Econometric Theory, **29**, 2013, 1136–1161.
doi:10.1017/S0266466613000029

GENERALIZED ADDITIVE PARTIAL LINEAR MODELS WITH HIGH-DIMENSIONAL COVARIATES

HENG LIAN
Nanyang Technological University

HUA LIANG
University of Rochester

This paper studies generalized additive partial linear models with high-dimensional covariates. We are interested in which components (including parametric and nonparametric components) are nonzero. The additive nonparametric functions are approximated by polynomial splines. We propose a doubly penalized procedure to obtain an initial estimate and then use the adaptive least absolute shrinkage and selection operator to identify nonzero components and to obtain the final selection and estimation results. We establish selection and estimation consistency of the estimator in addition to asymptotic normality for the estimator of the parametric components by employing a penalized quasi-likelihood. Thus our estimator is shown to have an asymptotic oracle property. Monte Carlo simulations show that the proposed procedure works well with moderate sample sizes.

1. INTRODUCTION

Generalized additive partially linear models (GAPLM) (Hiirdle et al., 2004) are perhaps the most realistic tool to analyze the relationship between a discrete response variable and covariates, for which some covariates are conjectured to contribute through a sum of smooth unknown functions of predictor variables instead of a linear combination. For example, in studying the impact of various possible determinants on the intention of East Germans to migrate to West Germany 1 year after the German unification (Härdle, Mammen, and Müller, 1998; Müller and Rönz, 2000), some covariates may affect the response variable linearly, for instance, discrete covariates, whereas other covariates will affect it nonlinearly. This generalization allows richer and more flexible model structures than generalized linear models and generalized additive models. This is appealing for both model interpretation and prediction. GAPLM also cover commonly used semiparametric

We sincerely thank Professor Oliver Linton and two anonymous reviewers for their insightful and constructive comments, which have improved the manuscript significantly. Lian's research was supported by Singapore MOE Tier 1 RG 62/11. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444 and by Award 11228103 from the National Natural Science Foundation of China. Address correspondence to Hua Liang, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA; e-mail: Hua˙Liang@URMC.Rochester.edu.

models. For example, when there is only one nonparametric term, they reduce to generalized partially linear models (Burda, Härdle, Müller, and Werwatz, 1998; Horowitz, 1998; Severini and Staniswalis, 1994; Wu and Xiao, 2002). Furthermore, if the link function is the identity, they become well studied partially linear models (PLM), for which there is much work in the literature (Chen, 1988; Engle, Granger, Rice, and Weiss, 1986; Härdle, Liang, and Gao, 2000; Heckman, 1986; Juhl and Xiao, 2005; Li and Stengos, 1996; Li and Wooldridge, 2002; Robinson, 1988; Speckman, 1988; Su and Jin, 2010). Yatchew (2003) is an excellent account of semiparametric regression with an emphasis on empirical applications. When the link function in GAPLM is the identity, we have partially additive models (Li, 2000; Liu, Wang, and Liang, 2011).

For the application of the GAPLM, a practical issue is to decide which covariates go into the nonlinear part and which ones go into the linear part. One strategy that has been used is simply to put the discrete covariates into the linear part and the continuous ones into the nonlinear part. Such a strategy has been used in the literature, for example, Du, Ma, and Liang (2010) and Kneib, Konrath, and Fahrmeir (2011). We note that misspecification of the semiparametric structure might have adverse effects on estimation and variable selection. Recently Zhang, Cheng, and Liu (2011) have carried out an interesting study on automatic discovery of partially linear structure in semiparametric models. It remains to be seen whether this work can be extended to our high-dimensional case using quasi-likelihood.

In the early literature, kernel-based backfitting and local scoring procedures have been proposed by Buja, Hastie, and Tibshirani (1989) to iteratively estimate the linear coefficients and nonparametric components by solving a large system of equations (Yu, Park, and Mammen, 2008). Linton and Nielsen (1995) applied the marginal integration approach (Linton and Härdle, 1996). This strategy may suffer from the "curse of dimensionality" (Hiirdle et al., 2004). As it is well known that the kernel-based backfitting and marginal integration approaches are computationally expensive, Marx and Eilers (1998), Ruppert, Wand, and Carroll (2003), and Wood (2004) suggested penalized regression splines, but no theoretical justifications are available for these procedures.

To overcome these limitations, Wang, Liu, Liang, and Carroll (2011) proposed estimating the nonparametric components by polynomial splines. Polynomial splines estimation is a popular approach for nonparametric regression in the literature, including Andrews (1991), Andrews and Whang (1990), Chen (2007), Donald and Newey (1994), Huang (1998), Newey (1997), Stone (1986, 1994), and Xue and Yang (2006), many of which considered general series expansion where spline is a special case. That is, the unknown functions are approximated via polynomial splines characterized by a linear combination of spline basis. The gain of the proposed procedure in terms of computational reduction is remarkable in contrast to its counterparts: local scoring backfitting and marginal integration. We aim to develop a more genuine method that is able to select important parametric and nonparametric components numerically stably, even when the numbers of the nonparametric and parametric components diverge. To the best of our knowledge, there has been no attempt to carry out estimation for GAPLM with diverging numbers of covariates, let alone variable selection in such a setting.

Li and Liang (2008) adopted the Fan and Li (2001) variable selection procedures for parametric models via nonconcave penalized quasi-likelihood, but their models do not cover the GAPLM and they considered only the fixed-dimensional case. In this paper, we adopt the adaptive least absolute shrinkage and selection operator (LASSO) proposed by Zou (2006), which remedies the drawback of the LASSO (Tibshirani, 1996), lack of oracle property, with appropriately selecting the weights. The strategy of Zou (2006) is not applicable here because we consider semiparametric models with diverging numbers of covariates. Consequently the full model may even be unidentifiable. As an alternative to obtain weights for the adaptive LASSO, we propose a doubly penalized initial LASSO estimator. The resulting adaptive LASSO has the oracle property as long as the initial estimator achieves certain convergence rates, which can be guaranteed under the mild assumptions. See the details in Section 2.2. This alternative strategy was also advocated by Huang, Horowitz, and Wei (2010) for removing the requirement of $\sqrt{n}$-consistent initial estimator when they studied the adaptive LASSO for variable selection in high-dimensional linear and nonparametric additive models.

The rest of the paper is organized as follows. In Section 2, we introduce the GAPLM model, propose polynomial spline estimators via a quasi-likelihood approach and variable selection for the parametric and nonparametric components, and study the asymptotic properties of the proposed procedure. Simulation studies are presented in Section 3. Section 4 concludes. The proofs of the main results are presented in the Appendix.

2. ESTIMATION METHODS AND ASYMPTOTICS

Let Y be the response and its conditional expectation given the covariates $W = \left(W_1, \ldots, W_{p_1}\right)^{\mathrm{T}}$ and $X = \left(X_1, \ldots, X_{p_2}\right)^{\mathrm{T}}$ is defined via a known link function g by an additive linear function

$$\mu = E(Y|W,X) = g^{-1}\left\{\sum_{j=1}^{p_1} \alpha_j(W_j) + \sum_{j=1}^{p_2} X_j\beta_j\right\}, \tag{1}$$

where $\alpha(\cdot) = \left(\alpha_1(\cdot), \ldots, \alpha_{p_1}(\cdot)\right)^{\mathrm{T}}$ and $\beta = \left(\beta_1, \ldots, \beta_{p_2}\right)^{\mathrm{T}}$ are the nonparametric components and the coefficients of the parametric components, respectively. The variance is assumed to be a function of the mean given by

$$\mathrm{Var}(Y|W,X) = \sigma^2 V(\mu), \tag{2}$$

where V is a known function. It is not necessary for the value of σ to be known, but it plays no role in the specification of the quasi-likelihood later, and thus we assume $\sigma = 1$ without loss of generality. The quasi-likelihood approach includes Poisson regression and logistic regression as special cases, in addition to the simpler least squares regression, and the cases where exact distributional information is not available. Thus our proposed methods can be used for data with either continuous or discrete responses. For identifiability of the models, we assume that $E\alpha_j\left(W_j\right) = 0$. Furthermore, without loss of generality, we assume the distribution of $W_j, 1 \le j \le p_1$ is supported on $[0, 1]$.

The data we observe for the ith subject or unit are $\left(W_i, X_i, Y_i\right), i = 1, \ldots, n$, with $W_i = \left(W_{i1}, \ldots, W_{ip_1}\right)^{\mathrm{T}}$ and $X_i = (X_{i1}, \ldots, X_{ip_2})^{\mathrm{T}}$, assumed to be independent and identically distributed (i.i.d.) and satisfy (1) and (2) with true parameters denoted by $\alpha_0(\cdot) = \left(\alpha_{01}(\cdot), \ldots, \alpha_{0p_1}(\cdot)\right)^{\mathrm{T}}$ and $\beta_0 = \left(\beta_{01}, \ldots, \beta_{0p_2}\right)^{\mathrm{T}}$, respectively. Both p_1 and p_2 may diverge from, and possibly be larger than, the sample size. However, to make efficient estimation possible, we are only interested in a sparse true model where only s_1 nonparametric components and s_2 coefficients are nonzero. Theoretically allowed divergence rates for s_1, s_2, p_1, and p_2 will be discussed later. Here we only emphasize that we will assume s_1 is bounded for technical simplicity. In practice this assumption also seems reasonable when most of the significant predictors have linear effects. Without loss of generality, it is assumed that these nonzero components are the first ones in α_0 and β_0, respectively.

We use polynomial splines to approximate the nonparametric components. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order u is a function whose restriction to each subinterval is a polynomial of degree $u - 1$ and globally $u - 2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + u$. Because of the centering constraint $E\alpha_j\left(W_j\right) = 0$, we instead focus on the subspace of spline functions $S_j^0 := \left\{s : s(x) = \sum_{k=1}^{\tilde{K}} a_{jk} B_k(x), \sum_{i=1}^{n} s\left(W_{ij}\right) = 0\right\}$ with normalized basis $\left\{B_{jk}(x) = \sqrt{K}\left(B_k(x) - \sum_{i=1}^{n} B_k\left(W_{ij}\right)/n\right), k = 1, \ldots, K = \tilde{K} - 1\right\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). Using spline expansions, we can approximate the nonparametric components by $\alpha_j(x) \approx \sum_k a_{jk} B_{jk}(x)$, $1 \le j \le p_1$. Finally, we denote by D_j the $K \times K$ matrix with entries given by $\int B_{jk}(x) B_{jk'}(x)\, dx$, $1 \le k, k' \le K$ so that the L_2 norm of α_j is approximated by $\|a_j\|_{D_j} = \sqrt{a_j^{\mathrm{T}} D_j a_j}$ with $a_j = \left(a_{j1}, \ldots, a_{jK}\right)^{\mathrm{T}}$.

