Statistica Sinica **21** (2011), 1225-1248

ESTIMATION AND VARIABLE SELECTION FOR SEMIPARAMETRIC ADDITIVE PARTIAL LINEAR MODELS

Xiang Liu[1], Li Wang[2] and Hua Liang[1]

[1] *University of Rochester and* [2] *University of Georgia*

Abstract: Semiparametric additive partial linear models, containing both linear and nonlinear additive components, are more flexible than linear models, and they are more efficient compared to general nonparametric regression models because they reduce the "curse of dimensionality". In this paper, we propose a new estimation approach for these models, in which we use polynomial splines to approximate the additive nonparametric components and derive the asymptotic normality for the resulting estimators of the parameters. We also develop a variable selection procedure to identify significant linear components using the smoothly clipped absolute deviation penalty (SCAD), and we show that the SCAD-based estimators of non-zero linear components have an oracle property. Simulations are performed to examine the performance of our approach as compared to several other variable selection methods, such as the Bayesian Information Criterion and Least Absolute Shrinkage and Selection Operator (LASSO). The proposed approach is also applied to data from a nutritional epidemiology study.

Key words and phrases: BIC, LASSO, penalized likelihood, regression spline, SCAD.

1. Introduction

Additive partial linear models (APLMs) are a generalization of multiple linear regression models, and can be regarded as a special case of generalized additive nonparametric regression models (Hastie and Tibshirani (1990)) as well. APLMs allow an easier interpretation of the effect of each variable, and are preferable to completely nonparametric additive models since they combine both parametric and nonparametric components when it is believed that the response variable depends on some variables in a linear way but is nonlinearly related to the remaining independent variables.

Estimation and inference for APLMs have been well studied in literature (Stone (1985); Opsomer and Ruppert (1997)), with a backfitting algorithm generally used for estimation. Opsomer and Ruppert (1999) studied the asymptotics of the kernel-based backfitting estimators. Liang et al. (2008) suggested that a kernel-based estimation procedure is available for APLMs without an undersmoothing requirement, and applied this to study the relationship between environmental chemical exposures and semen quality. When there are multiple nonparametric terms, it is both useful and necessary that estimation and inference methods be efficient and easily implemented. Additionally, implementation should be able to be achieved in a commonly used computational environment like R. Kernel-based procedures (Opsomer and Ruppert (1999); Liang et al. (2008)) are intuitively attractive and theoretically justifiable, but computationally inexpedient; spline-based procedures (Li (2000)) are computationally expedient, but theoretically unreliable. Challenged by these demands, we propose approximating the nonparametric components with polynomial splines. As the models become linear, the resulting estimators for the linear components are easily calculated and, of most importance, still *asymptotically normal*.

Motivated by a dataset from a nutritional epidemiology project (details in Section 4), we study variable selection for APLMs. To the best of our knowledge, no variable selection procedures are available for APLMs. Best subset selection is commonly used to select significant variables in regression models. But it has two basic limitations. First, it is computationally infeasible to do subset selection when the number of predictors is large; second, it is extremely variable because of its inherent discreteness (Breiman (1996); Fan and Li (2001)). Stepwise selection is often used to reduce the number of candidate subsets. However, it still suffers from the high variability. Instead, Tibshirani (1996) proposed a regression method using the L_1 penalty, the LASSO, that is similar to ridge regression but can shrink some coefficients to 0, and thus implement variable selection. Fan and Li (2001) proposed a very general variable selection framework by using a smoothly clipped absolute deviation (SCAD) penalty. The choice of the SCAD penalty function encompasses the commonly used variable selection approaches as special cases (see Section 2.2 for details). Most importantly the SCAD-based approach has appealing statistical properties, as Fan and Li (2001) demonstrated. This approach has become popular and been widely studied by, for instance, Fan and Li (2002) for Cox models, Li and Liang (2008) for semiparametric models, and Liang and Li (2009) for partially linear models with measurement errors. Xie and Huang (2009) and Ni, Zhang, and Zhang (2009) studied variable selection for partially linear models with a divergent number of linear covariates, and established selection consistency and asymptotic normality. The former used polynomial splines and the latter used smoothing splines to approximate the nonparametric function. Since partially linear models have only one nonparametric component, they are not as flexible as APLM. In contrast to that in partially linear models, estimation or variable selection is much more difficult in APLM. Ravikumar et al. (2008, 2009) investigated high-dimensional nonparametric sparse additive models (SpAM), developed a new class of algorithms for estimation and discussed asymptotic properties of their estimators. SpAMs are more general but lack of the simplicity property of APLM, which are more appropriate when some covariates are not continuous.

In this paper we develop a SCAD-based variable selection procedure for APLMs combining the spline approximation. This combination overcomes a potential problem of how to define the objective function if a backfitting algorithm is used. Furthermore, employing a spline approximation can still allow our variable selection procedure to have the oracle property, the best theoretical performance of any procedure.

The rest of the article is organized as follows. Section 2 introduces the estimation and SCAD-based variable selection procedures for APLMs, and presents the theoretical results. Numerical comparisons and simulation studies are given in Section 3. Section 4 examines in detail the nutritional data to illustrate the procedure. Section 5 concludes with a discussion. Technical details are given in the Appendix.

2. Estimation and Variable Selection Procedure

Suppose that $\{(\boldsymbol{X}_1, \boldsymbol{Z}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Z}_n, Y_n)\}$ is an i.i.d. random sample of size n from the APLM

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{K} g_k(Z_k) + \varepsilon, \tag{2.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^{\mathrm{T}}$ are the linear and nonparametric components, $g_1, \ldots, g_K$ are unknown smooth functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is a vector of unknown parameters, and the model error ε has conditional mean zero and finite variance σ^2 given $(\boldsymbol{X}, \boldsymbol{Z})$. To ensure identifiability of the nonparametric functions, we assume that $E\{g_k(Z_k)\} = 0$ for $k = 1, \ldots, K$.

2.1. Spline approximation

Let $g_0 = g_{01}(z_1) + \cdots + g_{0K}(z_K)$ and $\boldsymbol{\beta}_0$ be the true additive function and parameter. For simplicity, we assume that the covariate Z_k is distributed on a compact interval $[a_k, b_k]$, $k = 1, \ldots, K$, and without loss of generality, we take all intervals $[a_k, b_k] = [0, 1]$, $k = 1, \ldots, K$. Under some smoothness assumptions, the g_{0k}'s can be well-approximated by spline functions. Let $\mathcal{S}_n$ be the space of polynomial splines on $[0, 1]$ of degree $\varrho \geq 1$. We introduce a knot sequence with J_n interior knots,

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

where J_n increases with sample size n at the precise order given in Condition (C4). Then $\mathcal{S}_n$ consists of functions ξ satisfying

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

(ii) for $\varrho \geq 2$, ξ is $\varrho - 1$ continuously differentiable on $[0,1]$.

Equally spaced knots are used here for simplicity. However other regular knot sequences can also be used, with similar asymptotic results. Let $h = 1/(J_n + 1)$ be the distance between neighboring knots.

We consider the additive spline estimates $\widehat{g}$ of g_0 based on the independent random sample $(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i)$, $i = 1, \ldots, n$. Let $\mathcal{G}_n$ be the collection of functions g with the additive form $g(\mathbf{z}) = g_1(z_1) + \cdots + g_K(z_K)$, where each component function $g_k \in \mathcal{S}_n$ and $\sum_{i=1}^n g_k(Z_{ik}) = 0$.

We would like to find a function $g \in \mathcal{G}_n$ and a value of $\boldsymbol{\beta}$ that minimize the following sum of squared residuals function

$$L(g, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n [Y_i - \{g(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}\}]^2, \quad g \in \mathcal{G}_n. \tag{2.2}$$

For the k-th covariate z_k, let $b_{j,k}(z_k)$ be the B-spline basis functions of degree ϱ. For any $g \in \mathcal{G}_n$, one can write

$$g(\boldsymbol{z}) = \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{b}(\mathbf{z}), \tag{2.3}$$

where $\boldsymbol{b}(\boldsymbol{z}) = \{b_{j,k}(z_k), j = -\varrho, \ldots, J_n, k = 1, \ldots, K\}^{\mathrm{T}}$, and the spline coefficient vector $\boldsymbol{\gamma} = \{\gamma_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\}^{\mathrm{T}}$. Thus the minimization problem in (2.2) is equivalent to finding $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to minimize

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n [Y_i - \{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}\}]^2. \tag{2.4}$$

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

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

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

For simplicity of notation, write $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Z})$. Let $m_0(\boldsymbol{T}) = g_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}_0$, $\Gamma(\boldsymbol{z}) = E(\boldsymbol{X} | \boldsymbol{Z} = \boldsymbol{z})$, $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \Gamma(\boldsymbol{Z})$, and $\mathbf{Q}^{\otimes 2} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}$ for any matrix or vector $\mathbf{Q}$. The result is that the estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is root-n consistent and asymptotically normal, although the convergence rate of the estimators of the nonparametric component g_0 is slower than root-n (Lemma A.4). The proof is in the Appendix.

