Q_{L_n}(\alpha) = \{2\log(L_n + 1)\}^{1/2}d_{L_n}(\alpha)$.

THEOREM 4. *Under conditions* (C1)–(C5) *in the Appendix, and* $L_n \asymp J_n^{\mathcal{S}} \asymp n^{1/(2r+1)}$ *and* $n^{1/(2r+1)} \ll J_n^{\mathrm{ini}} \ll n^{2r/(2r+1)}\{\log(n)\}^{-1}$, *we have*

$$\lim\limits_{n \to \infty} P\Big\{ \sup\limits_{x_1 \in S_{n,1}} \big|\sigma_{n1}^{-1}(x_1)\{\widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) - \alpha_{\ell 1}(x_1)\}\big| \leq Q_{L_n}(\alpha) \Big\} = 1 - \alpha,$$

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

(15) $\widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) \pm \sigma_{n1}(x_1)Q_{L_n}(\alpha).$

REMARK 4. Compared to the pointwise confidence intervals with width $2Z_{1-\alpha/2}\sigma_n(x_1)$, the width of the confidence bands (15) is inflated by a rate $\{2\log(L_n + 1)\}^{1/2}d_{L_n}(\alpha)/Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the cut-off point of the $100(1 - \alpha)$th percentile of the standard normal.

3.7. *Bootstrap smoothing for calculating the standard error.* Theorem 4 establishes a thresholding value $Q_{L_n}(\alpha)$ for the SCB. One critical question is how to estimate the standard deviation $\sigma_{n1}(x_1)$ in order to construct the SCB. We can use a sample estimate of $\sigma_{n1}(x_1)$ according to the asymptotic formula given in (12), which may have approximation error and thus lead to inaccurate results for inference. The bootstrap estimate of the standard deviation provides an alternative way. We here propose a bootstrap smoothed confidence band by adopting the nonparametric bootstrap smoothing idea from Efron (2014), which can eliminates discontinuities in jumpy estimates. The procedure is described as follows.

Let $\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_n\}$ be the data we have, where $\mathbf{D}_i = \{Y_i, \mathbf{X}_i, (T_{i\ell}, \ell \in \widehat{I}_1)\}$. Denote $\mathbf{D}^* = \{\mathbf{D}_1^*, \ldots, \mathbf{D}_n^*\}$ as a nonparametric bootstrap sample from $\{\mathbf{D}_1, \ldots, \mathbf{D}_n\}$, and $\mathbf{D}_{(j)}^* = \{\mathbf{D}_{(j)1}^*, \ldots, \mathbf{D}_{(j)n}^*\}$ as the jth bootstrap sample in B draws. Let $\widehat{\alpha}_{\ell 1,(j)}^{*\mathcal{S}}(x_1)$ be the two-step estimator of $\alpha_{\ell 1}(x_1)$ by using the data

$\mathbf{D}^*_{(j)}$. We first present an empirical standard deviation by the traditional resampling method which is given as

$$(16) \qquad \widehat{\sigma}_{\ell 1, B}(x_1) = \left[\sum_{j=1}^{B} \{\widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1) - \widehat{\alpha}_{\ell 1, \cdot}^{*\mathcal{S}}(x_1)\}^2 \Big/ (B-1) \right]^{1/2},$$

where $\widehat{\alpha}_{\ell 1, \cdot}^{*\mathcal{S}}(x_1) = \sum_{j=1}^{B} \widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1)/B$. Then a $100(1-\alpha)\%$ unsmoothed bootstrap SCB for $\alpha_{\ell 1}(x_1)$ over $x_1 \in S_{n,1}$ is given as

$$(17) \qquad \widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) \pm \widehat{\sigma}_{\ell 1, B}(x_1) Q_{L_n}(\alpha).$$

Another choice is the smoothed bootstrap SCB which eliminates discontinuities in the estimates [Efron (2014)]. Let

$$\widetilde{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) = \sum_{j=1}^{B} \widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1)/B$$

be the smoothed estimate of $\alpha_{\ell 1}(x_1)$ obtained by averaging over the bootstrap replications. Let $C_{(j)i}^* = \#\{\mathbf{D}_{(j)i'}^* = \mathbf{D}_i\}$ be the number of elements in $\mathbf{D}_{(j)i'}^*$ equaling $\mathbf{D}_i$.

PROPOSITION 1. *At each point $x_1 \in S_{n,1}$, the nonparametric delta-method estimate of the standard deviation for the smoothed bootstrap statistic $\widetilde{\alpha}_{\ell 1}^{\mathcal{S}}(x_1)$ is $\widetilde{\sigma}_{\ell 1}(x_1) = \{\sum_{i=1}^{n} \mathrm{cov}_i^2(x_1)\}^{1/2}$, where $\mathrm{cov}_i(x_1) = \mathrm{cov}_*\{C_{(j)i}^*, \widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1)\}$ which is the bootstrap covariance between $C_{(j)i}^*$ and $\widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1)$.*

The proof of Proposition 1 essentially follows the same arguments as the proof for Theorem 1 in Efron (2014). Based on Proposition 1, to construct the smoothed bootstrap SCB, we use the nonparametric estimate of the standard deviation given as

$$(18) \qquad \widetilde{\sigma}_{\ell 1, B}(x_1) = \left\{ \sum_{i=1}^{n} \widehat{\mathrm{cov}}_{i, B}^2(x_1) \right\}^{1/2},$$

where

$$\widehat{\mathrm{cov}}_{\ell i, B}(x_1) = \sum_{j=1}^{B} (C_{(j)i}^* - C_{\cdot i}^*)(\widehat{\alpha}_{\ell 1, (j)}^{*\mathcal{S}}(x_1) - \widehat{\alpha}_{\ell 1, \cdot}^{*\mathcal{S}}(x_1))/B$$

with $C_{\cdot i}^* = \sum_{j=1}^{B} C_{(j)i}^*/B$. The $100(1-\alpha)\%$ smoothed bootstrap SCB for $\alpha_{\ell 1}(x_1)$ over $x_1 \in S_{n,1}$ is given as

$$(19) \qquad \widetilde{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) \pm \widetilde{\sigma}_{\ell 1, B}(x_1) Q_{L_n}(\alpha).$$

4. A simulation study. In this section, we present a simulation study to evaluate the finite sample performance of our proposed penalized estimation procedure and the simultaneous confidence bands. More numerical studies are located in the supplementary materials [Ma et al. (2015)].

EXAMPLE 1. In this example, we use 1286 SNPs located on the sixth chromosome from the Framingham Heart Study to simulate the binary response from the logistic model

$$(20) \quad \text{logit}\{P(Y_i = 1|\mathbf{X}_i, \mathbf{T}_i)\} = \sum_{\ell=1}^{p} \alpha_\ell(\mathbf{X}_i) T_{i\ell} = \sum_{\ell=1}^{p} \left\{ \alpha_{\ell 0} + \sum_{k=1}^{2} \alpha_{\ell k}(X_{ik}) \right\} T_{i\ell},$$