2.1. Doubly Penalized Adaptive Group LASSO Estimator

Our estimation procedure is based on the penalized quasi-likelihood. The (negative) quasi-likelihood function is defined by $Q(\mu, y) = \int_\mu^y (y - t)/V(t)\, dt$, and the negative quasi-likelihood of the observed n i.i.d. data is

$$\sum_{i=1}^{n} Q\left(g^{-1}\left(\sum_{j=1}^{p_1}\alpha_j\left(W_{ij}\right)+\sum_{j=1}^{p_2}X_{ij}\beta_j\right),Y_i\right)\approx\sum_i Q\left(g^{-1}\left(Z_i^{\mathrm{T}}a+X_i^{\mathrm{T}}\beta\right),Y_i\right),$$

using the notation $Z_i = \left(B_{11}\left(W_{i1}\right),\ldots,B_{1K}\left(W_{i1}\right),\ldots,B_{p_1K}\left(W_{ip_1}\right)\right)^{\mathrm{T}}$ and $a=\left(a_1^{\mathrm{T}},\ldots,a_{p_1}^{\mathrm{T}}\right)^{\mathrm{T}}=\left(a_{11},\ldots,a_{p_1K}\right)^{\mathrm{T}}$.

In this paper, we will use the adaptive LASSO penalty (Zou, 2006) for automatic variable selection and estimation. Other penalties such as the smoothly clipped absolute deviation penalty (Fan and Li, 2001; Zou and Li, 2008) could also be used. In particular, we propose the following estimation procedure based on penalized negative quasi-likelihood:

$$\begin{aligned}\left(\hat{a},\hat{\beta}\right)=\arg\min_{a,\beta}&\sum_{i=1}^{n}Q\left(g^{-1}\left(Z_i^{\mathrm{T}}a+X_i^{\mathrm{T}}\beta\right),Y_i\right)\\&+n\lambda_1\sum_{j=1}^{p_1}\omega_{1j}\|a_j\|_{D_j}+n\lambda_2\sum_{j=1}^{p_2}\omega_{2j}|\beta_j|,\end{aligned}\tag{3}$$

where λ_1 and λ_2 are the regularization parameters controlling the shrinkage of the nonparametric and parametric components, respectively, and $\omega_1 = \left(\omega_{11},\ldots,\omega_{1p_1}\right)^{\mathrm{T}}$, $\omega_2=\left(\omega_{21},\ldots,\omega_{2p_2}\right)^{\mathrm{T}}$ are two vectors of weights. These weights should be appropriately chosen (typically based on an initial consistent estimator) for the preceding estimator to enjoy desirable asymptotic properties. In Section 2.1 these weights are assumed to be given, and we defer the discussion on their choices to Section 2.2. This penalty combination enables us to simultaneously select significant parametric and nonparametric components, and also to estimate these parameters and nonparametric functions. The first penalty is the same as groupwise selection (Xue, 2009; Yuan and Lin, 2006) in spirit, because each additive component is treated as a whole group. If we are only interested in selecting X-variables, then we set $\lambda_1\equiv 0$, whereas if we are only interested in selecting W-variables, then we set $\lambda_2\equiv 0$.

We first introduce some notation. Define $q_l(x,y)=\left(\partial^l/\partial x^l\right)Q\left(g^{-1}(x),y\right)$, $l=1,2$. We have $q_1(x,y)=-(y-g^{-1}(x))\rho_1(x)$ and $q_2(x,y)=\rho_2(x)-\left(y-g^{-1}(x)\right)\rho_1'(x)$, with $\rho_l(x)=\left[\left(dg^{-1}(x)\right)/dx\right]^l/V(g^{-1}(x))$. Denote the true nonparametric components by $\alpha_0=(\alpha_{01},\ldots,\alpha_{0p_1})^{\mathrm{T}}$ and the true linear parameter by $\beta_0=(\beta_{01},\ldots,\beta_{0p_2})^{\mathrm{T}}$. Let $m_0(W,X)=\sum_j\alpha_{0j}\left(W_j\right)+\sum_j X_j\beta_{0j}$. A vector with superscript (1) generally denotes its subvector associated with nonzero components. For example, $X_i^{(1)}=\left(X_{i1},\ldots,X_{is_2}\right)^{\mathrm{T}}$, $Z_i^{(1)}=\left(B_{11}\left(W_{i1}\right),\ldots,B_{s_1K}\left(W_{is_1}\right)\right)^{\mathrm{T}}$.

Define

$$\Gamma(w)=\left(\Gamma_{(1)}(w),\ldots,\Gamma_{(s_2)}(w)\right)^{\mathrm{T}}=\frac{E\left\{X^{(1)}\rho_2(m_0(W,X))|W^{(1)}=w\right\}}{E\left\{\rho_2(m_0(W,X))|W^{(1)}=w\right\}}.$$

Let $\mathcal{H}$ denote the subspace of functions on R^{s_1} that take an additive form

$$\mathcal{H} = \left\{h : h(\mathbf{w}) = h_1(w_1) + \cdots + h_{s_1}(w_{s_1}), Eh_j(W_j)^2 < \infty \text{ and } \right.$$
$$\left. Eh_j(W_j) = 0\right\},$$

and for any function f on R^{s_1}, the projection of f onto $\mathcal{H}$ is defined as the minimizer of

$$\inf_{h \in \mathcal{H}} E\left\{\rho_2\big(m_0(W, X)\big)\big(f(W^{(1)}) - h(W^{(1)})\big)^2\right\}.$$

The preceding definition of projection trivially extends to the case where f is a vector by componentwise projection.

Let $\Gamma^{add}(w) = \sum_{l=1}^{s_1} \Gamma_l(w_l)$, $\Gamma_l(w_l) = \{\Gamma_{(1)l}(w_l), \ldots, \Gamma_{(s_2)l}(w_l)\}^{\mathrm{T}}$, be the projection of Γ onto $\mathcal{H}$ as defined before. In what follows, write $A^{\otimes 2} = AA^{\mathrm{T}}$ for any matrix or vector A. Define the $s_2 \times s_2$ matrix $\Xi = E\big[\rho_2\big(m_0(W, X), Y\big)\big\{X^{(1)} - \Gamma^{add}\big(W^{(1)}\big)\big\}^{\otimes 2}\big]$. Let $U_i^{(1)\mathrm{T}} := \big(Z_i^{(1)\mathrm{T}}, X_i^{(1)\mathrm{T}}\big)$. The following assumptions are used in the proofs.

(A1) The covariates in the nonparametric part satisfy $W_j \in [0, 1]$, and the covariates X_j in the parametric part have finite fourth-order moments.

(A2) The eigenvalues of $E\big\{U_i^{(1)} U_i^{(1)\mathrm{T}}\big\}$ are bounded away from zero and infinity.

(A3) $q_1(m_0(W, X), Y)$ have finite second moments, and for some constants $C > c > 0$ we have $c < \rho_2(m_0(W, X), Y), \rho_2'(m_0(W, X), Y) < C$.

(A4) For $1 \le j \le s_1$, α_{0j} satisfies a Lipschitz condition of order $d > \frac{1}{2}$: $|\alpha_{0j}^{(\lfloor d \rfloor)}(t) - \alpha_{0j}^{(\lfloor d \rfloor)}(s)| \le C|s - t|^{d - \lfloor d \rfloor}$, where $\lfloor d \rfloor$ is the biggest integer strictly smaller than d and $\alpha_{0j}^{(\lfloor d \rfloor)}(t)$ is the $\lfloor d \rfloor$th derivative of $\alpha_{0j}(t)$. In addition, the same assumption holds for $\Gamma_{(l)j}, 1 \le j \le s_1, 1 \le l \le s_2$. The order of the B-spline used satisfies $u \ge d + 2$.

(A5) $E[|q_1^m(m_0(W, X), Y)|] \le m!/2J^{m-2}R^2, m = 2, 3, \ldots,$ for some constants $J, R > 0$.

(A6) The weights in the adaptive group LASSO satisfy

$$\sqrt{n}\left\{\sqrt{\log(p_1 K \vee n)} + \sqrt{K + s_2 + n/K^{2d}} + \sqrt{n}\left(\lambda_1 \|\omega_1^{(1)}\|^2 + \lambda_2 \|\omega_2^{(1)}\|^2\right)\right\}$$
$$= o\left(n\lambda_1 \omega_{1j}\right), \qquad \text{for } s_1 + 1 \le j \le p_1;$$

$$\sqrt{n}\left\{\sqrt{\log(p_2 \vee n)} + \sqrt{K + s_2 + n/K^{2d}} + \sqrt{n}\left(\lambda_1 \|\omega_1^{(1)}\|^2 + \lambda_2 \|\omega_2^{(1)}\|^2\right)\right\}$$
$$= o\big(n\lambda_2 \omega_{2j}\big), \qquad \text{for } s_2 + 1 \le j \le p_2.$$

Most of the conditions imposed are quite standard in the literature, in particular in Wang et al. (2011). Because the basis B_{jk} is appropriately normalized, Assumption (A2) on the eigenvalues of $EU_i^{(1)}U_i^{(1)\mathrm{T}}$ is natural. When $(K+s_2)^2/n \to 0$, this implies that the eigenvalues of the sample version, $\sum_i U_i^{(1)}U_i^{(1)\mathrm{T}}/n$, are also bounded away from zero and infinity with probability converging to 1. On the other hand, (A2) might not always hold because the dimension of $X^{(1)}$ diverges with sample size. However, the boundedness of eigenvalues of $EX_i^{(1)}X_i^{(1)\mathrm{T}}$ is typically assumed even in models with a diverging number of parameters, for example, in Fan and Peng (2004) and Lam and Fan (2008). This of course holds when the covariates are independent of each other. When (A2) holds the convergence rate of the estimators as stated in Theorem 2 later in this section is very similar to the fixed-dimensional case. If the minimum eigenvalue of $EU_i^{(1)}U_i^{(1)\mathrm{T}}$ is however of order t_n, say, with $t_n = o(1)$, then from the proof of Theorem 2 it is easy to see that the convergence rate will need to be multiplied by a factor t_n^{-1}, and this factor will also appear in various places such as in Assumption (A6). Furthermore, in general it is hard to imagine at what rate the minimum eigenvalue might converge to zero even if it does, and so we assume (A2) for simplicity, following the existing literature. Because $q_1(x,y) = -(y - g^{-1}(x))\rho_1(x)$, (A5) is an assumption on the moments of the noise. Such an assumption related to the tail decay rate of the noise distribution seems necessary in high-dimensional problems. Huang et al. (2010) used a sub-Gaussian noise assumption that is more stringent than our assumption here, which roughly assumes an exponential decay in the tail of the noise distribution. For example, for the Poisson model, the tail has exponential decay but is not sub-Gaussian.