Theorem 1. *Under the conditions* (C1)−(C5) *given in the Appendix,* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *converges to* $N(\mathbf{0}, \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1})$ *in distribution, where* $\mathbf{D} = E(\widetilde{\boldsymbol{X}}^{\otimes 2})$ *and* $\boldsymbol{\Sigma} = E(\varepsilon^2 \widetilde{\boldsymbol{X}}^{\otimes 2})$. *Furthermore, if* ε *and* $(\boldsymbol{X}, \boldsymbol{Z})$ *are independent,* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \rightarrow N\left(0, \sigma^2\mathbf{D}^{-1}\right)$, *where* $\sigma^2 = E\left(\varepsilon^2\right)$.

2.2. SCAD-penalty variable selection procedure

Penalized likelihood has been widely used for non- and semi-parametric models to trade off model complexity and estimation accuracy; comprehensive survey was given by Ruppert, Wand and Carroll (2003). The penalized objective function we use is

$$\mathcal{L}_P(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{b}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}\right]^2 + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{2.5}$$

where $p_{\lambda_j}(\cdot)$ is a penalty function with a tuning parameter λ_j that may be chosen by a data-driven method. See Liang and Li (2009) for a detailed discussion of the choice of the tuning parameter. Minimizing $\mathcal{L}_P(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\beta}$ results in a penalized least squares estimator $\widehat{\boldsymbol{\beta}}$. It is worth noting that the penalty functions and the tuning parameters are not necessarily the same for all coefficients. For instance, we wish to keep important variables in the final model, and therefore do not penalize their coefficients.

The use of (2.5) gives a general framework of variable selection for APLMs. Taking the penalty function to be the L_0-penalty (also called the entropy penalty in the literature), $p_{\lambda_j}(|\beta_j|) = 0.5\lambda_j^2 I\{|\beta_j| \neq 0\}$, where $I\{\cdot\}$ is an indicator function, we may extend the traditional variable selection criteria, including the AIC (Akaike (1973)), BIC (Schwarz (1978)), and RIC (Foster and George (1994)), for the APLM:

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) + \frac{n}{2}\sum_{j=1}^{d}\lambda_j^2 I\{|\beta_j| \neq 0\}, \tag{2.6}$$

as $\sum_{j=1}^{d} I\{|\beta_j| \neq 0\}$ is the size of the selected model. Specifically, the AIC, BIC, and RIC correspond to $\lambda_j \equiv \sqrt{2/n}\sigma$, $\sqrt{\log(n)/n}\sigma$, and $\sqrt{\log(d)/n}\sigma$, respectively. Note that Bridge regression (Frank and Friedman (1993)) is equivalent to the L_q-penalty $p_\lambda(|\beta_j|) = q^{-1}\lambda|\beta_j|^q$; the LASSO (Tibshirani (1996); Zou (2006)) corresponds to the L_1-penalty, and SCAD corresponds to the following smoothly clipped absolute deviation penalty function

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \leq |\beta| < \lambda, \\ \frac{(a^2-1)\lambda^2 - (|\beta| - a\lambda)^2}{2(a-1)}, & \text{if } \lambda \leq |\beta| < a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta| \geq a\lambda, \end{cases}$$

where $a = 3.7$. As demonstrated in Fan and Li (2001), SCAD is an improvement of LASSO in terms of modeling bias, and of bridge regression with $q < 1$ in terms of stability. It also has an oracle property.

We turn to the sampling properties of the resulting penalized least squares estimate. Let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{d0})^{\mathrm{T}} = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ be the true value of $\boldsymbol{\beta}$. Without loss of generality, assume that $\boldsymbol{\beta}_{10}$ consists of all nonzero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{20} = \mathbf{0}$. Let s denote the length of $\boldsymbol{\beta}_{10}$. Write $a_n = \max_{1\le j\le d}\{|p'_{\lambda_j}(|\beta_{j0}|)|, \beta_{j0} \neq 0\}$, $b_n = \max_{1\le j\le d}\{|p''_{\lambda_j}(|\beta_{j0}|)|, \beta_{j0} \neq 0\}$, $\kappa_n = \{p'_{\lambda_1}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \ldots, p'_{\lambda_s}(|\beta_{s0}|)\ \mathrm{sgn}(\beta_{s0})\}^{\mathrm{T}}$, and $\boldsymbol{\Sigma}_\lambda = \mathrm{diag}\{p''_{\lambda_1}(|\beta_{10}|), \ldots, p''_{\lambda_s}(|\beta_{s0}|)\}$. Denote $\boldsymbol{X}_1$ as the vector comprised by the first s elements of $\boldsymbol{X}$.

Theorem 2. *Suppose that* $a_n = O(n^{-1/2})$, $b_n \to 0$, *and* (C1)−(C5) *in the Appendix hold. Then* (I) *With probability approaching one, there exists a local minimizer* $\widehat{\boldsymbol{\beta}}$ *of* $\mathcal{L}_P(\boldsymbol{\beta}, \boldsymbol{\gamma})$ *such that* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$. (II) *If* $\lambda_j \to 0$, $n^{1/2}\lambda_j \to \infty$, *and*

$$\liminf_{n\to\infty} \liminf_{u\to 0^+} \frac{p'_{\lambda_j}(u)}{\lambda_j} > 0, \tag{2.7}$$

then, with probability approaching one, the root-n consistent estimator $\widehat{\boldsymbol{\beta}}$ *in* (I) *satisfies* (a) $\widehat{\boldsymbol{\beta}}_2 = 0$, *and* (b) $\widehat{\boldsymbol{\beta}}_1$ *has an asymptotic normal distribution*

$$\sqrt{n}\{E(\widetilde{\boldsymbol{X}}_1^{\otimes 2}) + \Sigma_\lambda\}[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{E(\widetilde{\boldsymbol{X}}_1^{\otimes 2}) + \boldsymbol{\Sigma}_\lambda\}^{-1}\kappa_n] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_s),$$

where $\boldsymbol{\Sigma}_s = \mathrm{Var}\left(\varepsilon\widetilde{\boldsymbol{X}}_1\right)$.

Theorem 2 indicates that the SCAD-penalty variable selection procedure can effectively identify the significant components, with the associated estimators holding the oracle property. The proof is in the Apprendix.

3. Simulation Studies

In this section, the finite sample performance of the proposed procedure is investigated by Monte Carlo simulations. We numerically compare estimation accuracy and complexity of models selected by SCAD, LASSO, and BIC. We use the local quadratic approximation algorithm of Fan and Li (2001) to implement the SCAD and LASSO procedures, and select the tuning parameter by generalized cross-validation (GCV) in both simulation studies and in the data example in Section 4.

Let $g_1(z) = 5\sin(4\pi z)$ and $g_2(z) = 100\{\exp(-3.25z) - 4\exp(-6.5z) + 3\exp(-9.75z)\}$. We generated 100 data sets consisting of $n = 60$, 100, and 200 observations from the model

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}) + \sigma\varepsilon,$$

Table 1. Simulation Results for Case (i)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.49	0.00	0.852	4.39	0.12	0.899	4.29	0.61	0.903
	lasso	3.38	0.00	0.882	3.49	0.02	0.750	3.41	0.31	0.723
	bic	4.66	0.00	0.869	4.76	0.14	0.948	4.57	0.86	0.969
	oracle	5.00	0.00	0.662	5.00	0.00	0.680	5.00	0.00	0.635
100	scad	4.45	0.00	0.838	4.44	0.03	0.870	4.35	0.32	0.947
	lasso	3.31	0.00	0.906	3.53	0.00	0.775	3.40	0.09	0.768
	bic	4.84	0.00	0.876	4.80	0.03	0.880	4.77	0.60	1.036
	oracle	5.00	0.00	0.717	5.00	0.00	0.704	5.00	0.00	0.706
200	scad	4.40	0.00	0.798	4.37	0.00	0.818	4.38	0.03	0.788
	lasso	3.27	0.00	0.884	3.37	0.00	0.829	3.37	0.00	0.797
	bic	4.91	0.00	0.803	4.88	0.00	0.916	4.90	0.06	0.772
	oracle	5.00	0.00	0.723	5.00	0.00	0.668	5.00	0.00	0.693

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$, $\sigma = 1, 3, 5$, and the components of $\boldsymbol{X}$ and ε are standard normal with $\boldsymbol{X}$ and ε independent. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$ for $i, j = 1, \cdots, 8$. We considered three cases: (i) $g(\mathbf{Z}) = g_1(Z_1)$; (ii) $g(\mathbf{Z}) = g_2(Z_2)$; and (iii) $g(\mathbf{Z}) = g_1(Z_1) + g_2(Z_2)$. Here Z_1 and Z_2 are independent uniformly distributed on $[0, 1]$ so in the first two cases there is only one nonparametric component, while in the third case there are two.