with the four SNPs ss66063578, ss66236230, ss66194604 and ss66533844 selected from the real data analysis in Section 5 as important covariates and the other SNPs as unimportant covariates, so that $s = 4$ (the number of important covariates), $p = 1286$ and the sample size $n = 300$. The three possible allele combinations are coded as 1, 0 and -1 for each SNP. The covariates X_{ik}, $k = 1, 2$, are simulated environmental effects, which are generated from independent uniform distributions on $[0, 1]$. We generate the coefficient functions as $\alpha_{10} = 0.5$, $\alpha_{11}(x_1) = 4\cos(2\pi x_1)$, $\alpha_{12}(x_2) = 5\{(2x_2 - 1)^2 - 1/3\}$, $\alpha_{20} = 0.5$, $\alpha_{21}(x_1) = 6x_1 - 3$, $\alpha_{22}(x_2) = 4\{\sin(2\pi x_2) + \cos(2\pi x_2)\}$, $\alpha_{30} = 0.5$, $\alpha_{31}(x_1) = 4\sin(2\pi x_1)$, $\alpha_{32}(x_2) = 6x_2 - 3$, $\alpha_{40} = 0.5$, $\alpha_{41}(x_1) = 4\cos(2\pi x_1)$, $\alpha_{42}(x_2) = 5\{(2x_2 - 1)^2 - 1/3\}$ and $\alpha_\ell(\mathbf{X}_i) = 0$ for $l = 5, \ldots, 1286$. We conducted 500 replications for each simulation. We fit the data with the GACM (20) by using the adaptive group lasso (AGL) and group lasso (GL). In the literature, the generalized varying coefficient model [GVCM; Lian (2012)], which considers one index variable in the coefficient function for each predictor $T_{i\ell}$, has been widely used to study nonlinear interactions. To apply the GVCM method [Lian (2012)] in this setting, we first perform principal component analysis (PCA) on $\mathbf{X}_i$ and then use the first principal component as the index variable in the GVCM. Then we apply the AGL and GL methods to the GVCM: $\text{logit}\{P(Y_i = 1|\mathbf{X}_i, \mathbf{T}_i)\} = \sum_{\ell=1}^{p} \alpha_\ell(U_i) T_{i\ell}$, where U_i is the first principal component obtained by PCA on $\mathbf{X}_i$. Moreover, we also fit the data with the parametric logistic regression by assuming linear coefficient functions (3) with the AGL method. We also compare our proposed method with the conventional screening method by parametric logistic regression for Genome-Wide Association Studies [GWAS; Murcray, Lewinger and Gauderman (2009)]. In the screening method, we fit a logistic model for each SNP: $\text{logit}\{P(Y_i = 1|\mathbf{X}_i, T_{i\ell})\} = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X}_i + \beta_\ell T_{i\ell} + \sum_{k=1}^{2} \beta_{\ell k} X_{ik} T_{i\ell}$, for $\ell = 1, \ldots, 1286$. Then we conduct a likelihood ratio test for the genetic and interaction effects of $H_0 : \beta_\ell = \beta_{\ell 1} = \beta_{\ell 2} = \beta_{\ell 3} = 0$. Let $\alpha_0 = 0.05$ be the overall type I error for the study and $M = 1286$ be the number of SNPs in this study. We apply the multiple testing correction procedure for GWAS with H_0 rejected when the p-value $< \alpha_0/M_{\text{eff}}$, where M_{eff} is the Cheverud–Nyholt estimate of

TABLE 1

Variable selection and estimation results by the adaptive group lasso and the group lasso with the GACM and GVCM, respectively, and parametric logistic regression with adaptive group lasso and screening methods based on 500 *replications. The columns of C, O and I show the percentage of correct-fitting, over-fitting and incorrect-fitting. The columns TP, FP and MR show true positives, false positives and model errors, respectively*

		C	O	I	TP	FP	MR
GACM	AGL	0.410	0.460	0.130	3.860	0.870	0.059
	GL	0.140	0.764	0.096	3.904	2.540	0.083
GVCM	AGL	0.030	0.000	0.970	1.636	5.685	0.142
	GL	0.060	0.000	0.940	2.076	20.670	0.120
Logistic regression	AGL	0.000	0.000	1.000	1.872	1.174	0.159
	Screening	0.000	0.000	1.000	1.056	0.786	0.141

the effective number of tests [Cheverud (2001), Nyholt (2004)] calculated by $M_{\text{eff}} = 1 + M^{-1} \sum_{j=1}^{M} \sum_{k=1}^{M} (1 - r_{jk}^2)$ and r_{jk} are the correlation coefficients of the SNPs, and we obtain $M_{\text{eff}} = 1275.65$.

Table 1 presents the percentages of correct-fitting (C) (exactly the important covariates are selected), over-fitting (O) (both the important covariates and some unimportant covariates are selected) and incorrect-fitting (I) (some of the important covariates are not selected), the average true positives (TP), that is, the average number of selected covariates among the important covariates, the average false positives (FP), that is, the average number of selected covariates among the unimportant covariates, and the average model errors (MR), the latter defined as $\sum_{i=1}^{n} \{\hat{\mu}_i(\mathbf{X}_i, \mathbf{T}_i) - \mu_i(\mathbf{X}_i, \mathbf{T}_i)\}^2 / n$, where $\hat{\mu}_i(\mathbf{X}_i, \mathbf{T}_i)$ and $\mu_i(\mathbf{X}_i, \mathbf{T}_i)$ are the estimated and true conditional means for Y_i, respectively. We see that by fitting the proposed GACM, the GL method has larger percentage of over-fitting as well as larger average false positives than the AGL methods. The AGL improves the correct-fitting percentage by 26%. As a result, the AGL reduces the model fitting error by $(0.083 - 0.059)/0.059 = 40.7\%$ compared to the GL method. Moreover, both the logistic model and the GVCM fail to identify those important covariates with incorrect-fitting percentage close to or being 1. Furthermore, by using the screening method with logistic regression, the average true positive is 1.056, which is much less than 4 (the number of those important SNPs). This further illustrates that the traditional screening method is not an effective tool to identify important genetic factors in this context. In addition, we observe that the results for the AGL method in Table 1 are comparable to the results in Table S.1 of Example 2 (in the supplementary materials) at $p = 1000$ with the simulated SNPs in terms of having similar correct-fitting percentages and MR values.

Next, we investigate the empirical coverage rates of the unsmoothed and smoothed SCBs given in (17) and (19). To calculate the unsmoothed and smoothed

TABLE 2
The empirical coverage rates (cov) and the sample average of median and mean of the standard deviations (sd.median and sd.mean) for the unsmoothed SCB (17) and smoothed SCB (19) for the coefficient functions $\alpha_{\ell 1}(x_1)$ for $\ell = 1, 2, 3, 4$

	Unsmoothed bootstrap			Smoothed bootstrap		
	cov	sd.median	sd.mean	cov	sd.median	sd.mean
α_{11}	0.610	0.689	0.809	0.818	0.735	0.982
α_{21}	0.628	0.563	0.725	0.846	0.666	0.932
α_{31}	0.636	0.736	0.832	0.869	0.837	1.053
α_{41}	0.646	0.768	0.843	0.882	0.891	1.064

bootstrap standard deviations (16) and (18), we use $B = 500$ bootstrap replications. The confidence bands are constructed at $L_n = 20$ equally spaced points. At 95% confidence level, Table 2 reports the empirical coverage rates (cov) and the sample averages of median and mean standard deviations (sd.median and sd.mean), respectively, for the unsmoothed SCB (17) and smoothed SCB (19) for coefficient functions $\alpha_{\ell 1}(x_1)$, $\ell = 1, 2, 3, 4$. We see that the smoothed bootstrap method leads to better performance, having empirical coverage rates closer to the nominal confidence level 0.95.

5. Data application. We illustrate our method via analysis of the Framingham Heart Study [Dawber, Meadors and Moore (1951)] to investigate the effects of G × E interactions on obesity. People are defined as obese when their body mass index (BMI) is 30 or greater: this is the definition of being obese made by the U.S. Centers for Disease Control and Prevention; see http://www.cdc.gov/obesity/adult/defining.html. We defined the response variable to be $Y = 1$ for BMI ≥ 30; and $Y = 0$ for BMI < 30. We use $X_1 =$ sleeping hours per day; $X_2 =$ activity hours per day; and $X_3 =$ diastolic blood pressure as the environmental factors, and use single nucleotide polymorphisms (SNPs) located in the sixth chromosome as the genetic factors. The three possible allele combinations are coded as 1, 0 and -1. As in the simulation, we thus are fitting a multiplicative risk model in the SNPs. For details on genotyping, see http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?studyid=phs000007.v3.p2. A total of 1286 SNPs remain in our analysis after eliminating SNPs with minor allele frequency <0.05, those with departure from Hardy–Weinberg equilibrium and those having correlation coefficient with the response between -0.1 and 0.1. We have $n = 300$ individuals left in our study after deleting observations with missing values.