Assumption (A6) on the weights is crucial for the adaptive group LASSO estimator to achieve model selection consistency. Roughly speaking, these two expressions require that weights associated with zero components be large enough, for the nonparametric and parametric parts, respectively. These conditions imply some constraints on the allowable values of p_1 and p_2, as we will discuss in Section 2.2.

We first show that when the weights are appropriately specified, the adaptive LASSO correctly identifies the zero coefficients.

THEOREM 1. *Suppose s_1 is bounded. Under Assumptions (A1)–(A6), we have $\hat{a}_j = 0, s_1+1 \le j \le p_1$, and $\hat{\beta}_j = 0, s_2+1 \le j \le p_2$ with probability converging to 1.*

Let $\omega_1^{(1)} = (\omega_{11}, \dots, \omega_{1s_1})^{\mathrm{T}}$, $\omega_2^{(1)} = (\omega_{21}, \dots, \omega_{2s_2})^{\mathrm{T}}$ be the weights associated with the nonzero components. The exact requirements on the weights are stated in Assumption (A6). The next theorem states the convergence rate for the estimator in terms of sum of squares of estimation error of all nonzero components.

THEOREM 2. *Under the same assumptions as in Theorem 1, and in addition*

$$K \to \infty, \qquad (K+s_2)^2/n \to 0, \qquad \lambda_1^2\|\omega_1^{(1)}\|^2 + \lambda_2^2\|\omega_2^{(1)}\|^2 \to 0,$$

the estimator $(\hat{\alpha}, \hat{\beta})$ *in (3) satisfies*

$$\sum_{j=1}^{s_1} \|\hat{\alpha}_j(t) - \alpha_{0j}(t)\|^2 + \sum_{j=1}^{s_2} |\hat{\beta}_j - \beta_{0j}|^2$$
$$= O_P\left(\frac{K+s_2}{n} + \frac{1}{K^{2d}} + \lambda_1^2\|\omega_1^{(1)}\|^2 + \lambda_2^2\|\omega_2^{(1)}\|^2\right), \qquad \textbf{(4)}$$

where $\hat{\alpha}_j(t) = \sum_k \hat{\alpha}_{jk} B_{jk}(t)$.

Theorem 2 indicates that the rate of convergence is dominated by the stochastic error of estimating nonparametric and parametric components (the first term), the spline approximation (the second term), and the bias arisen from penalization (the last two terms). Under slightly stronger assumptions, the estimator for the parametric components can be shown to be asymptotically normal. The limiting distribution is the same as that of Wang et al. (2011). We also note that our assumptions imply that Ξ is invertible. The arguments are as follows. Note that for an $s_2 \times s_1 K$ matrix A, we have $(A, I)U_i^{(1)} = X_i^{(1)} + AZ_i^{(1)}$, where I is the $s_2 \times s_2$ identity matrix. By the smoothness assumption (A4), there is a matrix A such that the difference between $E(A, I)U_i^{(1)}U_i^{(1)\mathrm{T}}(A, I)^{\mathrm{T}}$ and $E\{X_i^{(1)} - \Gamma^{add}(W_i^{(1)})\}^{\otimes 2}$ is $o(1)$, elementwise. Furthermore, given any s_2-vector u with $\|u\| = 1$, $E\{u^{\mathrm{T}}(A, I)U_i^{(1)}\}^{\otimes}$ is bounded away from zero by Assumption (A2) and assuming that $\|u^{\mathrm{T}}(A, I)\| \geq \|u\| = 1$. This implies that the eigenvalues of $E(A, I)U_i^{(1)}U_i^{(1)\mathrm{T}}(A, I)^{\mathrm{T}}$ are bounded away from zero, and so are those of $E\{X_i^{(1)} - \Gamma^{add}(W_i^{(1)})\}^{\otimes 2}$. Thus Ξ is invertible given that $\rho_2(m_0(W, X))$ is bounded below by a positive constant.

THEOREM 3 (Asymptotic normality). *Under the same assumptions as in Theorem 1, and assuming that* $\sqrt{n}(\lambda_1\|\omega_1^{(1)}\| + \lambda_2\|\omega_2^{(1)}\|) \to 0$, $\sqrt{n}/K^d \to 0$,

$$\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0^{(1)}\right) \to N\left(0, \Xi^{-1}\right) \quad \textit{in distribution},$$

where $\hat{\beta}^{(1)}$, $\beta_0^{(1)}$ *contains the first* s_2 *components of* $\hat{\beta}$, β_0, *respectively.*

With regard to the consistency and asymptotic normality results presented previously, we also mention some recent works (Leeb and Pötscher, 2005, 2008) that have studied the (uniform) consistency of estimates of distribution functions of the penalized estimators. In particular, they have shown that if an estimator is consistent in model selection, it is impossible to possess the oracle property (asymptotic normality) *uniformly* in a neighborhood of zero coefficients. These results are of great interest, and uniform convergence is more relevant when many of the true regression parameters are very close to zero, which commonly arises in cases where the number of parameters diverges with sample size. Thus one should be careful when using the preceding results to conduct inferences in practice. Although we acknowledge that taking the parameters in our model to be fixed is partly subjective and in some cases modeling the parameters as changing with sample size is more appropriate, the oracle property for fixed parameter as studied here is still an important issue.

2.2. Doubly Penalized Initial Lasso Estimator

In the adaptive LASSO penalty, the weights associated with the zero components are required to be larger than those associated with the nonzero components as indicated in Assumption (A6). Of course, it remains unknown to the analysts which components are zeros before statistical analysis. Following Zou (2006), we can first obtain an initial estimator with LASSO penalty (all weights set to be 1), by solving the following penalized quasi-likelihood:

$$(\tilde{a},\tilde{\beta}) = \arg\min_{a,\beta}\sum_i Q\left(g^{-1}\left(Z_i^{\mathrm{T}}a+X_i^{\mathrm{T}}\beta\right),Y_i\right)+n\lambda_{01}\sum_{j=1}^{p_1}\|a_j\|_{D_j}+n\lambda_{02}\sum_{j=1}^{p_2}|\beta_j| \tag{5}$$

and then setting

$$\omega_{1j} = \begin{cases} 1/\|\tilde{\alpha}_j\|_{D_j} & \text{if } \|\tilde{\alpha}_j\|_{D_j}>0 \\ \infty & \text{if } \|\tilde{\alpha}_j\|_{D_j}=0 \end{cases} \text{ and } \omega_{2j} = \begin{cases} 1/|\tilde{\beta}_j| & \text{if } |\tilde{\beta}_j|>0 \\ \infty & \text{if } |\tilde{\beta}_j|=0 \end{cases},$$

with the hope that if some component is zero, its estimate will also be close to zero, making the weight larger.

Let $S_1 := \{1,\dots,s_1\}$, $S_1^c := \{s_1+1,\dots,p_1\}$, $S_2 := \{1,\dots,s_2\}$, $S_2^c := \{s_2+1,\dots,p_2\}$ so that S_1 and S_2 contain the indexes of the nonzero components. Indeed, the following theorem presents the estimation accuracy of such an initial estimator.

The following additional condition is needed in Theorem 4, which follows, for the initial LASSO estimator. This assumption is a variant of similar assumptions in Bickel et al. (2009) that only focused on parametric models. We defer a more detailed discussion on sufficient conditions for it to hold to Section A.4 of the Appendix. For simplicity of discussion in the paper, we assume that k_1 defined in what follows are bounded away from zero. However we keep it explicit in the proofs for generality.

(A7) (Restricted eigenvalue condition). For some $\gamma > 0$, $\inf\sum_i v^{\mathrm{T}}U_iU_i^{\mathrm{T}}v/n =: k_1 > 0$, where $U_i^{\mathrm{T}} = \left(Z_i^{\mathrm{T}},X_i^{\mathrm{T}}\right)^{\mathrm{T}}$ and the infimum is taken over the vector $v = \left(v_{a1},\dots,v_{ap_1},v_{\beta 1},\dots,v_{\beta p_2}\right)^{\mathrm{T}}$ where $v_{aj}\in R^K$ and $v_{\beta j}$ is a scalar (thus v is (p_1K+p_2)-dimensional), which satisfies $\|v\|=1$ and

$$\sum_{j\in S_1^c}\|v_{aj}\|_{D_j}+\left(\lambda_{02}/\lambda_{01}\right)\sum_{j\in S_2^c}|v_{\beta j}| \le (1+\gamma)\left\{\sum_{j\in S_1}\|v_{aj}\|_{D_j}+(\lambda_{02}/\lambda_{01})\sum_{j\in S_2}|v_{\beta j}|\right\}.$$

THEOREM 4. *Under regularity assumptions (A1)–(A5) and (A7), in addition to*

$$\sqrt{\frac{\log(p_1 K \vee n)}{n}} = o(\lambda_{01}), \qquad \sqrt{\frac{\log(p_2 \vee n)}{n}} = o(\lambda_{02}),$$
$$\sqrt{s_1}\lambda_{01} + \sqrt{s_2}\lambda_{02} \to 0,$$

we have the convergence rate

$$\|\tilde{\alpha} - \alpha_0\| + \|\tilde{\beta} - \beta_0\| = O_P\big(\sqrt{s_1}\lambda_{01} + \sqrt{s_2}\lambda_{02}\big),$$

where $\tilde{\alpha}_j = \sum_k \tilde{a}_{jk} B_{jk}$.