Cubic B-splines were used to approximate the nonparametric functions as described in Section 2.1. To determine the number of knots in the approximation, we examined several (say M) models, with the number of knots from 2 to 12 for each nonparametric component. Thus, $M = 11$ in both Case (i) and Case (ii), and $M = 11^2 = 121$ in Case (iii). In each case, the M linear prediction models were taken into account, and the model with the smallest median relative model error when compared to the full model, which includes all covariates, was taken as the selected model.

Simulation results are presented in Tables 1−3; the columns labeled "C" give the average number of the 5 zero coefficients correctly set to 0, the columns labeled "I" give the average number of the 3 nonzero coefficients incorrectly set to 0, and the columns labeled "MRME" give the median of relative model errors, which is defined as the ratio of model error comparing the selected model to the full model. Rows refer to procedures, where "Oracle" stands for the oracle estimates computed by using the true model $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + g(\mathbf{Z}) + \sigma\varepsilon$. The oracle estimates always set the 5 zero coefficients correctly to zero and do not set any of the 3 nonzero coefficients to zero.

The results for SCAD, BIC, and LASSO of correctly and incorrectly selected covariates show a similar pattern as those obtained by Fan and Li (2001)

Table 2. Simulation Results for Case (ii)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.44	0.00	0.774	4.48	0.14	0.937	4.32	0.69	1.028
	lasso	3.28	0.00	1.017	3.41	0.02	1.003	3.47	0.35	0.889
	bic	4.60	0.00	0.792	4.74	0.19	0.983	4.58	0.88	1.058
	oracle	5.00	0.00	0.673	5.00	0.00	0.674	5.00	0.00	0.662
100	scad	4.49	0.00	0.784	4.46	0.03	0.874	4.47	0.38	1.017
	lasso	3.58	0.00	1.044	3.50	0.00	0.996	3.58	0.11	0.963
	bic	4.85	0.00	0.784	4.76	0.03	0.907	4.78	0.61	1.041
	oracle	5.00	0.00	0.747	5.00	0.00	0.655	5.00	0.00	0.681
200	scad	4.40	0.00	0.768	4.31	0.00	0.805	4.31	0.03	0.870
	lasso	3.29	0.00	1.006	3.38	0.00	0.983	3.36	0.00	0.910
	bic	4.89	0.00	0.767	4.89	0.01	0.839	4.84	0.08	0.954
	oracle	5.00	0.00	0.716	5.00	0.00	0.644	5.00	0.00	0.677

Table 3. Simulation Results for Case (iii)

n	method	$\sigma = 1$			$\sigma = 3$			$\sigma = 5$		
		C	I	MRME	C	I	MRME	C	I	MRME
60	scad	4.43	0.00	0.924	4.39	0.28	1.045	4.37	0.74	1.010
	lasso	3.52	0.00	1.072	3.68	0.10	0.922	3.67	0.26	0.783
	bic	4.32	0.00	0.939	4.42	0.31	1.077	4.43	0.87	1.012
	oracle	5.00	0.00	0.802	5.00	0.00	0.802	5.00	0.00	0.752
100	scad	4.41	0.00	0.926	4.49	0.02	0.957	4.28	0.35	1.052
	lasso	3.58	0.00	1.028	3.58	0.00	0.964	3.56	0.09	0.883
	bic	4.60	0.00	0.939	4.77	0.05	0.977	4.66	0.65	1.112
	oracle	5.00	0.00	0.800	5.00	0.00	0.782	5.00	0.00	0.784
200	scad	4.44	0.00	0.881	4.45	0.00	0.953	4.45	0.06	0.973
	lasso	3.42	0.00	1.022	3.48	0.00	0.995	3.41	0.01	0.891
	bic	4.84	0.00	0.900	4.79	0.00	0.988	4.86	0.11	1.021
	oracle	5.00	0.00	0.821	5.00	0.00	0.806	5.00	0.00	0.797

for linear models. In all three cases, BIC performed the best in correctly setting coefficients to 0, followed by SCAD and LASSO. However, BIC had the highest average number of coefficients erroneously set to 0, followed by SCAD and LASSO. This indicates that BIC was the most aggressive method in terms of excluding variables, while LASSO was the most conservative and tended to include more variables.

As for the MRME, SCAD performed the best when the sample size was large or the error variance small , while LASSO performed the best when the sample size was small or the error variance large. The performance of BIC was worse than, although sometimes close to, SCAD.

Overall, SCAD and BIC had the best performances in our simulations. Compared to BIC, SCAD had the higher prediction accuracy by slightly increasing the model complexity. In other words, SCAD selected more variables to reduce prediction error. Furthermore, SCAD was much more computationally efficient than the best subset selection method using BIC.

4. A Nutritional Study

It is well known that there is a direct relationship between beta-carotene and cancers such as lung, colon, breast, and prostate cancer (Fairfield and Fletcher (2002)). Some observational epidemiological studies have shown that beta carotene cannot only effectively prevent cancer because beta carotene has powerful antioxidant properties, but can also help cleanse the body of free radicals that can cause cancer. Sufficient beta carotene supply can also strengthen the body's autoimmune system, making it more effective in fighting degenerative diseases such as cancer. Clinicians and nutritionists are therefore interested in the relationship between serum concentrations of beta-carotene and other factors such as age, smoking status, alcohol consumption, and dietary intake because this information may be potentially useful in clinical decision-making and individualization of therapy. For example, Nierenberg et al. (1989) found that dietary carotene and female were positively related to beta-carotene levels, while cigarette smoking and body mass index (BMI) were negatively related to beta-carotene levels. Age was not associated with beta-carotene levels to a statistically significant extent. Faure et al. (2006) recently found that beta-carotene concentration depends on gender, age, smoking status, dietary intake, and location of residence. Examination of this relationship therefore shows diverse results so far, and there is insufficient evidence to draw a convincing conclusion regarding the relationship between beta-carotene and these factors.

A closer investigation of these publications finds either a simple analysis of variance(ANOVA) method or linear models used to explore the relationship between beta-carotene and other factors to determine those that influence beta-carotene concentration. However, the use of advanced statistical techniques seems necessary in appropriately modeling the relationship. We examine a dataset from a nutritional epidemiology study where the interest is in the relationships between the plasma beta-carotene levels and personal characteristics, including AGE, GENDER, BMI, and other factors: CALORIES (number of calories consumed per day), FAT (grams of fat consumed per day), FIBER (grams of fiber consumed per day), ALCOHOL (number of alcoholic drinks consumed per week), CHOL (cholesterol consumed mg per day), BETADIET (dietary beta-carotene consumed mcg per day), SMOKE2 (smoking status [1 = former smoker, 0 = never smoked], and SMOKE3 (smoking status [1 = current smoker, 0 = never smoked]). There was one extremely high leverage point in alcohol consumption that was deleted prior to analysis. See Nierenberg et al. (1989) for a detailed description of the data. A general linear model was used to fit this dataset and the results are presented in the left panel of Table 4. These results indicate that only BMI, FIBER, GENDER, and SMOKE3 are statistically significant, while the other seven variables are not. However, a closer study shows that the relationship between the logarithm of beta-carotene levels and AGE and CHOL may be nonlinear. We therefore fitted the same dataset using the R function gam and found that the beta-carotene level seems to be linearly related to BMI, CALORIES, FAT, FIBER, ALCOHOL and BETADIET, but nonlinearly related to AGE and CHOL. Figure 1 shows the patterns of AGE and CHOL that indicate a positive correlation before 45 or after 65, and a slightly negative correlation between 45 and 65. Interestingly, we discern a concave curve in the pattern of CHOL.

We use an APLM and the proposed procedures to study the relationship between beta-carotene and other factors, assuming beta-carotene concentration depends linearly on covariates BMI, CALORIES, FAT, FIBER, ALCOHOL, BETADIET, GENDER, and SMOKE2/3 but is nonlinearly related to AGE and CHOL. We look to which linear covariates should be included in our model and appropriately fit the nonlinear unknown functions that can objectively reflect impact on beta-carotene level and avoid misleading conclusions. To this end, we used a model to fit the nutritional dataset and applied SCAD, LASSO, and BIC procedures for variable selection:

$$\begin{aligned}\log(\text{beta-carotene}) = \beta_0 &+ \beta_1\text{BMI} + \beta_2\text{CALORIES} + \beta_3\text{FAT} + \beta_4\text{FIBER}\\ &+\beta_5\text{BETADIET} + \beta_6\text{GENDER} + \beta_7\text{ALCOHOL} + \beta_8\text{SMOKE2}\\ &+\beta_9\text{SMOKE3} + g_1(\text{AGE}) + g_2(\text{CHOL}) + \varepsilon.\end{aligned}$$

To determine the number of knots in the cubic B-splines approximation of the nonparametric components AGE and CHOL, we examined the number of knots from 2 to 9 for each component and chose the number that gave the smallest relative mean squared error when compared to the full model. As a result, for the nonparametric component AGE, 2 and 5 knots were chosen using SCAD and LASSO, respectively; for the nonparametric component CHOL, 2 knots were used in both SCAD and LASSO. The tuning parameters selected by GCV were 0.035 and 0.015 for SCAD and LASSO, respectively.