To see possible nonlinear main effects of the environmental factors, we first fit a generalized additive model by using X_1, X_2 and X_3 as predictors such that

$$(21) \qquad E(Y_i | \mathbf{X}_i, \mathbf{T}_i) = g^{-1}\{\eta(\mathbf{X}_i)\} \qquad \text{with } \eta(\mathbf{X}_i) = m_0 + \sum_{k=1}^{3} m_k(X_{ik}).$$

Figure S.1 given in the supplementary material [Ma et al. (2015)] depicts the plots of $\widehat{m}_k(\cdot)$ for $k = 1, 2, 3$ by one-step cubic spline estimation. Clearly the estimate of each nonparametric function has a nonlinear pattern. We refer to Section S.2 for the detailed description of this figure. Based on the plots shown in Figure S.1, we fit the GACM model

$$(22) \qquad \eta(\mathbf{X}_i, \mathbf{T}_i) = \sum_{\ell=1}^{1287} \left\{ \alpha_{\ell 0} + \sum_{k=1}^{3} \alpha_{\ell k}(X_{ik}) \right\} T_{i\ell},$$

where $\mathbf{T}_i = (T_{i1}, T_{i2}, \ldots, T_{i\,1287})^{\mathrm{T}}$ with $T_{i1} = 1$, and $T_{i\ell}$ are the SNP covariates for $\ell = 2, \ldots, 1287$. The nonparametric function $\alpha_{\ell k}(\cdot)$ is estimated by cubic splines, and the number of interior knots for each step is selected based on the criterion described in Section 2.4. We select variables in model (22) by the proposed adaptive group lasso (AGL) and the group lasso (GL). To compare the proposed model with linear models, we perform the group lasso by assuming linear interaction effects (Linear) such that $\alpha_\ell(\mathbf{X}_i) = \alpha_{\ell 0} + \sum_{k=1}^{3} \beta_{\ell k} X_{ik}$, and we also perform the lasso by assuming no interaction effects (No interaction) such that $\alpha_\ell(\mathbf{X}_i) = \alpha_{\ell 0}$. We also apply the screening method with parametric logistic regression (Screening) as described in Example 2. Table 3 reports the variable selection results in these five scenarios. After model selection, we calculate the estimated leave-one-out cross-validation prediction error (CVPE) for the model with the selected variables as shown in the last row of Table 3. Among the selected SNPs by the AGL method, two SNPs, rs4714924 and rs6543930, have been scientifically confirmed by Randall et al. (2013) to have strong associations with obesity. Moreover, compared to the linear, no interaction and screening methods, our proposed AGL with GACM method enables us to identify more genetic factors, which may be important to the response but missed out by other methods. As a result, it has the smallest CVPE (0.078), so that it significantly improves model prediction compared to other methods. We also see that the logistic model that completely ignores interactions has the largest CVPE (0.152). The screening method has the second largest CVPE (0.149), which is larger than that of the penalization method (0.124) obtained by fitting the same logistic regression model but including interaction considered. This result demonstrates that the screening method is not as effective as the penalization method for analysis of this data set, a result which also agrees with our simulations.

Next we fit the final GACM selected variables from the AGL procedure as

$$(23) \qquad \eta(\mathbf{X}_i, \mathbf{T}_i) = \sum_{\ell=1}^{10} \left\{ \alpha_{\ell 0} + \sum_{k=1}^{3} \alpha_{\ell k}(X_{ik}) \right\} T_{i\ell}.$$

MA, CARROLL, LIANG AND XU

TABLE 3
Variable selection results for the group lasso (GL) and the adaptive group lasso (AGL) in model (22), *the group lasso by assuming linear interaction effects* (linear), *the lasso by assuming no interaction effects* (no interaction) *and the screening method* (screening). *The symbol $\sqrt{}$ indicates that the SNP was selected into the model. The last row shows the cross validation prediction errors* (CVPE)

SNPs	GL	AGL	Linear	No interaction	Screening
rs9296244	$\sqrt{}$	$\sqrt{}$			
rs6910353	$\sqrt{}$	$\sqrt{}$			
rs3130813	$\sqrt{}$	$\sqrt{}$			
rs9353447				$\sqrt{}$	$\sqrt{}$
rs4714924	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$		
rs242263	$\sqrt{}$	$\sqrt{}$	$\sqrt{}$		$\sqrt{}$
rs282123	$\sqrt{}$				
rs282128	$\sqrt{}$	$\sqrt{}$			
rs6929006			$\sqrt{}$		
rs9353711	$\sqrt{}$				
rs12199154	$\sqrt{}$	$\sqrt{}$			
rs2277114	$\sqrt{}$				
rs749517	$\sqrt{}$				
rs729888	$\sqrt{}$				
rs203139			$\sqrt{}$		
rs6914589	$\sqrt{}$	$\sqrt{}$			
rs6543930	$\sqrt{}$	$\sqrt{}$			
CVPE	0.099	0.078	0.124	0.152	0.149

To illustrate the main effects of the environmental factors, Figure 2 plots the smoothed two-step estimated functions $\widetilde{\alpha}_{1k}^{\mathcal{S}}(\cdot)$ of the functions $\alpha_{1k}^{\mathcal{S}}(\cdot)$, for $k = 1, 2, 3$, and the associated 95% smoothed SCBs (upper and lower solid lines). The plots of the functional estimates have the same nonlinear change patterns as the corresponding plots in Figure S.1, although because of the addition of the SCBs, the scale of the plot has changed.

To illustrate the effects of the genetic factors changing with the environmental factors, in Figure 3 we plot the smoothed two-step estimated functions $\widetilde{\alpha}_{6k}^{\mathcal{S}}(\cdot)$ and the associated 95% smoothed SCBs of the coefficient functions $\alpha_{6k}^{\mathcal{S}}(\cdot)$ for the SNP rs242263. To further demonstrate how the probability of developing obesity changes with the environmental factors for each category of SNP rs242263, Figure 4 plots the estimated conditional probability of obesity against each environmental factor by letting $T_{i\ell} = 0$ for $\ell \neq 6$. Letting A be the minor allele, the curves are for aa (solid line), Aa (dashed line) and AA (dotted line). Figure 3 indicates different changing patterns of the interaction effects under different environments. For example, sleeping hours seem to have an overall more significant interaction effect with this particular SNP than the other two variables. The effect of this SNP
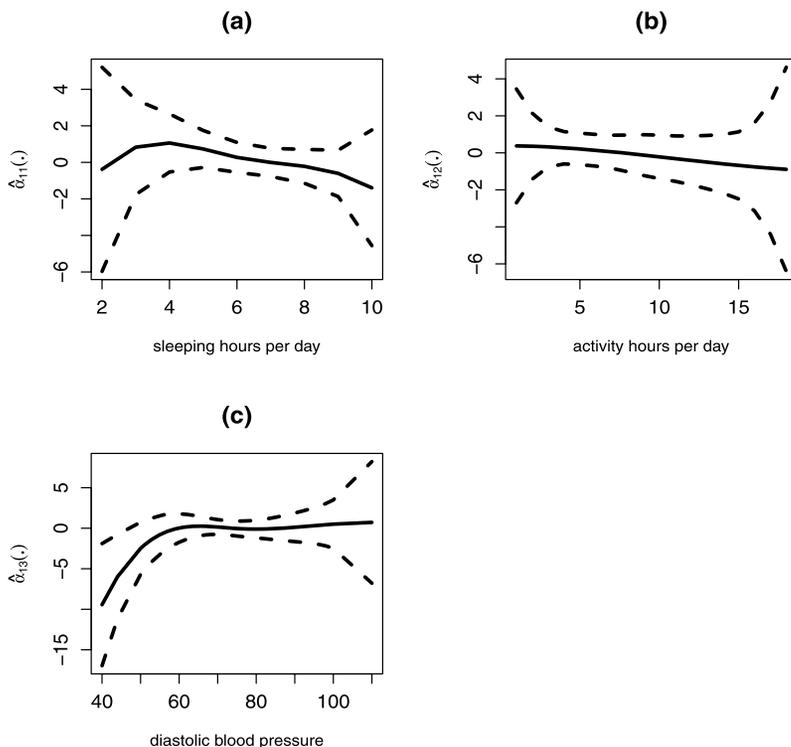
(a)

(b)

(c)

FIG. 2. *Plots of the smoothed two-step estimated functions $\widetilde{\alpha}_{1k}^{\mathcal{S}}(\cdot)$ for $k = 1, 2, 3$ and the associated 95% SCBs based on model* (23).

changes from positive to negative and then to positive again as the sleeping hours increase. The coefficient functions of the SNP have an increasing pattern along with the activity hours and diastolic blood pressure, respectively. From Figure 4, we observe that there are stronger differences among the levels AA, Aa, and aa of SNP rs242263 for both large and small values of the environmental factors.There are other interesting results worth further study. For example, in the 2–6 hours per day sleeping range, the AA group (dotted lines) have much higher rates of obesity than the aa group (solid line), but the opposite occurs in the 6–9 hour range. For those with low amounts of activity per day, again the AA group is more obese than the aa group, while when activity increases, the AA group is less obese than the aa group. A similar noticeable difference occurs between the <60 diastolic blood pressure group, those who are hypotensive, and the >90 group, those who are hypertensive, although there are few subjects in the former group.