Assumption (A6) is stated in terms of general weights. When the weights are determined by the initial LASSO estimator as $\omega_{1j} = 1/\|\tilde{a}_j\|, \omega_{2j} = 1/|\tilde{\beta}_j|$, along with other simplifying assumptions, we will now see that (A6) can be reduced to simpler and clearer expressions. For simplicity of discussion we will assume $\|\alpha_{0j}\|, 1 \le j \le s_1, |\beta_{0j}|, 1 \le j \le s_2$ are bounded away from zero, s_1 and s_2 are bounded, and $K \sim n^{1/(2d+1)}$, which is the optimal choice asymptotically according to the rates (4). For notational simplicity in the following discussion, we also assume $p_1 K = p_2$, which means that the X's and Z's parts after approximating the nonparametric functions contain the same number of parameters. Choosing $\lambda_{01} = \lambda_{02} \sim \sqrt{\log(p_2 \vee n) b_n / n}$ with some $b_n \to \infty$ arbitrarily slowly, the convergence rate of the initial estimator is $O_P(\lambda_{01}) = O_P(\sqrt{\log(p_2 \vee n) b_n / n}) = o_P(1)$. This implies $\|\omega_1^{(1)}\| = O_P(1), \|\omega_2^{(1)}\| = O_P(1)$ when the LASSO estimator is used to specify the weights. Thus if

$$\lambda_1 = \lambda_2 = O\left(\sqrt{\frac{K}{n}}\right), \tag{6}$$

the final term in the rate (4) is small enough and the penalization does not adversely affect the convergence rate. Using that $\omega_{1j}, j > s_1, \omega_{2j}, j > s_2$ are at least of order $\sqrt{n/(\log(p_2 \vee n) b_n)}$, it is easy to see that the second equation in Assumption (A6) is satisfied if

$$\sqrt{n}\big\{\sqrt{\log(p_2 \vee n)} + \sqrt{K} + \sqrt{n}\big(\lambda_1 + \lambda_2\big)\big\} << n\lambda_2 \sqrt{\frac{n}{\log(p_2 \vee n)}}. \tag{7}$$

This obviously imposes some constraint on p_2. In particular, by some simple calculations, for λ_2 to exist that satisfies both (6) and (7), we would require that $p_2 = o(\exp\{n^{(d+1)/(2d+1)}\})$. Similarly Assumption (A6) implies $p_1 = o(\exp\{n^{(d+1)/(2d+1)}\}/K)$.

2.3. Tuning Parameters Selection and Implementation

As commonly adopted, we use cubic splines (spline order $u = 4$) in all our numerical examples. For the doubly penalized adaptive LASSO estimator, both the number of spline basis K and the regularization parameters λ_1 and λ_2 are important in tuning the performance of the estimator. In this study, we use fivefold cross-validation to choose all three parameters simultaneously. More specifically, for $K = 3, 4, 5, 6$ and a grid of values of λ_1 and λ_2, the estimates obtained on the training data are evaluated on test data in terms of the values of quasi-likelihood, and the combination of parameters achieving the maximum is chosen. Similarly we also apply fivefold cross-validation for the initial LASSO estimator.

The minimization problem (3) (and also (5)) is solved by locally quadratic approximation as discussed in Fan and Li (2001). Specifically, given initial values $(a^{(0)}, \beta^{(0)})$, after approximation by a quadratic function and getting rid of constants irrelevant for optimization, we obtain the function

$$\min \sum_i Q\big(g^{-1} Z_i^{\mathrm{T}} a + X_i^{\mathrm{T}} \beta\big), Y_i) + \frac{n\lambda_1}{2} \sum_j \omega_{1j} \frac{\|a_j\|_{D_j}^2}{\|a_j^{(0)}\|_{D_j}} + \frac{n\lambda_2}{2} \sum_j \omega_{2j} \frac{\beta_j^2}{|\beta_j^{(0)}|},$$

which can be solved by the Newton–Raphson algorithm.

3. SIMULATION STUDY

Here we assess the finite-sample performance of the proposed procedure. We consider the semiparametric logistic regression model

$$\operatorname{logit}\big\{P\big(Y = 1|X, W\big)\big\} = \sum_{j=1}^{p_1} \alpha_j\big(W_j\big) + \sum_{j=1}^{p_2} \beta_j X_j.$$

For the nonparametric part, we set $\alpha_1(x) = \alpha_3(x) = 8\sin(2\pi x)$, $\alpha_2(x) = \alpha_4(x) = 6\cos(2\pi x)$, and other components are zeros. For the parametric part, we set $\beta = (10, \ldots, 5, 0, \ldots, 0)^{\mathrm{T}}$. Thus in our simulation example, $s_1 = 4$, $s_2 = 6$. The covariates are generated as follows: we first generate a $p = p_1 + p_2$-dimensional random Gaussian vector $(V_1, \ldots, V_p)$ with $\operatorname{Cov}(V_j, V_{j'}) = (0.3)^{|j-j'|}$ and then apply the cumulative distribution function of standard normal distribution to each component of V_j, and finally we use the first p_1 components in the nonparametric part and the rest in the parametric part. Thus all the covariates are marginally uniformly distributed with correlations present.

We consider four different scenarios: (i) $n = 100$ and $p = 100 (p_1 = p_2 = 50)$, (ii) $n = 100$ and $p = 200 (p_1 = p_2 = 100)$, (iii) $n = 200$ and $p = 100 (p_1 = p_2 = 50)$, and (iv) $n = 200$ and $p = 200 (p_1 = p_2 = 100)$, respectively. For each scenario, 500 data sets are generated, and the model is fitted and tuning parameters are chosen as described in Section 2.3.

In Table 1, we show the number of true positives (TP) and false positives (FP) for both the LASSO estimator and the adaptive LASSO estimator. From the table, both the number of TPs and the number of FPs for the adaptive LASSO estimator is smaller, but the decrease in the number of FPs is more dramatic. The ability to correctly identify the nonzero components is significantly increased when n

TABLE 1. Model selection results for the simulation study.

n	p	Method	TP.NONPAR	FP.NONPAR	TP.PAR	FP.PAR
100	100	LAS	3.86 (0.60)	18.19 (7.26)	5.34 (0.77)	21.18 (4.99)
100	100	ALAS	3.60 (0.74)	4.36 (4.28)	4.08 (1.06)	5.87 (3.67)
100	200	LAS	3.10 (1.57)	26.17 (25.27)	5.02 (1.13)	44.88 (30.44)
100	200	ALAS	2.80 (1.51)	4.54 (3.89)	4.18 (1.39)	8.90 (5.61)
200	100	LAS	4 (0)	30.88 (5.79)	5.93 (0.25)	26.58 (3.80)
200	100	ALAS	4 (0)	10.80 (4.29)	5.51 (0.61)	6.74 (4.21)
200	200	LAS	4 (0)	42.58 (15.34)	5.82 (0.40)	47.04 (8.44)
200	200	ALAS	3.99 (0.06)	11.77 (6.77)	5.41 (0.67)	13.69 (6.23)

TP.NONPAR = number of estimated nonzero components in the nonparametric part that are truly nonzero. FP.NONPAR = number of components estimated to be nonzero but that are actually zero in the nonparametric part. TP.PAR and FP.PAR are interpreted similarly for the parametric part. LAS denotes the initial LASSO estimator, and ALAS denotes the adaptive LASSO estimator. The numbers in parentheses are the corresponding standard errors

increases from 100 to 200. To see how different significant components could be missed during estimation, we show the number of times each component is selected among the 500 generated data sets in Table 2. As expected, in general smaller coefficients are more likely missed by the estimator. Table 3 shows the average root mean squared errors (RMSE), together with standard deviations based on simulation, for the eight nonzero components (RMSEs for α_3 and α_4 are not shown because they are similar to α_1 and α_2, respectively). The table shows that the adaptive LASSO estimator performs better than the initial LASSO estimator.

Because our ultimate goal is to use the estimated model to predict future responses, it is important to look at the prediction accuracy of the LASSO and adaptive LASSO estimators. For a comparison, we also examine the oracle estimator where no penalties are used (we also use fivefold cross-validation to selection the number of knots here). For that purpose, in each scenario and for

TABLE 2. The number of times each nonzero component is selected among the 500 replicates.

n	p	Method	Nonparametric part				Parametric part					
100	100	LAS	490	483	489	470	477	468	469	457	434	366
100	100	ALAS	485	428	480	407	420	396	388	342	292	205
100	200	LAS	391	376	401	382	458	467	455	420	382	332
100	200	ALAS	382	320	388	313	404	408	384	333	311	254
200	100	LAS	500	500	500	500	500	500	500	495	500	470
200	100	ALAS	500	500	500	500	495	489	490	468	448	368
200	200	LAS	500	500	500	500	500	498	494	493	477	450
200	200	ALAS	500	499	500	499	495	492	477	463	429	353

There are 4 nonzero components in the nonparametric part and 6 nonzero components in the parametric part