The estimated coefficients and their standard errors are listed in the right panel of Table 4. SCAD found BMI, FIBER, BETADIET, GENDER, and SMOKE3 significant, while LASSO also identified FAT as being significant; BIC had it that only BMI, FIBER, and SMOKE3 are significant. The standard errors of non-zero coefficients based on the APLM are consistently smaller than

Table 4. Results for the nutritional study. Left panel: Estimated values, associated standard error, and P-value by using the ordinary least squares. Right panel: Estimates, associated standard errors of the coefficients using the APLM with the proposed variable selection procedures.

	LS				APLM		
	Est.	s.e	z value	Pr(> $\|z\|$)	SCAD (s.e.)	LASSO (s.e.)	BIC (s.e.)
BMI	-0.976	0.202	-4.829	$< 10^{-4}$	-0.947(0.189)	-0.948(0.173)	-1.001(0.188)
CALORIES	0	0	-0.457	0.648	0(0)	0(0)	0(0)
FAT	-0.002	0.003	-0.711	0.477	0(0)	-0.001(0.001)	0(0)
FIBER	0.027	0.012	2.352	0.019	0.021(0.007)	0.019(0.007)	0.025(0.008)
BETADIET	0.137	0.073	1.889	0.060	0.046(0.027)	0.101(0.051)	0(0)
GENDER	0.277	0.135	2.060	0.040	0.194(0.088)	0.201(0.096)	0(0)
ALCOHOL	0.043	0.048	0.901	0.368	0(0)	0(0)	0(0)
SMOKE2	-0.068	0.091	-0.742	0.458	0(0)	0(0)	0(0)
SMOKE3	-0.286	0.130	-2.191	0.029	-0.245(0.097)	-0.224(0.096)	-0.293(0.117)
AGE	0.005	0.003	1.724	0.086			
CHOL	-0.015	0.114	-0.133	0.894			

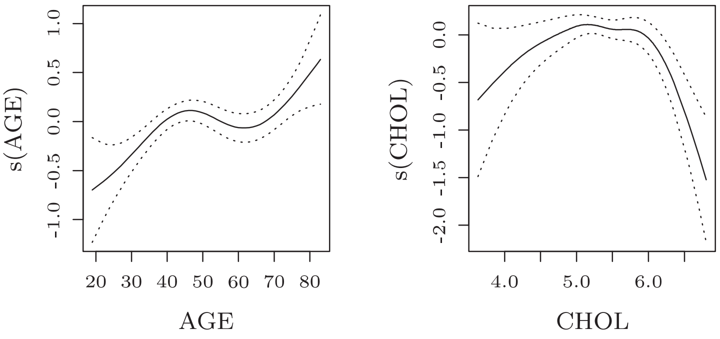

Figure 1. The patterns of AGE and CHOL with $\pm s.e.$ using the R function, gam, for the dataset from a nutritional study.

the corresponding ones based on the linear fit. The estimated values using different methods under the APLM setting are similar, but there are differences in magnitude. The estimated curves of the two nonparametric components, AGE and CHOL, are similar to those in Figure 1, and therefore are not shown here. It is worthwhile to mention that the effects of AGE and CHOL are not only significant, but also should not be described by linear functions.

5. Discussion

We proposed an effective routine using a regression spline technique, then used an advanced variable selection procedure to identify which linear predictors should be included in our model fitting. There are three principal advantages of our method: it avoids iterative algorithms and computational challenges;

the estimators of the linear components, which are of primary interest, are still asymptotically normal; the variable selection procedure has the oracle property. Combined with ideas of Liang and Li (2009), we believe a similar procedure can be developed for partially linear additive models with error-prone linear covariates. The approach possibly extends to generalized additive partial linear models and the situation with longitudinal data (Lin and Carroll (2001)). However, these extensions are by no means straightforward.

It appears possible, at least in principle, to extend our methods to cases in which the numbers of linear components and nonparametric components diverge. An alternative is a combination of the methods of Xie and Huang (2009) and Ravikumar et al. (2009). One main challenge is the establishment of the asymptotic properties of the methods. A detailed investigation of these issues is beyond the scope of this article.

An important question to nutritionists is whether available scientific data support an important role of beta-carotene in the prevention of pathologic conditions such as cancer. In this paper, we have proposed the use of an APLM to describe such a relationship since the APLM model can parsimoniously reflect the influence of covariates in linear or nonlinear forms.

Acknowledgement

Liu's Research was supported by Merck Quantitative Sciences Fellowship Program. Wang's Research was supported by NSF grant DMS-0905730. Liang's research was supported by NSF grants DMS-0806097 and DMS-1007167. The authors are grateful to an associate editor and two referees for valuable comments and suggestions that lead to a much improved paper.

Appendix

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

Following Stone (1985) and Huang (2003), for any measurable functions φ_1, φ_2 on $[0,1]^K$, we take the empirical inner product and the corresponding norm to be

$$\langle \varphi_1, \varphi_2 \rangle_n = n^{-1} \sum_{i=1}^n \varphi_1(\boldsymbol{Z}_i)\, \varphi_2(\boldsymbol{Z}_i), \quad \|\varphi\|_n^2 = n^{-1} \sum_{i=1}^n \varphi^2(\boldsymbol{Z}_i),$$

where $\{\boldsymbol{Z}_i\}$ is a sample from density f. If φ_1 and φ_2 are L^2-integrable, take the inner product

$$\langle \varphi_1, \varphi_2 \rangle = \int_{[0,1]^K} \varphi_1(\boldsymbol{z})\, \varphi_2(\boldsymbol{z})\, f(\boldsymbol{z}) d\boldsymbol{z}$$

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

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

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

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

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

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

Notice that finding the $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ that minimizes (2.4) is equivalent to finding the $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ that minimizes

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

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

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

In practice, basis $\{b_{j,k}, \; j = -\varrho + 1, \ldots, J_n, \; k = 1, \ldots, K\}^{\mathrm{T}}$ is used for data-analytic implementation, and (A.2) is convenient for asymptotic analysis.

A.1. Assumptions

The following conditions are necessary for Theorems 1 and 2. Let r be a positive integer and $\nu \in (0,1]$ be such that $p = r + \nu > 1.5$. Let $\mathcal{H}$ be the collection of functions g on $[0,1]$ whose r-th derivative, $g^{(r)}$ exists and satisfies the Lipschitz condition of order ν:

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

where and below C is a generic positive constant.

(C1) Each component function $g_{0k} \in \mathcal{H}$, $k = 1, \ldots, K$.

(C2) The distribution of $\boldsymbol{Z}$ is absolutely continuous and its density f is bounded away from zero and infinity on $[0,1]^K$.

(C3) The random vector $\boldsymbol{X}$ satisfies that for any vector $\boldsymbol{w} \in R^d$

$$c\left\|\boldsymbol{w}\right\|^2 \leq \boldsymbol{w}^{\mathrm{T}} E\left(\boldsymbol{X}^{\otimes 2} | \boldsymbol{Z} = \boldsymbol{z}\right) \boldsymbol{w} \leq C \left\|\boldsymbol{w}\right\|^2,$$

where c is a positive constant.

(C4) The number of interior knots J_n satisfies: $n^{1/(2p)} \ll J_n \ll n^{1/3}$.

(C5) The projection function $\Gamma(\boldsymbol{z})$ has the additive form $\Gamma(\boldsymbol{z}) = \Gamma_1(z_1) + \cdots + \Gamma_K(z_K)$, where $\Gamma_k \in \mathcal{H}$, $E[\Gamma_k(Z_k)] = 0$, and $E[\Gamma_k(Z_k)]^2 < \infty$, $k = 1, \ldots, K$.