6. Discussions. The generalized additive coefficient model (GACM) proposed by Xue and Yang (2006) and Xue and Liang (2010) has been demonstrated to be a powerful tool for studying nonlinear interaction effects of variables. To promote the use of the GACM in modern data applications such as gene-environment
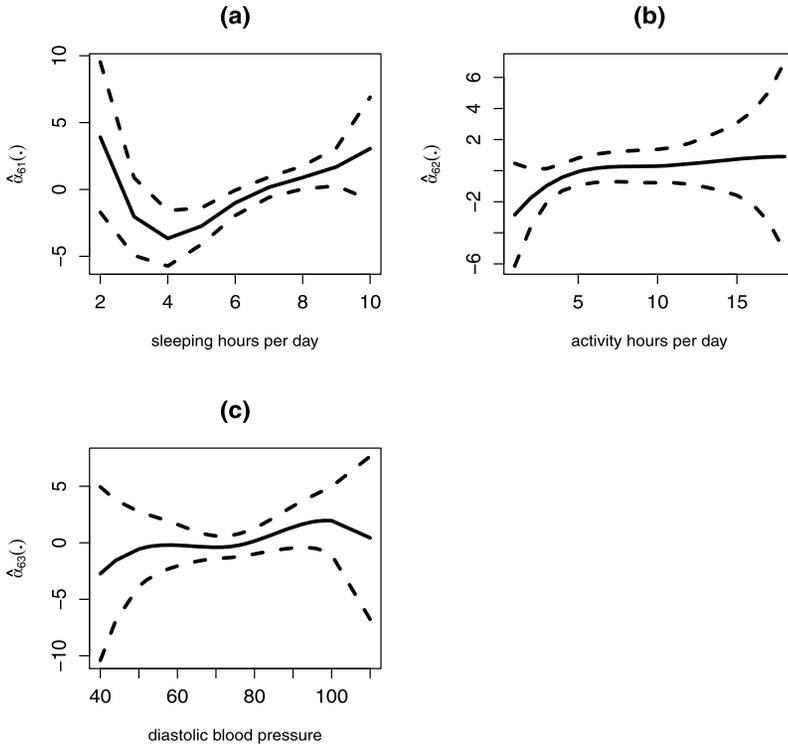
FIG. 3. *Plots of the smoothed two-step estimated functions $\widetilde{\alpha}_{5k}^{\mathcal{S}}(\cdot)$ for $k = 1, 2, 3$ and the associated 95% SCBs based on model* (23).

(G $\times$ E) interaction effects in GWAS, we have proposed estimation and inference procedures for the GACM when the dimension of the variables is high. Specifically, we have devised a groupwise penalization method in the GACM for simultaneous model selection and estimation. We showed by numerical studies that we can effectively identify important genetic factors by using the proposed nonparametric model while traditional generalized parametric models such as logistic regression model fails to do so when nonlinear interactions exist. Moreover, by comparing with the conventional screening method with logistic regression as commonly used in the GWAS community, our proposed groupwise penalization method with the GACM has been demonstrated to be more effective for variable selection and model estimation. After identifying those important covariates, we have further constructed simultaneous confidence bands for the nonzero coefficient functions based on a refined two-step estimator. We estimate the standard deviation of the functional estimator by a smoothed bootstrap method as proposed in Efron (2014). The method was shown to have good numerical performance by reducing variability as well as improving the empirical coverage rate of the proposed simultaneous confidence bands. Our methods can be extended to longitudinal data
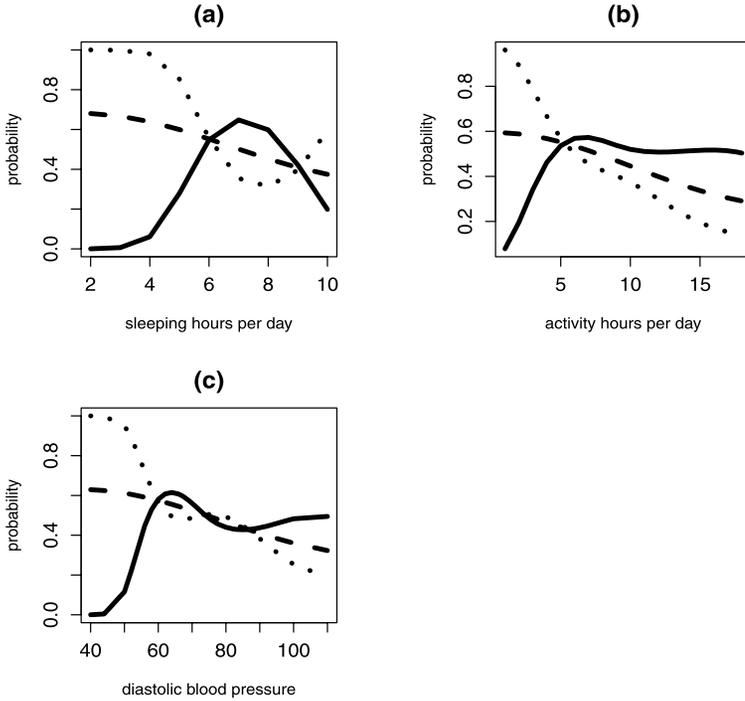
(a)

(b)

(c)

FIG. 4. *Plots of the estimated conditional probability of obesity against each environmental factor by letting $T_{i\ell} = 0$ for $\ell \neq 5$. With A being the minor allele, the curves are aa (solid line), Aa (dashed line) and AA (dotted line), based on model* (23).

settings through marginal models or mixed-effects models. More work, however, is needed to understand the properties of the estimators in such new settings. Moreover, extending this work to the setting with the dimensions for both genetic and environmental factors growing with the sample size can be a future project to be considered. Some associated theoretical properties with respect to model selection and estimation as well as inference need to be carefully investigated.

APPENDIX

Denote the space of the qth order smooth functions as $C^{(q)}([0, 1]) = \{\phi | \phi^{(q)} \in C[0, 1]\}$. For any $s \times s$ symmetric matrix $\mathbf{A}$, denote its L_q norm as $\|\mathbf{A}\|_q = \max_{\varsigma \in R^s, \|\varsigma\|_2 = 1} \|\mathbf{A}\varsigma\|_q$. Let $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^{s} |a_{ij}|$. For a vector $\mathbf{a}$, let $\|\mathbf{a}\|_\infty = \max_{1 \leq i \leq s} |a_i|$.

Let $C^{0,1}(\mathcal{X}_w)$ be the space of Lipschitz continuous functions on $\mathcal{X}_w$, that is,

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

in which $\|\varphi\|_{0,1}$ is the $C^{0,1}$-norm of φ. Denote $q_j(\eta, y) = \partial^j Q\{g^{-1}(\eta), y\}/\partial \eta^j$, so that

$$q_1(\eta, y) = \frac{\partial}{\partial \eta} Q\{g^{-1}(\eta), y\} = -\{y - g^{-1}(\eta)\}\rho_1(\eta),$$

$$q_2(\eta, y) = \frac{\partial^2}{\partial \eta^2} Q\{g^{-1}(\eta), y\} = \rho_2(\eta) - \{y - g^{-1}(\eta)\}\rho_1'(\eta),$$

where $\rho_j(\eta) = \{\dot{g}^{-1}(\eta)\}^j / V\{g^{-1}(\eta)\}$.