TABLE 3. Average root mean squared errors for each of the 8 nonzero components based on 500 replications in each scenario. The numbers in parentheses are the corresponding standard deviations

n	p	Method	α_1	α_2	β_1	β_2	β_3	β_4	β_5	β_6
100	100	ORA	0.80 (0.87)	0.62 (0.61)	0.38 (0.44)	0.33 (0.39)	0.32 (0.37)	0.27 (0.33)	0.23 (0.26)	0.21 (0.25)
100	100	LAS	1.87 (0.14)	1.47 (0.11)	0.84 (0.12)	0.75 (0.12)	0.67 (0.12)	0.59 (0.11)	0.52 (0.11)	0.45 (0.11)
100	100	ALAS	0.90 (0.40)	0.89 (0.24)	0.43 (0.29)	0.38 (0.26)	0.39 (0.26)	0.37 (0.25)	0.36 (0.24)	0.37 (0.21)
100	100	ORA	0.84 (0.88)	0.62 (0.60)	0.40 (0.45)	0.36 (0.40)	0.30 (0.36)	0.30 (0.35)	0.25 (0.30)	0.21 (0.25)
100	200	LAS	2.01 (0.21)	1.56 (0.14)	0.89 (0.14)	0.78 (0.15)	0.71 (0.13)	0.61 (0.13)	0.54 (0.12)	0.47 (0.10)
100	200	ALAS	0.96 (0.44)	0.92 (0.26)	0.50 (0.28)	0.45 (0.26)	0.44 (0.26)	0.41 (0.25)	0.39 (0.23)	0.40 (0.20)
200	100	ORA	0.21 (0.64)	0.16 (0.51)	0.08 (0.31)	0.08 (0.27)	0.07 (0.25)	0.08 (0.30)	0.06 (0.20)	0.05 (0.17)
200	100	LAS	0.35 (0.82)	0.27 (0.63)	0.16 (0.38)	0.13 (0.32)	0.12 (0.30)	0.11 (0.26)	0.10 (0.23)	0.08 (0.19)
200	100	ALAS	0.23 (0.59)	0.21 (0.51)	0.13 (0.34)	0.10 (0.28)	0.11 (0.28)	0.09 (0.24)	0.09 (0.23)	0.07 (0.19)
200	200	ORA	0.20 (0.53)	0.15 (0.39)	0.09 (0.28)	0.08 (0.24)	0.07 (0.23)	0.07 (0.20)	0.06 (0.19)	0.05 (0.16)
200	200	LAS	0.41 (0.90)	0.31 (0.69)	0.18 (0.40)	0.16 (0.35)	0.14 (0.32)	0.12 (0.28)	0.11 (0.25)	0.09 (0.21)
200	200	ALAS	0.32 (0.73)	0.28 (0.63)	0.15 (0.36)	0.12 (0.30)	0.12 (0.29)	0.11 (0.26)	0.10 (0.24)	0.08 (0.20)

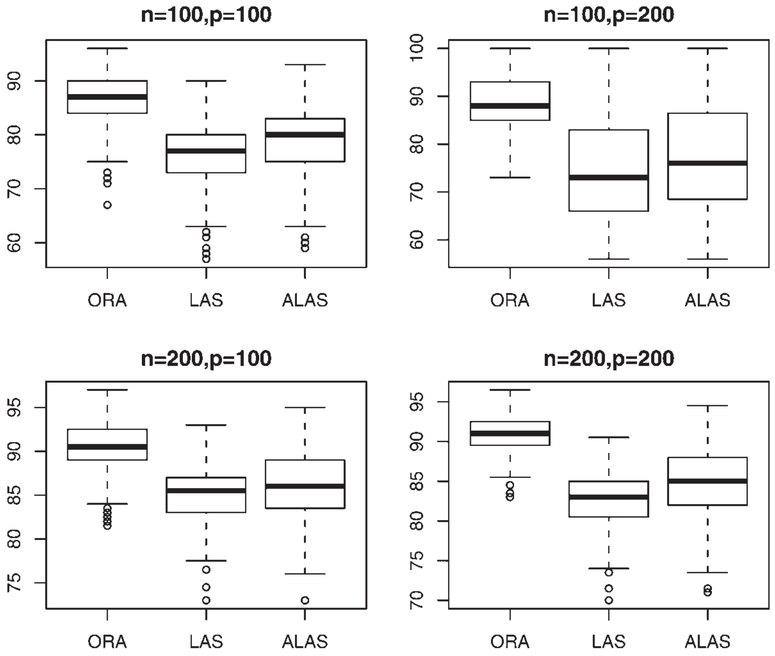

FIGURE 1. Number of correct predictions for 500 replicates. Three estimators, the oracle estimator, the LASSO estimator, and the adaptive LASSO estimator, are compared.

each of the 500 generated data sets, another 100 observations are simulated and the number of times correct predictions are made is recorded. Figure 1 shows the percentage of correct predictions for 500 generated data sets in each of the four scenarios. We see that the adaptive LASSO estimator performs better than LASSO, even though Table 1 shows it has more false negatives compared to LASSO.

4. CONCLUDING REMARKS

We have studied variable selection for GAPLM when the dimensions of parametric and nonparametric components diverge at exponential rates of the sample size. We used polynomial splines to approximate nonparametric functions and the adaptive LASSO penalty to eliminate insignificant components. An important result is that the proposed procedure is selection consistent and the resulting estimators for the selected parametric components are asymptotically normal with a proper choice of tuning parameters. The methods were shown to be promising through numerical examples.

In implementation, we combine locally quadratic approximation with the Newton–Raphson algorithm in solving the objective function. In large dimensions, this algorithm is relatively slow. More efficient algorithms, such as coordinate descent (Friedman, Hastie, Höfling, and Tibshirani, 2007), may be adapted to the proposed procedure for gains in reduce of computational burden.

Based on our numerical experience, we have observed that the number of false positives is relatively large when using fivefold cross-validation to select the tuning parameters. In the literature, both generalized cross-validation (GCV) and Bayesian information criterion (BIC) have been advocated for selection of tuning parameters for penalized sparse models (Huang et al., 2010; Wang, Li, and Tsai, 2007; Wang and Xia, 2009; Wang et al., 2011). But our experience advises that GCV usually results in even more false positives, whereas BIC seems to be too stringent in model selection, sometimes even choosing the null model. It remains a challenging endeavor to achieve a compromise between these two prospects and needs further investigation.

REFERENCES

Andrews, D. (1991) Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–345.

Andrews, D. & Y. Whang (1990) Additive interactive regression models: Circumvention of the curse of dimensionality. *Econometric Theory* 6, 466–479.

Belloni, A. & V. Chernozhukov (2012) Post-l1-penalized estimators in high-dimensional linear regression models. *Bernoulli*, forthcoming.

Bickel, P., Y. Ritov, & A. Tsybakov (2009) Simultaneous analysis of LASSO and Dantzig selector. *Annals of Statistics* 37, 1705–1732.

Buja, A., T. Hastie, & R. Tibshirani (1989) Linear smoothers and additive models. *Annals of Statistics* 17, 453–510.

Burda, M.C., W. Härdle, M. Müller, & A. Werwatz (1998) Semiparametric analysis of German East-West migration intentions: Facts and theory. *Journal of Applied Econometrics* 13, 525–541.

Chen, H. (1988) Convergence rates for parametric components in a partly linear model. *Annals of Statistics* 16, 136–146.

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6, 5549–5632.

Donald, S. & W. Newey (1994) Series estimation of semilinear models. *Journal of Multivariate Analysis* 50, 30–40.

Du, P., S.G. Ma, & H. Liang (2010) Penalized variable selection procedure for Cox models with semiparametric relative risk. *Annals of Statistics* 38, 2092–2117.

Engle, R., C. Granger, J. Rice, & A. Weiss (1986) Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–320.

Fan, J.Q. & R. Z. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Fan, J.Q. & H. Peng (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32, 928–961.

Friedman, J., T. Hastie, H. Höfling, & R. Tibshirani (2007) Pathwise coordinate optimization. *Annals of Applied Statistics* 1, 302–332.

Härdle, W., H. Liang, & J.T. Gao (2000) *Partially Linear Models*. Springer Physica.

Härdle, W., E. Mammen, & M. Müller (1998) Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association* 93, 1461–1474.

Heckman, N.E. (1986) Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B* 48, 244–248.

Hiirdle, W., M. Müller, S. Sperlich, & A. Werwatz (2004) *Nonparametric and Semiparametric Models*. Springer-Verlag.

Horowitz, J. (1998) *Semiparametric Methods in Econometrics*. Springer-Verlag.

Huang, J. (1998) Functional ANOVA models for generalized regression. *Journal of Multivariate Analysis* 67, 49–71.

Huang, J., J. L. Horowitz, & F. Wei (2010) Variable selection in nonparametric additive models. *Annals of Statistics* 38, 2282–2313.

Huang, J.H.Z., C.O. Wu, & L. Zhou (2004) Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14, 763–788.

Juhl, T. & Z. Xiao (2005) Partially linear models with unit roots. *Econometric Theory* 21, 877–906.

Kneib, T., S. Konrath, & L. Fahrmeir (2011) High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society, Series C* 60, 51–70.

Lam, C. & J. Fan (2008) Profile-kernel likelihood inference with diverging number of parameters. *Annals of Statistics* 36, 2232–2260.

Leeb, H. & B. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.

Leeb, H. & B. Pötscher (2008) Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142, 201–211.

Li, Q. (2000) Efficient estimation of additive partially linear models. *International Economic Review* 41, 1073–1092.

Li, Q. & T. Stengos (1996) Semiparametric estimation of partially linear panel data models. *Journal of Econometrics* 71, 389–397.

Li, Q. & J.M. Wooldridge (2002) Semiparametric estimation of partially linear models for dependent data with generated regressors. *Econometric Theory* 18, 625–645.

Li, R. & H. Liang (2008) Variable selection in semiparametric regression modeling. *Annals of Statistics* 36, 261–286.

Linton, O. & W. Härdle (1996) Estimation of additive regression models with known links. *Biometrika* 83, 529–540.

Linton, O. & J. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.

Liu, X., L. Wang, & H. Liang (2011) Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica* 21, 1225–1248.

Marx, B. & P. Eilers (1998) Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* 28, 193–209.

Müller, M. & B. Rönz (2000) *Credit Scoring Using Semiparametric Methods*. Springer Lecture Notes in Statistics. Springer-Verlag.

Newey, W. (1997) Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147–168.

Robinson, P.M. (1988) Root n-consistent semiparametric regression. *Econometrica* 56, 931–954.

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

Severini, T.A. & J.G. Staniswalis (1994) Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* 89, 501–511.

Speckman, P.E. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B* 50, 413–436.

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

Stone, C. (1994) The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics* 22, 118–171.

Su, L. & S. Jin (2010) Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics* 157, 18–33.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

van der Geer, S.A. (2000) *Applications of Empirical Process Theory*. Cambridge University Press.