A.2. Technical lemmas

According to the result of de Boor (2001, p.149), for any function $\eta \in \mathcal{H}$ and $n \geq 1$, there exists a function $\widetilde{\eta} \in \mathcal{S}_n$ such that $\left\|\widetilde{\eta} - \eta\right\|_\infty \leq Ch^p$. Recall that $\mathbf{B}(\boldsymbol{z}) = \{B_{j,k}(z_k), j = -\varrho + 1, \ldots, J_n, k = 1, \ldots, K\}^{\mathrm{T}}$. For g_0 satisfying (C1), we can find $\widetilde{\boldsymbol{\gamma}} = \{\widetilde{\gamma}_{j,k}, j = -\varrho + 1, \ldots, J_n, k = 1, \ldots, K\}^{\mathrm{T}}$ and an additive spline function $\widetilde{g} = \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{B}(\boldsymbol{z}) \in \mathcal{G}_n$, such that

$$\left\|\widetilde{g} - g_0\right\|_\infty = O\left(h^p\right). \tag{A.3}$$

In the following, let

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^n \left[Y_i - \left\{\widetilde{g}\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}\right\}\right]^2. \tag{A.4}$$

Write $m_{0i} \equiv m_0\left(\boldsymbol{T}_i\right) = g_0\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0$, and

$$\widetilde{m}_0\left(\boldsymbol{t}\right) = \widetilde{g}\left(\boldsymbol{z}\right) + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_0, \quad \widetilde{m}_{0i} \equiv \widetilde{m}_0\left(\boldsymbol{T}_i\right) = \widetilde{g}\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0. \tag{A.5}$$

Lemma A.1. *Under Conditions* (C1)–(C4), $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow N(\mathbf{0}, \mathbf{A}^{-1} \boldsymbol{\Sigma}_1 \mathbf{A}^{-1})$, *where* $\mathbf{A} = E\left(\boldsymbol{X}^{\otimes 2}\right)$ *and* $\boldsymbol{\Sigma}_1 = E\left(\varepsilon^2 \boldsymbol{X}^{\otimes 2}\right)$.

Proof. Let $\widehat{\boldsymbol{\delta}} = \sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$. According to (A.4), $\widehat{\boldsymbol{\delta}}$ minimizes

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

By expansion, one has

$$\widetilde{l}_n\left(\boldsymbol{\delta}\right) = -n^{-1/2} \sum_{i=1}^n \left(Y_i - \widetilde{m}_{0i}\right) \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{X}_i + 2^{-1} \boldsymbol{\delta}^{\mathrm{T}} \mathbf{A}_n \boldsymbol{\delta},$$

where $\mathbf{A}_n = (1/n)\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} = \mathbf{A} + o_P(1)$. Observe that

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^n (Y_i - \widetilde{m}_{0i})\boldsymbol{X}_i \\ &= n^{-1/2}\sum_{i=1}^n \left[Y_i - \{\widetilde{g}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\}\right]\boldsymbol{X}_i \\ &= n^{-1/2}\sum_{i=1}^n \left[Y_i - \{g_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\}\right]\boldsymbol{X}_i + n^{-1/2}\sum_{i=1}^n \{g_0(\boldsymbol{Z}_i) - \widetilde{g}(\boldsymbol{Z}_i)\}\boldsymbol{X}_i \\ &= n^{-1/2}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i + n^{-1/2}\sum_{i=1}^n \{g_0(\boldsymbol{Z}_i) - \widetilde{g}(\boldsymbol{Z}_i)\}\boldsymbol{X}_i. \end{aligned}$$

By (A.3) and Condition (C3), the absolute value of the second term on the right-hand side of the above equation is

$$|n^{-1/2}\sum_{i=1}^n \{g_0(\boldsymbol{Z}_i) - \widetilde{g}(\boldsymbol{Z}_i)\}\boldsymbol{X}_i| \le n^{-1/2}\sum_{i=1}^n |\boldsymbol{X}_i| \, \|\widetilde{g} - g_0\|_\infty = o_P(1).$$

Thus,

$$\widetilde{l}_n(\boldsymbol{\delta}) = -n^{-1/2}\sum_{i=1}^n \varepsilon_i \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{X}_i + 2^{-1}\boldsymbol{\delta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\delta} + o_P(1),$$

and the convexity lemma of Pollard (1991) implies that

$$\widehat{\boldsymbol{\delta}} = \mathbf{A}^{-1}n^{-1/2}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i + o_P(1).$$

It follows that $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \to N\left(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}_1\mathbf{A}^{-1}\right)$.

Let

$$\mathbf{V}_n = n^{-1}\sum_{i=1}^n \left\{\begin{matrix} \{\mathbf{B}(\boldsymbol{Z}_i)\}^{\otimes 2} & \mathbf{B}(\boldsymbol{Z}_i)\boldsymbol{X}_i^{\mathrm{T}} \\ \boldsymbol{X}_i\mathbf{B}^{\mathrm{T}}(\boldsymbol{Z}_i) & \boldsymbol{X}_i^{\otimes 2} \end{matrix}\right\}. \tag{A.6}$$

Lemma A.2. *Under* (C1)−(C4)*, there exists a positive constant* C *such that* $\left\|\mathbf{V}_n^{-1}\right\|_2 \le C$*, a.s..*

Proof. We first derive the lower and upper bound of the eigenvalues of $\mathbf{V}_n$. For any vectors $\boldsymbol{\omega}_1 = \{\omega_{j,k}, j = -\varrho + 1, \ldots, J_n, k = 1, \ldots, K\} \in R^{(J_n+\varrho)K}$ and $\boldsymbol{\omega}_2 \in R^d$, let $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^{\mathrm{T}}, \boldsymbol{\omega}_2^{\mathrm{T}})^{\mathrm{T}}$. Then one has

$$n\boldsymbol{\omega}^{\mathrm{T}}\mathbf{V}_n\boldsymbol{\omega} = \boldsymbol{\omega}_1^{\mathrm{T}}\sum_{i=1}^n \{\mathbf{B}(\boldsymbol{Z}_i)\}^{\otimes 2}\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2^{\mathrm{T}}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2}\boldsymbol{\omega}_2 + 2\boldsymbol{\omega}_1^{\mathrm{T}}\sum_{i=1}^n \mathbf{B}(\boldsymbol{Z}_i)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\omega}_2.$$

Lemma 1 of Stone (1985) provides a constant $c > 0$ such that

$$\left\| \sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 \geq c \sum_{k=1}^{K} \left\| \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 .$$

According to Theorem 5.4.2 of DeVore and Lorentz (1993), Condition (C2) and the definition of $B_{j,k}$ in (A.2), there exist constants $C'_k > c'_k > 0$ such that, for any $k = 1, \ldots, K$,

$$c'_k \sum_{j=-\varrho+1}^{J_n} \omega_{j,k}^2 \leq \left\| \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 \leq C'_k \sum_{j=-\varrho+1}^{J_n} \omega_{j,k}^2 .$$

Thus there exist constants $C_0 > c_0 > 0$ such that

$$c_0 \left\| \boldsymbol{\omega}_1 \right\|^2 \leq \left\| \sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 \leq C_0 \left\| \boldsymbol{\omega}_1 \right\|^2 .$$

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

$$A_n \equiv \sup_{g_1, g_2 \in \mathcal{G}_n} \left| \frac{\langle g_1, g_2 \rangle_n - \langle g_1, g_2 \rangle}{\|g_1\|_2 \|g_2\|_2} \right| = O\left\{ \left(\log \frac{n}{nh} \right)^{1/2} \right\}, \ a.s.. \tag{A.7}$$

It is clear to see that

$$\begin{aligned} (1 - A_n) \left\| \sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 &\leq \boldsymbol{\omega}_1^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \left\{ \mathbf{B}\left(\boldsymbol{Z}_i\right) \right\}^{\otimes 2} \boldsymbol{\omega}_1 \\ = \left\| \sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_{2,n}^2 &\leq (1 + A_n) \left\| \sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \omega_{j,k} B_{j,k} \right\|_2^2 . \end{aligned}$$

Therefore, $c \left\| \boldsymbol{\omega}_1 \right\|^2 \leq \boldsymbol{\omega}_1^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \left\{ \mathbf{B}\left(\boldsymbol{Z}_i\right) \right\}^{\otimes 2} \boldsymbol{\omega}_1 \leq C \left\| \boldsymbol{\omega}_1 \right\|^2$, a.s.. Next,

$$\begin{aligned} \boldsymbol{\omega}_2^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} \boldsymbol{\omega}_2 &= \boldsymbol{\omega}_2^{\mathrm{T}} E\left(\boldsymbol{X}^{\otimes 2}\right) \boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^{\mathrm{T}} \left[n^{-1} \textstyle\sum_{i=1}^{n} \left\{ \boldsymbol{X}_i^{\otimes 2} - E\left(\boldsymbol{X}^{\otimes 2}\right) \right\} \right] \boldsymbol{\omega}_2 \\ &= \boldsymbol{\omega}_2^{\mathrm{T}} E\left(\boldsymbol{X}^{\otimes 2}\right) \boldsymbol{\omega}_2 + \left\| \boldsymbol{\omega}_2 \right\|^2 o(1), \ a.s.. \end{aligned}$$

and, according to (C3), $c \left\| \boldsymbol{\omega}_2 \right\|^2 \leq \boldsymbol{\omega}_2^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} \boldsymbol{\omega}_2 \leq C \left\| \boldsymbol{\omega}_2 \right\|^2$, a.s.. Then $\left| \boldsymbol{\omega}_1^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \mathbf{B}\left(\boldsymbol{Z}_i\right) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\omega}_2 \right| = o(\left\| \boldsymbol{\omega}_1 \right\| \left\| \boldsymbol{\omega}_2 \right\|)$, a.s.. Thus

$$c \left\| \boldsymbol{\omega} \right\|^2 \leq \boldsymbol{\omega}^{\mathrm{T}} \mathbf{V}_n \boldsymbol{\omega} \leq C \left\| \boldsymbol{\omega} \right\|^2 , \ a.s.. \tag{A.8}$$

Let $\lambda_{\max}\left(\mathbf{V}_n\right)$ and $\lambda_{\min}\left(\mathbf{V}_n\right)$ be the maximum and minimum eigenvalues of $\mathbf{V}_n$. Algebra and (A.8) show that $\left\| \mathbf{V}_n \right\|_2 = \lambda_{\max}\left(\mathbf{V}_n\right) \leq C$ and $\left\| \mathbf{V}_n^{-1} \right\|_2 = \lambda_{\min}^{-1}\left(\mathbf{V}_n\right) \leq c^{-1}$, a.s..