A.1. Assumptions.
Throughout the paper, we assume the following regularity conditions:

(C1) The joint density of $\mathbf{X}$, denoted by $f(\mathbf{x})$, is absolutely continuous, and there exist constants $0 < c_f \leq C_f < \infty$, such that $c_f \leq \min_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq \max_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq C_f$.

(C2) The function V is twice continuously differentiable, and the link function g is three times continuously differentiable. The function $q_2(\eta, y) < 0$ for $\eta \in R$ and y in the range of the response variable.

(C3) For $1 \leq \ell \leq p$, $1 \leq k \leq d$, $\alpha_{\ell k}^{(r-1)}(x_k) \in C^{0,1}[0, 1]$, for given integer $r \geq 1$. The spline order satisfies $q \geq r$.

(C4) Let $\varepsilon_i = Y_i - \mu(\mathbf{X}_i, \mathbf{T}_i)$, $1 \leq i \leq n$. The random variables $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i | \mathbf{X}_i, \mathbf{T}_i) = \sigma^2(\mathbf{X}_i, \mathbf{T}_i)$. Furthermore, their tail probabilities satisfy $P(|\varepsilon_i| > x) < K \exp(-Cx^2)$, $i = 1, \ldots, n$, for all $x \geq 0$ and for some positive constants C and K.

(C5) The eigenvalues of $E(\mathbf{T}_{I_1} \mathbf{T}_{I_1}^{\mathrm{T}} | \mathbf{X} = \mathbf{x})$, where $\mathbf{T}_{I_1} = (T_\ell, \ell \in I_1)^{\mathrm{T}}$, are uniformly bounded away from 0 and ∞ for all $\mathbf{x} \in [0, 1]^d$. There exist constants $0 < c_1 < C_1 < \infty$, such that $c_1 \leq E(T_\ell^2 | \mathbf{X} = \mathbf{x}) \leq C_1$, for all $\mathbf{x} \in [0, 1]^d$, $\ell \in I_2$.

Conditions (C1)–(C5) are standard conditions for nonparametric estimation. Condition (C1) is the same as condition (C1) in Xue and Yang (2006) and condition (C5) in Xue and Liang (2010). The first condition in (C2) gives the assumptions on V and the link function g, which can be found in condition (E) of Lam and Fan (2008). The second condition in (C2) guarantees that the negative quasi-likelihood function $Q\{g^{-1}(\eta), y\}$ is convex in $\eta \in R$, which is also given in condition (D) of Lam and Fan (2008) and (a) of condition 1 in Carroll et al. (1997). Condition (C3) is typical for polynomial spline smoothing; see the same condition given in Section 5.2 of Huang (2003). Condition (C4) is the same as assumption (A2) given in Huang, Horowitz and Wei (2010). Condition (C5) is given in condition (C5) of Xue and Liang (2010) and condition (A5) in Ma and Yang (2011b).

A.2. Preliminary lemmas.
Define $\alpha_\ell^0(\mathbf{x}) = \sum_{k=1}^d \alpha_{\ell k}^0(x_k) = B(\mathbf{x})^{\mathrm{T}} \boldsymbol{\gamma}_\ell$, where $\alpha_{\ell k}^0(x_k)$ is defined in (6). Let $\boldsymbol{\gamma}_{I_1} = (\boldsymbol{\gamma}_\ell : \ell \in I_1)^{\mathrm{T}}$. To prove Theorem 1, we next

define the oracle estimator of $\boldsymbol{\gamma}_{I_1}$ by minimizing the penalized negative quasi-likelihood with all irrelevant predictors eliminated as such

$$(24) \quad L_n(\boldsymbol{\gamma}_{I_1}) = \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{\ell \in I_1} B(\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{\gamma}_\ell T_\ell\right\}, Y_i\right] + n\lambda_n \sum_{\ell \in I_1} w_{n\ell} \|\boldsymbol{\gamma}_\ell\|_2,$$

so that $\widehat{\boldsymbol{\gamma}}_{I_1}^0 = (\widehat{\boldsymbol{\gamma}}_\ell^0 : \ell \in I_1)^{\mathrm{T}} = \arg\min_{\boldsymbol{\gamma}_{I_1}} L_n(\boldsymbol{\gamma}_{I_1})$. Define $\widehat{\boldsymbol{\gamma}}_{I_2}^0 = (\widehat{\boldsymbol{\gamma}}_\ell^0 : \ell \in I_2)^{\mathrm{T}}$ with $\widehat{\boldsymbol{\gamma}}_\ell^0 \equiv \mathbf{0}_{dJ_n+1}$ for $\ell \in I_2$, where $\mathbf{0}_{dJ_n+1}$ is a $(dJ_n + 1)$-dimensional zero vector. We next present several lemmas, whose detailed proofs are given in the online supplementary materials [Ma et al. (2015)]. Lemma A.1 is used for the proof of Theorem 1, while Lemma A.2 is needed in the proof of Theorem 3.

LEMMA A.1. *Under the conditions of Theorem* 1, *one has*

$$(25) \quad \|\widehat{\boldsymbol{\gamma}}_{I_1}^0 - \boldsymbol{\gamma}_{I_1}\|_2 = O_p(\lambda_n \|w_{n,I_1}\| + n^{-1/2} J_n^{1/2} + J_n^{-r}),$$

and as $n \to \infty$,

$$(26) \quad P\{\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_{I_1}^{0\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_{I_2}^{0\mathrm{T}})^{\mathrm{T}}\} \to 1.$$

LEMMA A.2. *Under conditions* (C1)–(C5) *and Assumptions* 1–3,

$$(27) \quad \|\widehat{\boldsymbol{\gamma}}_{,1}^{\mathcal{S}} - \widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}}\|_\infty = O_p\left(\sqrt{\log n/(J_n^{\mathcal{S}} n)} + (J_n^{\mathcal{S}})^{-1/2} (J_n^{\mathrm{ini}})^{-r}\right).$$

A.3. Proof of Theorem 1. By (25) and (26),

$$\sum_{\ell \in I_1} \|\widehat{\alpha}_\ell - \alpha_\ell\| \asymp \|\widehat{\boldsymbol{\gamma}}_{I_1} - \boldsymbol{\gamma}_{I_1}\|_2 = O_p(\lambda_n \|w_{n,I_1}\| + n^{-1/2} J_n^{1/2} + J_n^{-r}),$$

$$P(\|\widehat{\alpha}_\ell\| > 0, \ell \in I_1 \text{ and } \|\widehat{\alpha}_\ell\| = 0, \ell \in I_2) \to 1.$$

A.4. Proof of Theorem 2. Let $\boldsymbol{\gamma}_{,1} = (\boldsymbol{\gamma}_{\ell 1}, \ell \in \widehat{I}_1)^{\mathrm{T}}$, where $\boldsymbol{\gamma}_{\ell 1}$ is defined in (7). By Taylor's expansion, from (10), one has

$$\widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}} - \boldsymbol{\gamma}_{,1} = \left[\sum_{i=1}^{n} Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i^*)\}^2 \Big/ \sigma_i^2\right]^{-1}$$

$$\times \left[\sum_{i=1}^{n} Z_{i,1} \{Y_i - g^{-1}(\eta_i^0)\} (\dot{g}^{-1}(\eta_i^0)/\sigma_i^2)\right],$$

where $\eta_i^0 = \sum_{\ell=1}^{p} \{\alpha_{\ell 0} + \sum_{k=2}^{d} \alpha_{\ell k}(X_{ik})\} T_{i\ell} + \sum_{\ell=1}^{p} B^{\mathcal{S}}(x_1)^{\mathrm{T}} \boldsymbol{\gamma}_{\ell 1} T_{i\ell}$ and

$$\eta_i^* = \sum_{\ell=1}^{p} \left\{\alpha_{\ell 0} + \sum_{k=2}^{d} \alpha_{\ell k}(X_{ik})\right\} T_{i\ell} + \sum_{\ell=1}^{p} B^{\mathcal{S}}(x_1)^{\mathrm{T}} \boldsymbol{\gamma}_{\ell 1}^* T_{i\ell},$$