Wang, H., R. Li, & C.L. Tsai (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94, 553–568.

Wang, H.S. & Y.C. Xia (2009) Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* 104, 747–757.

Wang, L., X. Liu, H. Liang, & R. Carroll (2011) Estimation and variable selection for generalized additive partially linear models. *Annals of Statistics* 39, 1827–1851.

Wood, S. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99, 673–686.

Wu, G. & Z. Xiao (2002) A generalized partially linear model of asymmetric volatility. *Journal of Empirical Finance* 9, 287–319.

Xue, L. (2009) Consistent variable selection in additive models. *Statistica Sinica* 19, 1281–1296.

Xue, L. & L. Yang (2006) Additive coefficient modeling via polynomial spline. *Statistica Sinica* 16, 1423–1446.

Yatchew, A. (2003) *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

Yu, K., B. Park, & E. Mammen (2008) Smooth backfitting in generalized additive models. *Annals of Statistics* 36, 228–260.

Yuan, M. & Y. Lin (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68, 49–67.

Zhang, C.H. & J. Huang (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* 36, 1567–1594.

Zhang, H., G. Cheng, & Y. Liu (2011) Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* 106, 1099–1112.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.

Zou, H. & R.Z. Li (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 36, 1509–1533.

APPENDIX

In this Appendix, we prepare several preliminary results and give the proofs of the main results.

A.1. Preliminary Results. Let $a_{0j} = (a_{0j1}, \ldots, a_{0jK})^{\mathrm{T}}, 1 \le j \le p_1$ be the coefficients in spline approximation of α_{0j} that satisfies the approximation property $\| \sum_k a_{0jk} B_{jk}(x) - \alpha_{0j}(x) \|_\infty = O(K^{-d})$ and set $a_0 = (a_{01}^{\mathrm{T}}, \ldots, a_{0p_1}^{\mathrm{T}})^{\mathrm{T}}$. Denote $a_0^* = (a_0^{\mathrm{T}}, \beta_0^{\mathrm{T}})^{\mathrm{T}}, \hat{a}^* = (\hat{a}^{\mathrm{T}}, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$. Let $a_0^{(1)*} = (a_0^{(1)\mathrm{T}}, \beta_0^{(1)\mathrm{T}})^{\mathrm{T}} = (a_{01}^{\mathrm{T}}, \ldots, a_{0s_1}^{\mathrm{T}}, \beta_{01}, \ldots, \beta_{0s_2})^{\mathrm{T}}$.

In some of the proofs that follow we will make use of the concept of subdifferential and subgradient. For a scalar ξ, it is well known that the subdifferential $\partial|\xi| = \{1\}, \{-1\}$, or $[-1, 1]$ depending on the sign of ξ. For any matrix A and vector b (as long as the dimensions are compatible), the subdifferential of $\|Ab\|$ with respect to b is

$$\partial\|Ab\| = \begin{cases} \{A^{\mathrm{T}}Ab/\|Ab\|\} & \text{if } Ab \neq 0 \\ \{A^{\mathrm{T}}Aa : ||Aa|| \le 1\} & \text{if } Ab = 0. \end{cases}$$

Note that the subdifferential at a point is defined to be a set and its elements are called subgradients. When $Ab = 0$ the subgradient is not unique. In what follows we will use the same notation, $\partial\|Ab\|$, to denote either subdifferential or a subgradient when the specific element picked has no significance in our proofs.

The strategy of proof is as follows. If it is known beforehand which coefficients are zero, we can optimize (3) over only the nonzero components in a and β, and the optimization problem (3) reduces to

$$\begin{aligned}(\hat{a}^{(1)}, \hat{\beta}^{(1)}) = \arg \min_{a^{(1)},\beta^{(1)}} \sum_{i=1}^{n} Q\big(g^{-1}\big(Z_i^{(1)\mathrm{T}} a^{(1)} + X_i^{(1)\mathrm{T}} \beta^{(1)}\big), Y_i\big) \\ + n\lambda_1 \sum_{j=1}^{s_1} \omega_{1j} \|a_j\|_{D_j} + n\lambda_2 \sum_{j=1}^{s_2} \omega_{2j} |\beta_j|. \qquad \textbf{(A.1)}\end{aligned}$$

We will first demonstrate the convergence rate and the asymptotic normality of the estimator (A.1) in Propositions A.1 and A.2, respectively; these convergence results are the same as stated in Theorems 2 and 3. Then we will proceed to show that under Assumption (A6) and relevant regularity conditions, the solution of (3) is exactly the same as the solution $(\hat{a}^{(1)}, \hat{\beta}^{(1)})$ of (A.1) if we set the rest of the components that do not appear in (A.1) as zero. This immediately implies that the zero components are identified correctly and also that the convergence rates and asymptotic normality for (A.1) carry over to the solution of (3). In fact, this relationship between the solution of (3) and that of (A.1) is the reason that we use the same notation for both minimizers.

Now as we have planned, consider the solution of (A.1).

PROPOSITION A.1. *Under Assumptions (A1)–(A5) and assuming that s_1 is bounded,*

$$\frac{K+s_2}{n} + \frac{1}{K^{2d}} + \lambda_1^2 \|\omega_1^{(1)}\|^2 + \lambda_2^2 \|\omega_2^{(1)}\|^2 \to 0,$$

the estimator $(\hat{a}^{(1)}, \hat{\beta}^{(1)})$ in (A.1) satisfies

$$\begin{aligned}&\sum_{j=1}^{s_1} \|\hat{\alpha}_j(t) - \alpha_{0j}(t)\|^2 + \sum_{j=1}^{s_2} |\hat{\beta}_j - \beta_{0j}|^2 \\ &\quad = O_P\left(\frac{K+s_2}{n} + \frac{1}{K^{2d}} + \lambda_1^2 \|\omega_1^{(1)}\|^2 + \lambda_2^2 \|\omega_2^{(1)}\|^2\right),\end{aligned}$$

where $\hat{\alpha}_j(t) = \sum_k \hat{a}_{jk} B_{jk}(t)$.

Proof. Define $U_i^{(1)\mathrm{T}} = \left(Z_i^{(1)\mathrm{T}}, X_i^{(1)\mathrm{T}}\right)$ as before, and also let $a_0^{(1)*} = (a_0^{(1)}, \beta_0^{(1)})^{\mathrm{T}}$. Because here we only consider the estimator in (A.1), we will omit the superscript $(\cdot)^{(1)}$ in what follows. Because $\hat{a}^* := (\hat{a}, \hat{\beta})^{\mathrm{T}}$ minimizes

$$\sum_i Q\left(g^{-1}\left(U_i^{\mathrm{T}} a^*\right), Y_i\right) + n\lambda_1 \sum_{j=1}^{s_1} \omega_{1j} \|a_j\|_{D_j} + n\lambda_2 \sum_{j=1}^{s_2} \omega_{2j} |\beta_j|$$

with respect to $a^* = (a, \beta)$, $\hat{a}^*$ satisfies the first-order condition

$$\sum_i q_1(U_i^{\mathrm{T}} \hat{a}^*, Y_i) U_i + v\left(\hat{a}^*\right) = 0, \qquad \textbf{(A.2)}$$

where $v(\hat{a}^*)$ is the (s_1K+s_2)-dimensional vector with the first Ks_1 components being $n\lambda_1\omega_{1j}\partial\|a_j\|_{D_j}$, $j=1,\dots,s_1$ and the last s_2 components being $n\lambda_2\omega_{2j}\partial|\beta_j|$, $1\le j\le s_2$. Obviously $\|v(\hat{a}^*)\| = O_P(n\lambda_1\|\omega_1^{(1)}\|+n\lambda_2\|\omega_2^{(1)}\|)$. Using Taylor expansion at $U_i^{\mathrm{T}}a_0^*$ for the first term in (A.2), we get

$$\left\|\sum_i q_1\left(U_i^{\mathrm{T}}a_0^*, Y_i\right)U_i + q_2\left(\cdot, Y_i\right)U_iU_i^{\mathrm{T}}\left(\hat{a}^*-a_0^*\right)\right\| + O_P\left(n\lambda_1\|\omega_1^{(1)}\|+n\lambda_2\|\omega_2^{(1)}\|\right)=0,$$

or

$$\left|\sum_i q_1(U_i^{\mathrm{T}}a_0^*, Y_i)U_i^{\mathrm{T}}(\hat{a}^*-a_0^*) + q_2(\cdot, Y_i)(\hat{a}^*-a_0^*)^{\mathrm{T}}U_iU_i^{\mathrm{T}}(\hat{a}^*-a_0^*)\right| \\ + O_P((n\lambda_1\|\omega_1^{(1)}\|+n\lambda_2\|\omega_2^{(1)}\|)\cdot\|\hat{a}^*-a_0^*\|)=0, \qquad \textbf{(A.3)}$$

where $q_2(\cdot, Y_i)$ is evaluated at some point between $U_i^{\mathrm{T}}a_0^*$ and $U_i^{\mathrm{T}}\hat{a}^*$.

Write $\mathbf{q}_1(\mathbf{U}a_0^*, Y) = \left\{q_1(U_1^{\mathrm{T}}a_0^*, Y_1),\dots,q_1(U_n^{\mathrm{T}}a_0^*, Y_n)\right\}$ and $\mathbf{U}=(U_1,\dots,U_n)^{\mathrm{T}}$. The first term in (A.3) can be written as $\mathbf{q}_1(\mathbf{U}a_0^*, Y)^{\mathrm{T}}\mathbf{U}(\hat{a}^*-a_0^*)$.