In the following, take $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}$, $\widetilde{\boldsymbol{\theta}} = \begin{pmatrix} \widetilde{\boldsymbol{\gamma}} \\ \widetilde{\boldsymbol{\beta}} \end{pmatrix}$, $\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix}$, $\widehat{l}_n(\boldsymbol{\theta}) = \ell(\boldsymbol{\gamma}, \boldsymbol{\beta})$, and

$$\widetilde{m}_i \equiv \widetilde{m}(\boldsymbol{T}_i) = \widetilde{g}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{B}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}. \tag{A.9}$$

Lemma A.3. *Under* (C1)−(C4), $\left\| \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \right\| = O_P \left\{ J_n^{1/2} \left(h^p + n^{-1/2} \right) \right\}$.

Proof. Note that

$$\left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} - \left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} = \left. \frac{\partial^2 \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}} \left(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \right),$$

where $\overline{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$. So

$$\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} = - \left(\left. \frac{\partial^2 \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}} \right)^{-1} \left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}}.$$

Next write

$$\left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} = \left\{ \left(\frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} \right)^{\mathrm{T}}, \left(\frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} \right\}^{\mathrm{T}} \Bigg|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} = - \sum_{i=1}^{n} (Y_i - \widetilde{m}_i) \{ \mathbf{B}(\boldsymbol{Z}_i), \boldsymbol{X}_i \},$$

where

$$\begin{aligned} \left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} &= - \sum_{i=1}^{n} (Y_i - m_{0i}) \mathbf{B}(\boldsymbol{Z}_i) + \sum_{i=1}^{n} \{ \widetilde{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \} \mathbf{B}(\boldsymbol{Z}_i) \\ &\quad + \sum_{i=1}^{n} \boldsymbol{X}_i^{\mathrm{T}} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) \mathbf{B}(\boldsymbol{Z}_i), \\ \left. \frac{\partial \widehat{l}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}} &= - \sum_{i=1}^{n} (Y_i - m_{0i}) \boldsymbol{X}_i + \sum_{i=1}^{n} \{ \widetilde{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \} \boldsymbol{X}_i \\ &\quad + \sum_{i=1}^{n} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right)^{\mathrm{T}} \boldsymbol{X}_i^{\otimes 2}. \end{aligned}$$

Observing that

$$\left\| - \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{0i}) \mathbf{B}(\boldsymbol{Z}_i) \right\| = \left[\sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i B_{j,k}(Z_{ik}) \right\}^2 \right]^{1/2},$$

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

we have $\|-(1/n)\sum_{i=1}^{n}(Y_i - m_{0i})\mathbf{B}(\boldsymbol{Z}_i)\| = O_P\left\{(J_n/n)^{1/2}\right\}$. By (C4), (A.3) and Lemma 1,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}\mathbf{B}(\boldsymbol{Z}_i)\right\| = O_P\left(J_n^{1/2}h^p\right),$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathrm{T}}\boldsymbol{X}_i\mathbf{B}(\boldsymbol{Z}_i)\right\| = O_P\left\{\left(\frac{J_n}{n}\right)^{1/2}\right\}.$$

Therefore, $\left\|(1/n)\,(\partial\widehat{l}_n(\boldsymbol{\theta})/\partial\boldsymbol{\gamma})\Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}\right\| = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}$. Similarly, one has

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(m_{0i} - Y_i)\boldsymbol{X}_i\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i\right\| = O_P\left\{\left(\frac{J_n}{n}\right)^{1/2}\right\},$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}\boldsymbol{X}_i\right\| = O_P\left(J_n^{1/2}h^p\right),$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right\| = O_P\left\{\left(\frac{J_n}{n}\right)^{1/2}\right\}.$$

Thus $\left\|(1/n)\,(\widehat{l}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}\right\| = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}$. For the second order derivative, one has

$$\frac{1}{n}\left.\frac{\partial^2\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} = \frac{1}{n}\left.\begin{Bmatrix}\frac{\partial^2\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}} & \frac{\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^{\mathrm{T}}}\\ \frac{\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}} & \frac{\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\end{Bmatrix}\right|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} = \mathbf{V}_n.$$

According to Lemma A2, $\left\|\mathbf{V}_n^{-1}\right\|_2 = O_P(1)$. Thus

$$\left\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\right\| \leq \left\|\mathbf{V}_n^{-1}\right\|_2\left\|\frac{1}{n}\left.\frac{\partial\widehat{l}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}\right\| = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}.$$

Lemma A.4. *Under* (C1)−(C4), $\|\widehat{g} - g_0\|_2 = O_P\{(J_n/n)^{1/2}\}$, $\|\widehat{g} - g_0\|_n = O_P\{(J_n/n)^{1/2}\}$, $\|\widehat{g}_k - g_{0k}\|_{2k} = O_P\{(J_n/n)^{1/2}\}$ *and* $\|\widehat{g}_k - g_{0k}\|_{nk} = O_P\{(J_n/n)^{1/2}\}$, *for* $k = 1, \ldots, K$.

Proof. According to Lemmas A2 and A3, $\|\widehat{g} - \widetilde{g}\|_2^2$ is equal to $\int_{[0,1]^K}(\widehat{g} - \widetilde{g})^2(\boldsymbol{z})f(\boldsymbol{z})d\boldsymbol{z} = (\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}})^{\mathrm{T}}\left(\langle B_{j,k}, B_{j',k'}\rangle\right)_{\substack{-\varrho\leq j,j'\leq J_n,\\ 1\leq k,k'\leq K}}(\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}) \leq C\,\|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_2^2$, thus $\|\widehat{g} - \widetilde{g}\|_2 = O_P\{J_n^{1/2}(h^p + n^{-1/2})\}$ and

$$\begin{aligned}\|\widehat{g} - g_0\|_2 \leq \|\widehat{g} - \widetilde{g}\|_2 + \|\widetilde{g} - g_0\|_2 &= O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\} + O_P(h^p)\\ &= O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}.\end{aligned}$$

By Lemma 1 of Stone (1985), $\|\widehat{g}_k - g_{0k}\|_{2k} = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}$, $1 \leq k \leq K$. Equation (A.7) then implies that $\|\widehat{g} - \widetilde{g}\|_n = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}$. Then

$$\begin{aligned}\|\widehat{g} - g_0\|_n \leq \|\widehat{g} - \widetilde{g}\|_n + \|\widetilde{g} - g_0\|_n &= O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\} + O_P\left(h^p\right) \\ &= O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}.\end{aligned}$$

Similar to (A.7),

$$\sup_{g \in \mathcal{S}_n}\left|\frac{\|g\|_{nk}}{\|g\|_{2k}} - 1\right| = O_P\left\{\left(\frac{\log n}{nh}\right)^{1/2}\right\}, \quad k = 1, \ldots, K.$$

Thus $\|\widehat{g}_k - g_{0k}\|_{nk} = O_P\left\{J_n^{1/2}\left(h^p + n^{-1/2}\right)\right\}$, for any $k = 1, \ldots, K$. The desired result follows by (C4).