where $\boldsymbol{\gamma}_{,1}^* = (\boldsymbol{\gamma}_{\ell 1}^*, \ell \in \widehat{I}_1)^{\mathrm{T}} \in (\boldsymbol{\gamma}_{,1}, \widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}})$. Following similar reasoning as the proofs for (25), we have $\|\widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}} - \boldsymbol{\gamma}_{,1}\|_2 = o_p(1)$. Then $\widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}} - \boldsymbol{\gamma}_{,1} = (\widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}} + \widehat{\boldsymbol{\gamma}}_{,1\mu}^{\mathrm{OR}}) + o_p(1)$, where

$$\widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}} = \left[\sum_{i=1}^n Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2 \Big/ \sigma_i^2\right]^{-1} \left[\sum_{i=1}^n Z_{i,1} \varepsilon_i \{\dot{g}^{-1}(\eta_i)/\sigma_i^2\}\right],$$

$$(28) \quad \widehat{\boldsymbol{\gamma}}_{,1\mu}^{\mathrm{OR}} = \left[\sum_{i=1}^n Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2 \Big/ \sigma_i^2\right]^{-1}$$

$$\times \left[\sum_{i=1}^n Z_{i,1} \{g^{-1}(\eta_i) - g^{-1}(\eta_i^0)\}\{\dot{g}^{-1}(\eta_i)/\sigma_i^2\}\right].$$

Therefore, $\mathrm{var}(\widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}}|\mathbb{X}, \mathbb{T}) = [\sum_{i=1}^n Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2/\sigma_i^2]^{-1}$. By Theorem 5.4.2 of DeVore and Lorentz (1993), for sufficiently large n, there exist constants $0 < c_B \le C_B < \infty$, such that $c_B \mathbf{I}_{J_n^{\mathcal{S}} \times J_n^{\mathcal{S}}} \le E(B_1^{\mathcal{S}}(X_{i1}) B_1^{\mathcal{S}}(X_{i1})^{\mathrm{T}}) \le C_B \mathbf{I}_{J_n^{\mathcal{S}} \times J_n^{\mathcal{S}}}$. By condition (C5), for n large enough, there are constants $0 < C_T, C' < \infty$, such that

$$E[Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2/\sigma_i^2]$$
$$\le C' E[\{B_1^{\mathcal{S}}(X_{i1}) B_1^{\mathcal{S}}(X_{i1})^{\mathrm{T}}\} \otimes \{E(T_\ell T_{\ell'}|\mathbf{X})\}_{\ell,\ell' \in \widehat{I}_1}]$$
$$\le C C_T s^* E\{B_1^{\mathcal{S}}(X_{i1}) B_1^{\mathcal{S}}(X_{i1})^{\mathrm{T}}\} \otimes \mathbf{I}_{s^* \times s^*} \le C' C_T C_B s^* \mathbf{I}_{J_n^{\mathcal{S}} \times J_n^{\mathcal{S}}} \otimes \mathbf{I}_{s^* \times s^*}$$
$$= C s^* \mathbf{I}_{J_n^{\mathcal{S}} s^* \times J_n^{\mathcal{S}} s^*},$$

where $C = C' C_T C_B$. Similarly, we have $E[Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2/\sigma_i^2] \ge c \mathbf{I}_{J_n^{\mathcal{S}} s^* \times J_n^{\mathcal{S}} s^*}$ for some constant $0 < c < \infty$. Thus, following the same reasoning as the proof for (S.5) in the supplementary materials [Ma et al. (2015)], we have with probability 1, for $n \to \infty$,

$$C^{-1}(s^*)^{-1} n^{-1} \mathbf{I}_{J_n^{\mathcal{S}} s^* \times J_n^{\mathcal{S}} s^*} \le \left[\sum_{i=1}^n Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2 \Big/ \sigma_i^2\right]^{-1}$$

$$(29)$$

$$\le c^{-1} n^{-1} \mathbf{I}_{J_n^{\mathcal{S}} s^* \times J_n^{\mathcal{S}} s^*}.$$

By the Lindeberg central limit theorem, it can be proved that

$$(30) \qquad \mathbf{a}^{\mathrm{T}} \sigma_n^{-1}(x_1) \{B^{\mathcal{S}}(x_1) \widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}}\} \to N(0, 1),$$

for any $\mathbf{a} \in R^{s^*}$ with $\|\mathbf{a}\|_2 = 1$. Since $\mathbf{a}^{\mathrm{T}} \sigma_n^{-1}(x_1) \{\widehat{\alpha}_1^{\mathrm{OR}}(x_1) - b_1(x_1)\} = \mathbf{a}^{\mathrm{T}} \sigma_n^{-1}(x_1) \{B^{\mathcal{S}}(x_1) \widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}}\} + o_p(1)$, by (30) and Slutsky's theorem, we have

$$(31) \qquad \mathbf{a}^{\mathrm{T}} \sigma_n^{-1}(x_1) \{\widehat{\alpha}_1^{\mathrm{OR}}(x_1) - b_1(x_1)\} \to N(0, 1).$$

By (28) and (29), with probability approaching 1,

$$\sum_{\ell \in I_1} \|\widehat{\alpha}_{\ell 1}^{\mathrm{OR}} - b_{\ell 1}\|^2 \asymp \|\widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}}\|_2^2$$

$$\leq c^{-2} n^{-2} \left[\sum_{i=1}^n \varepsilon_i Z_{i,1}^{\mathrm{T}} \big(\dot{g}^{-1}(\eta_i)/\sigma_i^2\big)\right]\left[\sum_{i=1}^n Z_{i,1} \varepsilon_i \big(\dot{g}^{-1}(\eta_i)/\sigma_i^2\big)\right]$$

$$\asymp c^{-2} n^{-1} E\big[Z_{i,1}^{\mathrm{T}} Z_{i,1} \{\dot{g}^{-1}(\eta_i)\}^2/\sigma_i^2\big] \asymp s^* J_n^{\mathcal{S}} n^{-1};$$

$$\|\mathbf{a}^{\mathrm{T}}(\widehat{\alpha}_1^{\mathrm{OR}} - b_1)\|^2 \leq C_a \|\widehat{\boldsymbol{\gamma}}_{,1e}^{\mathrm{OR}}\|_2^2 \leq C_a c^{-1} n^{-2}\left(\sum_{i=1}^n \varepsilon_i Z_{i,1}^{\mathrm{T}}\right)\left(\sum_{i=1}^n Z_{i,1}\varepsilon_i\right)$$

$$\asymp C_a c^{-1} n^{-1} E\big(Z_{i,1}^{\mathrm{T}} Z_{i,1}\big) \asymp s^* J_n^{\mathcal{S}} n^{-1}.$$

Since $\sup_{x_1 \in [0,1]} |\alpha_{\ell 1}(x_1) - B_1^{\mathcal{S}}(x_1)^{\mathrm{T}} \boldsymbol{\gamma}_{\ell 1}| = O\{(J_n^{\mathcal{S}})^{-r}\}$, it can be proved that $\|\mathbf{a}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{,1\mu}^{\mathrm{OR}}\| \leq \|\widehat{\boldsymbol{\gamma}}_{,1\mu}^{\mathrm{OR}}\|_2 = O_p\{(s^*)^{1/2}(J_n^{\mathcal{S}})^{-r}\}$, and $\|\mathbf{a}^{\mathrm{T}}(b_1 - \alpha_1^0)\| \asymp \|\mathbf{a}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{,1\mu 2}^{\mathrm{OR}}\| = O_p\{(s^*)^{1/2}(J_n^{\mathcal{S}})^{-r}\}$. Hence

$$\|\mathbf{a}^{\mathrm{T}}(b_1 - \alpha_1)\| \leq \|\mathbf{a}^{\mathrm{T}}(b_1 - \alpha_1^0)\| + \|\mathbf{a}^{\mathrm{T}}(\alpha_1^0 - \alpha_1)\| = O_p\{s^*(J_n^{\mathcal{S}})^{-r}\}.$$

By (31), $\{\mathbf{e}_\ell^{\mathrm{T}} \sigma_n^2(x_1) \mathbf{e}_\ell\}^{-1/2}\{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - b_{\ell 1}\} \to N(0,1)$, and $\sup_{\ell \in \widehat{I}_1} |\widehat{\alpha}_{\ell 0}^{\mathrm{OR}} - \alpha_{\ell 0}| = O_p(n^{-1/2})$ follows from the central limit theorem.