Note that $|\mathbf{q}_1(\mathbf{U}^{\mathrm{T}}a_0^*, Y)^{\mathrm{T}}\mathbf{U}(\hat{a}^*-a_0^*)|^2 \le \|P_U\mathbf{q}_1(\mathbf{U}^{\mathrm{T}}a_0^*, Y)\|^2\cdot\|\mathbf{U}(\hat{a}^*-a_0^*)\|^2$, where $P_U=\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}$ is the matrix of projection onto the columns of $\mathbf{U}$, and obviously $\|\mathbf{U}(\hat{a}^*-a_0^*)\|^2 = O_P(n\|\hat{a}^*-a_0^*)\|^2)$ by Assumption (A2). Besides, we have

$$\|P_U\mathbf{q}_1(\mathbf{U}a_0^*, Y)\|^2 \le 2\|P_U\mathbf{q}_1(\mathbf{m}, Y)\|^2 + 2\|P_U\{\mathbf{q}_1(\mathbf{U}a_0^*)-\mathbf{q}_1(\mathbf{m}, Y)\}\|^2,$$

where $\mathbf{m}=(m_1,\dots,m_n)^{\mathrm{T}}$ with $m_i=\sum_j\alpha_{0j}(W_{ij})+X_i^{\mathrm{T}}\beta_0$. The first term is of order $O_P(tr(P_U)) = O_P(K+s_2)$ because $\mathbf{q}_1(\mathbf{m}, Y)$ has mean zero conditional on the predictors. The second term is bounded by, using Taylor expansion and (A4), $\|\mathbf{Z}a_0-\alpha_0\|^2 = O_P(n/K^{2d})$. It follows that

$$|\mathbf{q}_1(\mathbf{U}a_0^*, Y)^{\mathrm{T}}\mathbf{U}(\hat{a}^*-a_0^*)|^2 = O_P\left\{\left(n(K+s_2)+\frac{n^2}{K^{2d}}\right)\|\hat{a}^*-a_0^*\|^2\right\}.$$

As a result, $\mathbf{q}_1(\mathbf{U}a_0^*, Y)^{\mathrm{T}}\mathbf{U}(\hat{a}^*-a_0^*) = O\left(\sqrt{K+s_2+nK^{-2d}}\cdot\sqrt{n}\|\hat{a}^*-a_0^*\|\right)$.

Using Assumptions (A2) and (A3), we have

$$c_1n\|\hat{a}^*-a_0^*\|^2 \le q_2\left(\cdot, Y_i\right)\left(\hat{a}^*-a_0^*\right)^{\mathrm{T}}U_iU_i^{\mathrm{T}}\left(\hat{a}^*-a_0^*\right) \le c_2n\|\hat{a}^*-a_0^*\|^2 \qquad \textbf{(A.4)}$$

for two positive constants c_1, c_2, with probability converging to 1. Thus (A.3) leads to $\|\hat{a}^*-a_0^*\| = O(\sqrt{(K+s_2)/n+K^{-2d}}+\lambda_1\|\omega_1^{(1)}\|+\lambda_2\|\omega_2^{(1)}\|)$, which in turn immediately implies the convergence rate (4). ■

PROPOSITION A.2 (Asymptotic normality). *In addition to the assumptions in Proposition A.1, suppose* $\sqrt{n}(\lambda_1\|\omega_1^{(1)}\|+\lambda_2\|\omega_2^{(1)}\|)\to 0$, $\sqrt{n}/K^{2d}\to 0$, $K/\sqrt{n}\to 0$, $s_2^2/n\to 0$. *Then*

$$\sqrt{n}\left(\hat{\beta}^{(1)}-\beta_0^{(1)}\right)\to N\left(0, \Xi^{-1}\right) \quad \textit{in distribution.}$$

Proof. Proof of asymptotic normality is similar to Theorem 2 in Wang et al. (2011), and we only sketch the proof here. The projection function $\Gamma_{(j)l}$ can be approximated by $\hat{\Gamma}_{(j)l} \in S_l^0$ with $\|\hat{\Gamma}_{(j)l} - \Gamma_{(j)l}\|_\infty = O(K^{-d})$, $1 \le j \le s_2, 1 \le l \le s_1$. Let $\hat{\Gamma}_l(W_{il}) = (\hat{\Gamma}_{(1)l}(W_{il}), \dots, \hat{\Gamma}_{(s_2)l}(W_{il}))^{\mathrm{T}}$ and $\hat{\Gamma}(W_i) = \sum_{l=1}^{s_1} \hat{\Gamma}_l(W_{il})$. Because $\hat{m}_i := Z_i^{\mathrm{T}}\hat{a} + X_i^{\mathrm{T}}\hat{\beta}, i = 1, \dots, n$ minimizes (A.1) over the set $\{\hat{m}_i + (X_i - \hat{\Gamma}(W_i))^{\mathrm{T}}\nu, i = 1, \dots, n : \nu \in R^{s_2}\}$, the first-order condition gives us

$$\sum_i q_1(\hat{m}_i, Y_i)(X_i - \hat{\Gamma}(Z_i)) + O_P(n\lambda_1\|\omega_1^{(1)}\| + n\lambda_2\|\omega_2^{(1)}\|) = 0.$$

Following the proof of Theorem 2 in Wang et al. (2011), this implies

$$\begin{aligned} 0 = {} & \frac{1}{\sqrt{n}}\sum_i q_1(m_{0i}, Y_i)\left\{X_i - \Gamma^{add}(W_i)\right\} \\ & + E\left[\rho_2(m_0)\left\{X_i - \Gamma^{add}(W_i)\right\}^{\otimes 2}\right]\sqrt{n}(\hat{\beta} - \beta_0) + o_p(1), \end{aligned}$$

and the asymptotic normality follows. ■

A.2. Proof of Theorem 1. As in Section 2.1 we define $Z_i = \left(B_{11}(W_{i1}), \dots, B_{p_1K}(W_{ip_1})\right)^{\mathrm{T}}$, and

$$Z_j = \begin{pmatrix} B_{j1}(W_{1j}) & \dots & B_{jK}(W_{1j}) \\ \vdots & \vdots & \vdots \\ B_{j1}(W_{nj}) & \dots & B_{jK}(W_{nj}) \end{pmatrix}_{n \times K}.$$

Although we somewhat abused our notation here, which one we mean is usually clear from the context or just from the subscript (i denotes sample and j denotes predictor). We also let $\mathbf{Z} = (Z_1, \dots, Z_i, \dots, Z_n)^{\mathrm{T}} = (Z_1, \dots, Z_j, \dots, Z_p)$.

Because $(\hat{a}^{(1)}, \hat{\beta}^{(1)})$ solves the optimization problem (A.1), we have that

$$Z_j^{\mathrm{T}}\mathbf{q}_1\left(\mathbf{Z}^{(1)}\hat{a}^{(1)} + X^{(1)}\hat{\beta}^{(1)}, Y\right) + n\lambda_1\omega_{1j}\partial\|\hat{a}_j\|_{D_j} = 0, j = 1, \dots, s_1, \tag{A.5}$$

$$X_j^{\mathrm{T}}\mathbf{q}_1\left(\mathbf{Z}^{(1)}\hat{a}^{(1)} + X^{(1)}\hat{\beta}^{(1)}, Y\right) + n\lambda_2\omega_{2j}\partial|\hat{\beta}_j| = 0, j = 1, \dots, s_2. \tag{A.6}$$

To show that $\left(\hat{a}^{(1)}, \hat{a}^{(2)} = 0, \hat{\beta}^{(1)}, \hat{\beta}^{(2)} = 0\right)$ also solves (3), we only need to verify the corresponding Karush–Kuhn–Tucker (KKT) conditions,

$$Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{Z}\hat{a} + X\hat{\beta}, Y) + n\lambda_1\omega_{1j}\partial\|\hat{a}_j\|_{D_j} = 0, \qquad j = 1, \dots, p_1, \tag{A.7}$$

$$X_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{Z}\hat{a} + X\hat{\beta}, Y) + n\lambda_2\omega_{2j}\partial|\hat{\beta}_j| = 0, \qquad j = 1, \dots, p_2. \tag{A.8}$$

First, equation (A.7) with $1 \le j \le s_1$ and (A.8) with $1 \le j \le s_2$ trivially follow from (A.5) and (A.6), respectively.

Next, we note that for $s_1 + 1 \le j \le p_1$, $\partial\|\hat{a}_j\|_{D_j} = D_j b_j$ for some b_j that satisfies $\|D_j^{1/2} b_j\| \le 1$ (because $\hat{a}_j = 0$). If we can show

$$\|Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{Z}\hat{a} + X\hat{\beta}, Y)\| = o_P(n\lambda_1\omega_{1j}), \qquad s_1 + 1 \le j \le p_1, \tag{A.9}$$

then (A.7) with $s_1 + 1 \le j \le p_1$ will follow. In fact, (A.9) implies that $Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{Z}\hat{a} + X\hat{\beta}, Y) = n\lambda_1\omega_{1j}d_j$ with $\|d_j\| = o_P(1)$. In addition, because the eigenvalues of D_j are of order $O(1)$, we have $d_j = D_j\left(D_j^{-1}d_j\right)$ with $D_j^{1/2}D_j^{-1}d_j = D_j^{-1/2}d_j = o_P(1)$ and thus $d_j \in \partial\|\hat{a}_j\|$ showing that the KKT condition is satisfied.

To show (A.9), we note that

$$\max_{s_1+1\le j\le p_1}\left\|Z_j^{\mathrm{T}}\mathbf{q}_1\left(\mathbf{Z}\hat{a}+X\hat{\beta},Y\right)\right\| \le \max_{s_1+1\le j\le p_1}\left\|Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{m},Y)\right\| + \max_{s_1+1\le j\le p_1}\left\|Z_j^{\mathrm{T}}\left[\mathbf{q}_2(\cdot,Y)\left(\mathbf{Z}\hat{a}+\mathbf{X}\hat{\beta}-\mathbf{m}\right)\right]\right\|,$$

where $\mathbf{m} = (m_1,\ldots,m_n)^{\mathrm{T}}$ with $m_i = \sum_j \alpha_{0j}\left(W_{ij}\right) + X_i^{\mathrm{T}}\beta_0$, $\mathbf{q}_2$ is evaluated somewhere between $\mathbf{m}$ and $\mathbf{Z}\hat{a}+X\hat{\beta}$, and the product of two n-vectors inside $[\cdot]$ in the preceding expression denotes componentwise product.