Lemma A.5. *Under* (C1)−(C4)*,*

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i \Gamma\left(\boldsymbol{Z}_i\right)^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = o_P\left(n^{-1/2}\right), \tag{A.10}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{g}\left(\boldsymbol{Z}_i\right) - g_0\left(\boldsymbol{Z}_i\right)\right\}\widetilde{\boldsymbol{X}}_i = o_P\left(n^{-1/2}\right). \tag{A.11}$$

Proof. We first show (A.11). Let $s\left(\boldsymbol{z}, g\right) = g(\boldsymbol{z})\widetilde{\boldsymbol{x}}$. Note that

$$E\left\{s\left(\boldsymbol{Z}, \widehat{g}\right) - s\left(\boldsymbol{Z}, g_0\right)\right\}^2 = E\left\{\left(\widehat{g} - g_0\right)\left(\boldsymbol{Z}_i\right)\widetilde{\boldsymbol{X}}_i\right\}^2 \leq O\left(\|\widehat{g} - g_0\|_2^2\right).$$

By Lemma A.2 of Huang (1999), the logarithm of the ε-bracketing number of the class of functions $\mathcal{A}_1(\delta) = \left\{s\left(\cdot, \widehat{g}\right) - s\left(\cdot, g_0\right) : g \in \mathcal{G}_n, \|g - g_0\|_2 \leq \delta\right\}$ is $c\left\{\left(J_n + \varrho\right)\log\left(\delta/\varepsilon\right) + \log\left(\delta^{-1}\right)\right\}$, so the corresponding entropy integral $J_{[]}(\delta, \mathcal{A}_1(\delta), \|\cdot\|_2) \leq c\delta\{(J_n + \varrho)^{1/2} + \log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of Stone (1986) and Lemma A4, $\|\widehat{g} - g_0\|_\infty \leq cJ_n^{1/2}\|\widehat{g} - g_0\|_2 = O_P\left(n^{-1/2}J_n\right)$. Lemma 3.4.2 of van der Vaar and Wellner (1996) implies that, for $r_n = (n/J_n)^{1/2}$,

$$\begin{aligned}&E\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{g}\left(\boldsymbol{Z}_i\right) - g_0\left(\boldsymbol{Z}_i\right)\right\}\widetilde{\boldsymbol{X}}_i - E\left[\left\{\widehat{g}\left(\boldsymbol{Z}\right) - g_0\left(\boldsymbol{Z}\right)\right\}\widetilde{\boldsymbol{X}}\right]\right| \\ &\leq n^{-1/2}Cr_n^{-1}\left\{\left(J_n + \varrho\right)^{1/2} + \log^{1/2}\left(r_n\right)\right\} \\ &\quad \times \left[1 + \frac{cr_n^{-1}\left\{\left(J_n + \varrho\right)^{1/2} + \log^{1/2}\left(r_n\right)\right\}}{r_n^{-2}\sqrt{n}}C_0\right] \\ &\leq O(1)n^{-1/2}r_n^{-1}\left\{\left(J_n + \varrho\right)^{1/2} + \log^{1/2}\left(r_n\right)\right\}.\end{aligned}$$

By Condition (C4), $O\left(n^{-1/2}J_n\right) = o\left(1\right)$. Thus, one has

$$E\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{g}\left(\boldsymbol{Z}_i\right) - g_0\left(\boldsymbol{Z}_i\right)\right\}\widetilde{\boldsymbol{X}}_i - E\left[\left\{\widehat{g}\left(\boldsymbol{Z}\right) - g_0\left(\boldsymbol{Z}\right)\right\}\widetilde{\boldsymbol{X}}\right]\right| = o\left(n^{-1/2}\right).$$

By the definition of $\widetilde{\boldsymbol{X}}$, for any measurable function ϕ, $E\left\{\phi\left(\boldsymbol{Z}\right)\widetilde{\boldsymbol{X}}\right\} = 0$. Hence (A.11) holds. Similarly, (A.10) follows from Lemma 3.4.2 of van der Vaart and Wellner (1996) and Lemma A3.

Lemma A.6. *Under the conditions of Theorem 2, with probability tending to* 1*, for any given* $\boldsymbol{\beta}_1$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ *and any constant* C*,*

$$\mathcal{L}_P\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}, \boldsymbol{\gamma}\right\} = \min_{\|\boldsymbol{\beta}_2\|\le Cn^{-1/2}} \mathcal{L}_P\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}, \boldsymbol{\gamma}\right\}.$$

Proof. To prove that the minimum is attained at $\boldsymbol{\beta}_2 = 0$, it suffices to show that with probability tending to 1, as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}$, $\partial\mathcal{L}_P(\boldsymbol{\beta})/\partial\beta_j$ and β_j have the same signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$, $j = s+1, \cdots, d$. It follows by arguments similar to those in the proof of Theorem 1 that

$$\ell_j'(\boldsymbol{\beta}) \equiv \frac{\partial\ell(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})}{\partial\beta_j} = n\left\{\frac{1}{n}\sum_{i=1}^{n}\Omega_j(Y_i, \boldsymbol{T}_i) - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}R_j + o_P(n^{-1/2})\right\},$$

where $\Omega_j(Y_i, \boldsymbol{T}_i)$ is the jth element of $-\varepsilon_i\widetilde{\boldsymbol{X}}_i$ and R_j is the jth column of $E(\widetilde{\boldsymbol{X}}^{\otimes 2})$. Note that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$ by assumption. Thus, $n^{-1}\ell_j'(\boldsymbol{\beta})$ is of the order $O_P(n^{-1/2})$. Therefore, for any zero β_j and $j = s+1, \cdots, d$,

$$\begin{aligned}\frac{\partial\mathcal{L}_P(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\beta_j} &= \ell_j'(\boldsymbol{\beta}) + np'_{\lambda_{jn}}\left(|\beta_j|\right)\mathrm{sgn}(\beta_j)\\ &= n\lambda_{jn}\{\lambda_{jn}^{-1}p'_{\lambda_{jn}}\left(|\beta_j|\right)\mathrm{sgn}(\beta_j) + O_P(\frac{1}{\sqrt{n}\lambda_n})\}.\end{aligned}$$

Because $\liminf_{n\to\infty}\liminf_{\beta_j\to 0^+}\lambda_{jn}^{-1}p'_{\lambda_{jn}}(|\beta_j|) > 0$ and $\sqrt{n}\lambda_{jn} \to \infty$, the sign of the derivative is completely determined by that of β_j. Thus the desired result is obtained.

A.3. Proof of Theorem 1

According to (C5), the projection function $\Gamma\left(\boldsymbol{z}\right) = \Gamma_1(z_1) + \cdots + \Gamma_K(z_K)$, where the theoretically centered function $\Gamma_k \in \mathcal{H}$. By the result of de Boor (2001, p.149), there exists an empirically centered function $\widetilde{\Gamma}_k \in \mathcal{S}_n$, such that

$\left\|\widetilde{\Gamma}_k - \Gamma_k\right\|_\infty = O_P(h^p)$, $k = 1, \ldots, K$. If $\widetilde{\Gamma}(\boldsymbol{z}) = \widetilde{\Gamma}_1(z_1) + \cdots + \widetilde{\Gamma}_K(z_K)$, $\widetilde{\Gamma} \in \mathcal{G}_n$. Define a class of functions

$$\mathcal{M}_n = \left\{m(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{z}) + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} : g \in \mathcal{G}_n\right\}. \tag{A.12}$$

For any $\boldsymbol{v} \in R^d$, let $\widehat{m}(\boldsymbol{x}, \boldsymbol{z}) = \widehat{g}(\boldsymbol{z}) + \boldsymbol{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ and $\widehat{m}_{\boldsymbol{v}} = \widehat{m}(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{v}^{\mathrm{T}}\{\boldsymbol{x} - \widetilde{\Gamma}(\boldsymbol{z})\}$. Then $\widehat{m}_{\boldsymbol{v}} = \{\widehat{g}(\boldsymbol{z}) - \boldsymbol{v}^{\mathrm{T}}\widetilde{\Gamma}(\boldsymbol{z})\} + (\widehat{\boldsymbol{\beta}} + \boldsymbol{v})^{\mathrm{T}}\boldsymbol{x} \in \mathcal{M}_n$. Note that $\widehat{m}_{\boldsymbol{v}}$ minimizes the function $\widehat{l}_n(m) = 1/2\sum_{i=1}^n \{Y_i - m(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}^2$ for all $m \in \mathcal{M}_n$ when $\boldsymbol{v} = \mathbf{0}$, thus $\frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_n(\widehat{m}_{\boldsymbol{v}})\Big|_{\boldsymbol{v}=\mathbf{0}} = \mathbf{0}$. Write

$$\widehat{m}_i \equiv \widehat{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i) = \widehat{g}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}, \tag{A.13}$$

and $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \Gamma(\boldsymbol{Z}_i)$. Then

$$\begin{aligned}\mathbf{0} \equiv \frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_n(\widehat{m}_{\boldsymbol{v}})\Big|_{\boldsymbol{v}=\mathbf{0}} &= -\sum_{i=1}^n (Y_i - \widehat{m}_i)\left\{\boldsymbol{X}_i - \widetilde{\Gamma}(\boldsymbol{Z}_i)\right\} \\ &= -\sum_{i=1}^n (Y_i - \widehat{m}_i)\widetilde{\boldsymbol{X}}_i + O_P(h^p) \\ &= -\sum_{i=1}^n \varepsilon_i\widetilde{\boldsymbol{X}}_i + \sum_{i=1}^n (\widehat{m}_i - m_{0i})\widetilde{\boldsymbol{X}}_i + O_P(h^p). \end{aligned} \tag{A.14}$$

Note that

$$\widehat{m}(\boldsymbol{x}, \boldsymbol{z}) - m_0(\boldsymbol{x}, \boldsymbol{z}) = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathrm{T}}\widetilde{\boldsymbol{x}} + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathrm{T}}\Gamma(\boldsymbol{z}) + \widehat{g}(\boldsymbol{z}) - g_0(\boldsymbol{z}).$$