A.5. Proof of Theorem 3. By (27) in Lemma A.2,

$$\sup_{x_1 \in [0,1]} \|\widehat{\alpha}_1^{\mathcal{S}}(x_1) - \widehat{\alpha}_1^{\mathrm{OR}}(x_1)\|_\infty \leq \sup_{x_1 \in [0,1]} \sum_{j=1}^{J_n^{\mathcal{S}}} |B_{j,1}^{\mathcal{S}}(x_1)| \|\widehat{\boldsymbol{\gamma}}_{,1}^{\mathcal{S}} - \widehat{\boldsymbol{\gamma}}_{,1}^{\mathrm{OR}}\|_\infty.$$

The right-hand side is bounded by $O_p\{(n^{-1}\log n)^{1/2} + (J_n^{\mathrm{ini}})^{-r}\}$. $\|\widehat{\boldsymbol{\alpha}}_0^{\mathcal{S}} - \widehat{\boldsymbol{\alpha}}_0^{\mathrm{OR}}\|_2 = o_p(n^{-1/2})$ can be proved following the same procedure and thus omitted. By (29), with probability approaching 1, for large enough n, for any $x_1 \in [0,1]$, and $\mathbf{a} \in R^{s^*}$ with $\|\mathbf{a}\|_2 = 1$, one has

$$\mathbf{a}^{\mathrm{T}} \sigma_n^2(x_1) \mathbf{a} \leq c_Z^{-1} n^{-1} \mathbf{a}^{\mathrm{T}} B^{\mathcal{S}}(x_1) B^{\mathcal{S}}(x_1)^{\mathrm{T}} \mathbf{a} \leq c^{-1} J_n^{\mathcal{S}} n^{-1} \mathbf{a}^{\mathrm{T}} \mathbf{a},$$

$$\mathbf{a}^{\mathrm{T}} \sigma_n^2(x_1) \mathbf{a} \geq C_Z^{-1}(s^*)^{-1} n^{-1} \mathbf{a}^{\mathrm{T}} B^{\mathcal{S}}(x_1) B^{\mathcal{S}}(x_1)^{\mathrm{T}} \mathbf{a} \geq C^{-1} J_n^{\mathcal{S}}(s^*)^{-1} n^{-1} \mathbf{a}^{\mathrm{T}} \mathbf{a},$$

where $\sigma_n^2(x_1)$ is defined in (12). Thus

$$\sup_{x_1 \in [0,1]} |\mathbf{a}^{\mathrm{T}} \sigma_n^{-1}(x_1)\{\widehat{\alpha}_1^{\mathcal{S}}(x_1) - \widehat{\alpha}_1^{\mathrm{OR}}(x_1)\}|$$

$$\leq \sup_{x_1 \in [0,1]} \|\sigma_n^{-1}(x_1)\|_2 \|\widehat{\alpha}_1^{\mathcal{S}}(x_1) - \widehat{\alpha}_1^{\mathrm{OR}}(x_1)\|_2$$

$$= O_p\big[s^*\{(\log n/J_n^{\mathcal{S}})^{1/2} + (n/J_n^{\mathcal{S}})^{1/2}(J_n^{\mathrm{ini}})^{-r}\}\big] = o_p(1).$$

A.6. Proof of Theorem 4. Using the strong approximation lemma given in Theorem 2.6.7 of Csörgő and Révész (1981), we can prove by the same procedure as Lemma A.7 in Ma, Yang and Carroll (2012) that

$$(32) \qquad \sup_{x_1 \in [0,1]} |\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - b_{\ell 1}(x_1) - \widehat{\alpha}_{\ell 1,\varepsilon}^0(x_1)| = o_{\mathrm{a.s.}}(n^t)$$

for some $t < -r/(2r+1) < 0$, where $\widehat{\alpha}_{\ell 1,\varepsilon}^0(x_1)$ is

$$\mathbf{e}_\ell^{\mathrm{T}} B^{\mathcal{S}}(x_1) \left[\sum_{i=1}^n Z_{i,1} Z_{i,1}^{\mathrm{T}} \{\dot{g}^{-1}(\eta_i)\}^2 \Big/ \sigma_i^2 \right]^{-1} \left[\sum_{i=1}^n Z_{i,1} e_i \{\dot{g}^{-1}(\eta_i)/\sigma_i^2\} \right],$$

and $e_i, 1 \le i \le n$, are i.i.d. $N(0,1)$ independent of $Z_{i,1}$. For $\sigma_n^2(x_1)$ defined in (12) and $\sigma_{n1}(x_1) \asymp (J_n^{\mathcal{S}}/n)^{1/2}\{1 + o_p(1)\}$ uniformly in $x_1 \in [0,1]$. By (32), $J_n^{\mathcal{S}} \asymp n^{1/(2r+1)}$ and $t < -r/(2r+1) < 0$, we have

$$\sup_{x_1 \in [0,1]} |\{\log(L_n+1)\}^{-1/2} \sigma_{n1}^{-1}(x_1)\{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - b_{\ell 1}(x_1) - \widehat{\alpha}_{\ell 1,\varepsilon}^0(x_1)\}|$$

$$(33) \qquad = o_{\mathrm{a.s.}}(\{\log(L_n+1)\}^{-1/2}(n/J_n^{\mathcal{S}})^{1/2} n^t)$$

$$= o_{\mathrm{a.s.}}(\{\log(L_n+1)\}^{-1/2} n^{r/(2r+1)-t}) = o_{\mathrm{a.s.}}(1).$$

Define $\eta(x_1) = \sigma_{n1}^{-1}(x_1)\widehat{\alpha}_{\ell 1,\varepsilon}^0(x_1)$. It is apparent that $\mathcal{L}\{\eta(\xi_J)|Z_{i,1}, 1 \le i \le n\} = N(0,1)$, so $\mathcal{L}\{\eta(\xi_J)\} = N(0,1)$ for $0 \le J \le L_n$. Moreover, the eigenvalues of $(EZ_{i,1}Z_{i,1}^{\mathrm{T}})^{-1} \asymp J_n^{\mathcal{S}}$. Then with probability approaching 1, for $J \ne J'$,

$$|E\{\eta(\xi_J)\eta(\xi_{J'})\}| \asymp (n/J_n^{\mathcal{S}})n^{-1}|\mathbf{e}_\ell^{\mathrm{T}} B^{\mathcal{S}}(\xi_J)(EZ_{i,1}Z_{i,1}^{\mathrm{T}})^{-1} B^{\mathcal{S}}(\xi_{J'})^{\mathrm{T}}\mathbf{e}_\ell|$$

$$\asymp |\mathbf{e}_\ell^{\mathrm{T}} B^{\mathcal{S}}(\xi_J) B^{\mathcal{S}}(\xi_{J'})^{\mathrm{T}}\mathbf{e}_\ell| = \sum_{j=1}^{J_n^{\mathcal{S}}} B_{j,1}^{\mathcal{S}}(\xi_J) B_{j,1}^{\mathcal{S}}(\xi_{J'})$$

and $\sum_{j=1}^{J_n^{\mathcal{S}}} B_{j,1}^{\mathcal{S}}(\xi_J) B_{j,1}^{\mathcal{S}}(\xi_{J'}) \asymp C$ for a constant $0 < C < \infty$ when $|j_J - j_{J'}| \le (q-1)$ and $\sum_{j=1}^{J_n^{\mathcal{S}}} B_{j,1}^{\mathcal{S}}(\xi_J) B_{j,1}^{\mathcal{S}}(\xi_{J'}) = 0$ when $|j_J - j_{J'}| > (q-1)$, in which j_J denotes the index of the knot closest to ξ_J from the left. Therefore, by $L_n \asymp J_n^{\mathcal{S}}$, there exist constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$ such that with probability approaching 1, for $J \ne J'$, $|E\{\eta(\xi_J)\eta(\xi_{J'})\}| \le C_1^{-|j_J - j_{J'}|} \le C_2^{-|J-J'|}$. By Lemma A1 given in Ma and Yang (2011a), we have

$$\lim_{n \to \infty} P\left\{ \sup_{0 \le J \le L_n} |\{2\log(L_n+1)\}^{-1/2}\eta(\xi_J)| \le d_{N_n}(\alpha) \right\} = 1 - \alpha,$$

and hence

$$(34) \quad \lim_{n \to \infty} P\left\{ \sup_{x_1 \in \mathcal{S}_{n,1}} |\{2\log(L_n+1)\}^{-1/2}\sigma_{n1}^{-1}(x_1)\widehat{\alpha}_{\ell 1,\varepsilon}^0(x_1)| \le d_{N_n}(\alpha) \right\} = 1 - \alpha.$$