We can rewrite the second term $\sum_{i=1}^n (\widehat{m}_i - m_{0i})\widetilde{\boldsymbol{X}}_i$ in (A.14) as

$$\left(\sum_{i=1}^n \widetilde{\boldsymbol{X}}_i^{\otimes 2}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + \left\{\sum_{i=1}^n \widetilde{\boldsymbol{X}}_i\Gamma(\boldsymbol{Z}_i)^{\mathrm{T}}\right\}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + \sum_{i=1}^n \{\widehat{g}(\mathbf{Z}_i) - g_0(\boldsymbol{Z}_i)\}\widetilde{\boldsymbol{X}}_i.$$

By Lemma A5, one has

$$\frac{1}{n}\sum_{i=1}^n (\widehat{m}_i - m_{0i})\widetilde{\boldsymbol{X}}_i = \left\{E(\widetilde{\boldsymbol{X}}^{\otimes 2}) + o_P(1)\right\}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + o_P\left(n^{-1/2}\right). \tag{A.15}$$

Combining (A.14), (A.15) and Condition (C4), one has

$$\mathbf{0} = -\frac{1}{n}\sum_{i=1}^n \varepsilon_i\widetilde{\boldsymbol{X}}_i + \left\{E(\widetilde{\boldsymbol{X}}^{\otimes 2}) + o_P(1)\right\}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + o_P\left(n^{-1/2}\right).$$

Thus the desired distribution of $\widehat{\boldsymbol{\beta}}$ follows.

A.4. Proof of Theorem 2

Let $\tau_n = n^{-1/2} + a_n$. It suffices to show that for any given $\zeta > 0$, there exists a large constant C such that

$$P\Big\{\sup_{\|\boldsymbol{v}\|=C}\mathcal{L}_P(\boldsymbol{\beta}_0+\tau_n\boldsymbol{v},\boldsymbol{\gamma})<\mathcal{L}_P(\boldsymbol{\beta}_0,\boldsymbol{\gamma})\Big\}\geq 1-\zeta. \tag{A.16}$$

Let

$$D_{n,1}=\frac{1}{2}\sum_{i=1}^{n}\left[\{Y_i-(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{Z}_i)+\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0+\tau_n\boldsymbol{v}))\}^2-\{Y_i-(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{Z}_i)+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2\right]$$

and $D_{n,2} = n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0}+\tau_n v_j|)-p_{\lambda_n}(|\beta_{j0}|)\}$, where s is the number of components of $\boldsymbol{\beta}_{10}$. Note that $p_{\lambda_n}(0)=0$ and $p_{\lambda_n}(|\beta|)\geq 0$ for all β. Thus, $\mathcal{L}_P(\boldsymbol{\beta}_0+\tau_n\boldsymbol{v},\boldsymbol{\gamma})-\mathcal{L}_P(\boldsymbol{\beta}_0,\boldsymbol{\gamma})\geq D_{n,1}+D_{n,2}$. Let $\widehat{m}_{0i}=\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{Z}_i)+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$. For $D_{n,1}$, note that $D_{n,1}=(1/2)\sum_{i=1}^{n}\Big[\{Y_i-(\widehat{m}_{0i}+\tau_n\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i)\}^2-(Y_i-\widehat{m}_{0i})^2\Big]$. Mimicking the proof of Theorem 1 indicates that

$$D_{n,1}=-\tau_n\boldsymbol{v}^{\mathrm{T}}\sum_{i=1}^{n}\varepsilon_i\widetilde{\boldsymbol{X}}_i+\frac{1}{2}\tau_n^2\boldsymbol{v}^{\mathrm{T}}\mathbf{D}\boldsymbol{v}+o_P(1), \tag{A.17}$$

where the orders of the first term and the second term are $O_P(n^{1/2}\tau_n)$ and $O_P(n\tau_n^2)$, respectively. For $D_{n,2}$, by a Taylor expansion and the Cauchy-Schwartz inequality, $n^{-1}D_{n,2}$ is bounded by $\sqrt{s}\tau_n a_n\|\boldsymbol{v}\|+\tau_n^2 w_n\|\boldsymbol{v}\|^2=C\tau_n^2(\sqrt{s}+w_n C)$. As $w_n\to 0$, both the first and second terms on the right-hand side of (A.17) dominate $D_{n,2}$, by making C sufficiently large. Hence (A.16) holds for a sufficiently large C.

We now prove part (II). From Lemma A6, it follows that $\widehat{\boldsymbol{\beta}}_2=0$. Let $\widehat{\boldsymbol{\beta}}_1^*=\sqrt{n}(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})$, $\widehat{m}_{i1}=\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{Z}_i)+\boldsymbol{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{10}$, and $m_{i1}=g_0^{\mathrm{T}}(\boldsymbol{Z}_i)+\boldsymbol{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{10}$. Then, $\widehat{\boldsymbol{\beta}}_1^*$ minimizes

$$\frac{1}{2}\sum_{i=1}^{n}\left[\left\{Y_i-(\widehat{m}_{i1}+n^{-1/2}\boldsymbol{X}_{i1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1^*)\right\}^2-(Y_i-\widehat{m}_{i1})^2\right]+n\sum_{j=1}^{s}p_{\lambda_{jn}}(|\beta_j|). \tag{A.18}$$

Let $\ell_{n1}(\boldsymbol{\beta}_1^*)$ be the first term in (A.18). Then

$$\ell_{n1}(\boldsymbol{\beta}_1^*)=-n^{-1/2}\sum_{i=1}^{n}(Y_i-\widehat{m}_{i1})\boldsymbol{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1^*+2^{-1}(\boldsymbol{\beta}_1^*)^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i1}^{\otimes 2}\right\}\boldsymbol{\beta}_1^*. \tag{A.19}$$

Using the arguments similar to the proofs for (A.15) and (A.17) yields

$$\ell_{n1}(\boldsymbol{\beta}_1^*)=-n^{-1/2}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_1^*(Y_i-m_{i1})\widetilde{\boldsymbol{X}}_{i1}+\frac{1}{2}\widehat{\boldsymbol{\beta}}_1^{*\mathrm{T}}E(\widetilde{\boldsymbol{X}}_1^{\otimes 2})\widehat{\boldsymbol{\beta}}_1^*+o_P(1).$$

Using the Convexity Lemma (Pollard (1991)) and combining (A.18), one has

$$\left(E(\widetilde{\boldsymbol{X}}_1^{\otimes 2}) + \boldsymbol{\Sigma}_\lambda\right)\widehat{\boldsymbol{\beta}}_1^* + n^{1/2}\kappa_n = n^{-1/2}\sum_{i=1}^{n}(Y_i - m_{i1})\,\widetilde{\boldsymbol{X}}_{i1} + o_P(1).$$

Hence the asymptotic normality is derived.

References

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika* **60**, 255-265.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350-2383.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag, New York.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation: Polynomials and Splines Approximation.* Springer-Verlag, Berlin.

Fairfield, K. M. and Fletcher, R. H. (2002). Vitamins for chronic disease prevention in adults. *J. Amer. Med. Assoc.* **287**, 3116-3226.

Fan, J. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. Z. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Faure, H., Preziosi, P., Roussel, A.-M., Bertrais, S., Galan, P., Hercberg, S. and Favie, A. (2006). Factors influencing blood concentration of retinol, α-tocopherol, vitamin C, and β-carotene in the French participants of the SU.VI.MAX trial. *European Journal of Clinical Nutrition* **60**, 706-717.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27**, 1536-1563.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31**, 1600-1635.

Li, Q. (2000). Efficient estimation of additive partially linear models. *Int. Econometric Rev.* **41**, 1073-1092.

Li, R. Z. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

Liang, H. and Li, R. Z. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104**, 234-248.

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

Lin, X. H. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96**, 1045-1056.

Ni, H., Zhang, H. H. and Zhang, D. (2009). Automatic model selection for partially linear models. *J. Multivariate Anal.* **100**, 2100-2111.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J. and Greenberg, E. R. (1989). Determinants of plasma levels of beta-carotene and retinol. *Am. J. Epidemiology* **130**, 511-521.

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25**, 186-211.

Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Statist.* **8**, 715-732.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**, 186-199.

Ravikumar, P., Liu, H., Lafferty, H. and Wasserman, L. (2008). Spam: Sparse additive models. *Advances in Neural Information Processing System* **20**, 1202-1208.

Ravikumar, P., Lafferty, H., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 1009-1030.

Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression.* Cambridge University Press, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590-606.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267-288.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* **35**, 2474-2503.

Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

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

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, U.S.A.

E-mail: xliu@bst.rochester.edu

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: lilywang@uga.edu

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, U.S.A.

E-mail: hliang@bst.rochester.edu

(Received May 2009; accepted January 2010)