Furthermore, according to the result on page 149 of de Boor (2001), we have

$$
\begin{aligned}
(35) \quad &\sup_{x_1 \in [0,1]} \left| \{\log(L_n + 1)\}^{-1/2} \sigma_{n1}^{-1}(x_1) \{b_{\ell 1}(x_1) - \alpha_{\ell 1}(x_1)\} \right| \\
&= O_p\big(\{\log(L_n + 1)\}^{-1/2} (n/J_n^{\mathcal{S}})^{1/2} (J_n^{\mathcal{S}})^{-r}\big) = o_p(1).
\end{aligned}
$$

Moreover, $\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - \alpha_{\ell 1}(x_1) = \widehat{\alpha}_{\ell 1, \varepsilon}^{0}(x_1) + \{\widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - b_{\ell 1}(x_1) - \widehat{\alpha}_{\ell 1, \varepsilon}^{0}(x_1)\} + \{b_{\ell 1}(x_1) - \alpha_{\ell 1}(x_1)\}$. Hence by (33) and (35), we have

$$
\begin{aligned}
(36) \quad &\lim_{n \to \infty} P\left\{ \sup_{x_1 \in S_{n,1}} \{\log(L_n + 1)\}^{-1/2} \sigma_{n1}^{-1}(x_1) \big| \widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) - \alpha_{\ell 1}(x_1) \big| \leq d_{N_n}(\alpha) \right\} \\
&= \lim_{n \to \infty} P\left\{ \sup_{x_1 \in S_{n,1}} \{\log(L_n + 1)\}^{-1/2} \sigma_{n1}^{-1}(x_1) \big| \widehat{\alpha}_{\ell 1, \varepsilon}^{0}(x_1) \big| \leq d_{N_n}(\alpha) \right\} \\
&= 1 - \alpha,
\end{aligned}
$$

where the last step follows from (34). By the oracle property given in Theorem 3, and $J_n^{\mathcal{S}} \asymp n^{1/(2r+1)}$ and $n^{1/(2r+1)} \ll J_n^{\mathrm{ini}}$, we have

$$
\begin{aligned}
(37) \quad &\sup_{x_1 \in [0,1]} \{\log(L_n + 1)\}^{-1/2} \sigma_{n1}^{-1}(x_1) \big| \widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) - \widehat{\alpha}_{\ell 1}^{\mathrm{OR}}(x_1) \big| \\
&= O_p\big[\log(L_n + 1)^{-1/2} (n/J_n^{\mathcal{S}})^{1/2} (n^{-1} \log n)^{1/2} + (J_n^{\mathrm{ini}})^{-r}\big] = o_p(1).
\end{aligned}
$$

Therefore, by (36) and (37), we have

$$
\lim_{n \to \infty} P\left\{ \sup_{x_1 \in S_{n,1}} \{\log(L_n + 1)\}^{-1/2} \sigma_{n1}^{-1}(x_1) \big| \widehat{\alpha}_{\ell 1}^{\mathcal{S}}(x_1) - \alpha_{\ell 1}(x_1) \big| \leq d_{N_n}(\alpha) \right\} = 1 - \alpha,
$$

and hence the result in Theorem 4 is proved.

SUPPLEMENTARY MATERIAL

Supplemental materials for "Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates" (DOI: 10.1214/15-AOS1344SUPP; .pdf). The supplementary material presents additional numerical results and the proofs of Lemmas A.1 and A.2.

REFERENCES

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

CHEVERUD, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity (Edinb)* **87** 52–58.

CLAESKENS, G. and VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** 1852–1884. MR2036392

CSÖRGŐ, M. and RÉVÉSZ, P. (1981). *Strong Approximations in Probability and Statistics*. Academic Press, New York. MR0666546

DAWBER, T. R., MEADORS, G. F. and MOORE, F. E. (1951). Epidemiological approaches to heart disease: The Framingham 660 study. *American Journal of Public Health* **41** 279–286.

DE BOOR, C. (2001). *A Practical Guide to Splines*, revised ed. *Applied Mathematical Sciences* **27**. Springer, New York. MR1900298

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. *Grundlehren der Mathematischen Wissenschaften* **303**. Springer, Berlin. MR1261635

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478

HALL, P. and TITTERINGTON, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27** 228–254. MR0971184

HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796. MR1105844

HOROWITZ, J., KLEMELÄ, J. and MAMMEN, E. (2006). Optimal estimation in additive regression models. *Bernoulli* **12** 271–298. MR2218556

HOROWITZ, J. L. and MAMMEN, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32** 2412–2443. MR2153990

HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

KNUTSON, K. L. (2012). Does inadequate sleep play a role in vulnerability to obesity? *Am. J. Hum. Biol.* **24** 361–371.

LAM, C. and FAN, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36** 2232–2260. MR2458186

LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048

LIAN, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statist. Sinica* **22** 1563–1588. MR3027099

LIU, R. and YANG, L. (2010). Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory* **26** 29–59. MR2587102

LIU, R., YANG, L. and HÄRDLE, W. K. (2013). Oracally efficient two-step estimation of generalized additive model. *J. Amer. Statist. Assoc.* **108** 619–631. MR3174646

MA, S. and YANG, L. (2011a). A jump-detecting procedure based on spline estimation. *J. Nonparametr. Stat.* **23** 67–81. MR2780816

MA, S. and YANG, L. (2011b). Spline-backfitted kernel smoothing of partially linear additive model. *J. Statist. Plann. Inference* **141** 204–219. MR2719488

MA, S., YANG, L. and CARROLL, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statist. Sinica* **22** 95–122. MR2933169

MA, S., CARROLL, R. J., LIANG, H. and XU, S. (2015). Supplement to "Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates." DOI:10.1214/15-AOS1344SUPP.

MEIER, L. and BÜHLMANN, P. (2007). Smoothing l_1-penalized estimators for high-dimensional time-course data. *Electron. J. Stat.* **1** 597–615. MR2369027

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

MURCRAY, C. E., LEWINGER, J. P. and GAUDERMAN, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169** 219–226.

NYHOLT, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74** 765–769.

RANDALL, J. C., WINKLER, T. M., KUTALIK, Z., BERNDT, S. I., JACKSON, A. U. et al. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLOS Genetics* **9** e1003500.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

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

WAREHAM, N. J., VAN SLUIJS, E. M. F. and EKELUND, U. (2005). Physical activity and obesity prevention: A review of the current evidence. *Proc Nutr Soc* **64** 229–247.

XUE, L. and LIANG, H. (2010). Polynomial spline estimation for a generalized additive coefficient model. *Scand. J. Stat.* **37** 26–46. MR2675938

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

ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

S. MA
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, RIVERSIDE
RIVERSIDE, CALIFORNIA 92521
USA
E-MAIL: shujie.ma@ucr.edu

R. J. CARROLL
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843
USA
AND
SCHOOL OF MATHEMATICAL SCIENCES
UNIVERSITY OF TECHNOLOGY
SYDNEY, BROADWAY NSW 2007
AUSTRALIA
E-MAIL: carroll@stat.tamu.edu

H. LIANG
DEPARTMENT OF STATISTICS
GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC 20052
USA
E-MAIL: hliang@gwu.edu

S. XU
CENTER FOR PLANT CELL BIOLOGY
UNIVERSITY OF CALIFORNIA, RIVERSIDE
RIVERSIDE, CA 92521
USA
E-MAIL: shizhong.xu@ucr.edu