We prove that the first term in the preceding display is $o(n\lambda_1\omega_{1j})$. Note that for fixed $s_1+1\le j\le p_1$, $1\le k\le K$, we have

$$P\left\{\left|\sum_i B_k\left(W_{ij}\right)q_1(m_i,Y_i)\right| > c\right\} \le 2\exp\left\{-c^2/\left(2Jc+2nR^2\right)\right\}, \quad \forall c>0,$$

by Bernstein's inequality, Assumption (A5) and Lemma 5.7 in van der Geer (2000). Using a simple union bound, we have

$$\begin{aligned}P\left\{\max_{s_1+1\le j\le p_1}\left\|Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{m},Y)\right\| > c\right\} &\le p_1K\max_{j,k}P\left\{\left|\sum_i B_k(W_{ij})q_1\left(m_i,Y_i\right)\right| > c\right\}\\ &\le 2p_1K\exp\left\{-c^2/(2Jc+2nR^2)\right\}.\end{aligned}$$

Taking $c = C\sqrt{n\log(p_1K\vee n)}$ for some $C>0$ large enough, we get

$$P\left\{\max_{s_1+1\le j\le p_1}\|Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{m},Y)\| > c\right\} \to 0,$$

and thus

$$\max_{s_1+1\le j\le p_1}\|Z_j^{\mathrm{T}}\mathbf{q}_1(\mathbf{m},Y)\| = O_P\left(\sqrt{n\log(p_1K\vee n)}\right) = o\left(n\lambda_1\omega_{1j}\right), \tag{A.10}$$

by Assumption (A6).

Furthermore, using Proposition A.1, we have

$$\begin{aligned}&\max_{s_1+1\le j\le p_1}\left\|Z_j^{\mathrm{T}}\left[\mathbf{q}_2(\cdot,Y)\left(\mathbf{Z}\hat{a}+\mathbf{X}\hat{\beta}-\mathbf{m}\right)\right]\right\|\\ &\quad = \text{Ø}_P\left(\sqrt{n\left(K+s_2+\frac{n}{K^{2d}}+n\lambda_1^2\left\|\omega_1^{(1)}\right\|^2+n\lambda_2^2\left\|\omega_2^{(1)}\right\|^2\right)}\right),\end{aligned} \tag{A.11}$$

and thus Assumption (A6) implies that the preceding expression is $o(n\lambda_1\omega_{1j})$ for $s_1+1\le j\le p_1$.

Finally, for (A.8) with $s_2+1\le j\le p_2$, by a similar argument we only need to show

$$|X_j^{\mathrm{T}}\mathbf{q}_1\left(\mathbf{Z}\hat{a}+\mathbf{X}\hat{\beta},Y\right)| = o\left(n\lambda_2\omega_{2j}\right). \tag{A.12}$$

We omit the proof because it is almost the same as that for (A.9). ■

A.3. Proof of Theorem 4. The proof is split into two steps.

Step 1. Let $\tilde{a}^* = (\tilde{a}^{\mathrm{T}}, \tilde{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and $\delta = \tilde{a}^* - a_0^*$. We can write $\delta = (\delta_{a1}^{\mathrm{T}}, \dots, \delta_{ap_1}^{\mathrm{T}}, \delta_{\beta 1}, \dots, \delta_{\beta p_2})^{\mathrm{T}}$ with $\delta_{aj} = \tilde{a}_j - a_{0j} \in R^K$ and $\delta_{\beta j} = \tilde{\beta}_j - \beta_{0j} \in R$. We will show that

$$\sum_{j \in S_1^c} \|\delta_{aj}\|_{D_j} + \frac{\lambda_{02}}{\lambda_{01}} \sum_{j \in S_2^c} |\delta_{\beta j}| \le (1+\gamma) \left(\sum_{j \in S_1} \|\delta_{aj}\|_{D_j} + \frac{\lambda_{02}}{\lambda_{01}} \sum_{j \in S_2} |\delta_{\beta j}| \right), \tag{A.13}$$

with probability converging to 1, for any $\gamma > 0$.

By the definition of $\tilde{a}^*$, we have

$$\begin{aligned} \sum_i Q\left(U_i^{\mathrm{T}} \tilde{a}^*, Y_i\right) - \sum_i Q\left(U_i^{\mathrm{T}} a_0^*, Y_i\right) \le n\lambda_{01} \sum_{j=1}^{p_1} \|a_{0j}\|_{D_j} - n\lambda_{01} \sum_{j=1}^{p_1} \|\tilde{a}_j\|_{D_j} \\ + n\lambda_{02} \sum_{j=1}^{p_2} |\beta_{0j}| - n\lambda_{02} \sum_{j=1}^{p_2} |\tilde{\beta}_j|. \end{aligned} \tag{A.14}$$

On the other hand, by the convexity of Q,

$$\sum_i Q(U_i^{\mathrm{T}} \tilde{a}^*, Y_i) - \sum_i Q(U_i^{\mathrm{T}} a_0^*, Y_i) \ge \sum_i q_1(U_i^{\mathrm{T}} a_0^*, Y_i) U_i^{\mathrm{T}} \delta.$$

Combining the previous two displayed equations, we get

$$\begin{aligned} &- \left(\sum_{j=1}^{p_1} \|\delta_{aj}\| + \frac{\lambda_{02}}{\lambda_{01}} \sum_{j=1}^{p_2} |\delta_{\beta j}| \right) \\ &\quad \times \max\left\{ \max_{1 \le j \le p_1} \left\| \sum_i q_1\left(U_i^{\mathrm{T}} a_0, Y_i\right) Z_{ij} \right\|, \frac{\lambda_{01}}{\lambda_{02}} \max_{1 \le j \le p_2} \left| \sum_i q_1\left(U_i^{\mathrm{T}} a_0, Y_i\right) X_{ij} \right| \right\} \\ &\qquad \le n\lambda_{01} \sum_{j=1}^{p_1} \|a_{0j}\|_{D_j} - n\lambda_{01} \sum_{j=1}^{p_1} \|\tilde{a}_j\|_{D_j} + n\lambda_{02} \sum_{j=1}^{p_2} |\beta_{0j}| - n\lambda_{02} \sum_{j=1}^{p_2} |\tilde{\beta}_j|. \end{aligned} \tag{A.15}$$

Using the same arguments as in the proof of (A.10), we have

$$\max_{1 \le j \le p_1} \left\| \sum_i q_1\left(U_i^{\mathrm{T}} a_0, Y_i\right) Z_{ij} \right\| = O_P\left(\sqrt{n \log(p_1 K \vee n)}\right) = o_P\left(n\lambda_{01}\right), \tag{A.16}$$

$$\max_{1 \le j \le p_1} \left| \sum_i q_1\left(U_i^{\mathrm{T}} a_0, Y_i\right) X_{ij} \right| = O_P\left(\sqrt{n \log\left(p_2 \vee n\right)}\right) = o_P\left(n\lambda_{02}\right). \tag{A.17}$$

Using (A.15)–(A.17), together with

$$\begin{aligned} \lambda_{01} \left(\sum_{j=1}^{p_1} \|a_{0j}\|_{D_j} - \sum_{j=1}^{p_1} \|\tilde{a}_j\|_{D_j} \right) &= \lambda_{01} \left(\sum_{j \in S_1} \|a_{0j}\|_{D_j} - \sum_{j \in S_1} \|\tilde{a}_j\|_{D_j} - \sum_{j \in S_1^c} \|\tilde{a}_j\|_{D_j} \right) \\ &\le \lambda_{01} \left(\sum_{j \in S_1} \|\delta_{aj}\|_{D_j} - \sum_{j \in S_1^c} \|\delta_{aj}\|_{D_j} \right), \end{aligned} \tag{A.18}$$

$$\begin{aligned}\lambda_{02}\left(\sum_{j=1}^{p_2}|\beta_{0j}|-\sum_{j=1}^{p_2}|\tilde{\beta}_j|\right)&=\lambda_{02}\left(\sum_{j\in S_2}|\beta_{0j}|-\sum_{j\in S_2}|\tilde{\beta}_j|-\sum_{j\in S_2^c}|\tilde{\beta}_j|\right)\\&\le\lambda_{02}\left(\sum_{j\in S_2}|\delta_{\beta j}|-\sum_{j\in S_2^c}|\delta_{\beta j}|\right),\qquad\textbf{(A.19)}\end{aligned}$$

we obtain (A.13).

Step 2. We prove

$$\|\tilde{a}^*-a_0^*\|=O_P\left(\frac{\sqrt{s_1}\lambda_{01}+\sqrt{s_2}\lambda_{02}}{k_1}\right)\text{ with probability converging to 1.}$$

Combining (A.14), (A.18), and (A.19), we get

$$\begin{aligned}&\sum_i Q\left(U_i^{\mathrm{T}}\tilde{a}^*,Y_i\right)-\sum_i Q\left(U_i^{\mathrm{T}}a_0^*,Y_i\right)\\&\quad\le n\lambda_{01}\left(\sum_{j\in S_1}\|\delta_{aj}\|_{D_j}-\sum_{j\in S_1^c}\|\delta_{aj}\|_{D_j}\right)+n\lambda_{02}\left(\sum_{j\in S_2}|\delta_{\beta j}|-\sum_{j\in S_2^c}|\delta_{\beta j}|\right).\end{aligned}$$

Using Taylor expansion for the left-hand side in the previous expression results in

$$\begin{aligned}&\sum_i q_1\left(U_i^{\mathrm{T}}a_0^*,Y_i\right)U_i^{\mathrm{T}}\delta+\frac{1}{2}q_2\left(\cdot,Y_i\right)\delta^{\mathrm{T}}U_iU_i^{\mathrm{T}}\delta\\&\quad\le n\lambda_{01}\left(\sum_{j\in S_1}\|\delta_{aj}\|_{D_j}-\sum_{j\in S_1^c}\|\delta_{aj}\|_{D_j}\right)+n\lambda_{02}\left(\sum_{j\in S_2}|\delta_{\beta j}|-\sum_{j\in S_2^c}|\delta_{\beta j}|\right),\qquad\textbf{(A.20)}\end{aligned}$$

where $q_2(\cdot,\cdot)$ is evaluated at some point that lies between $U_i^{\mathrm{T}}\tilde{a}^*$ and $U_i^{\mathrm{T}}a_0^*$.

As shown in step 1, $|\sum_i q_1(U_i^{\mathrm{T}}a_0,Y_i)U_i^{\mathrm{T}}\delta|=o_P(n\lambda_{01}\sum_j\|\delta_{aj}\|_{D_j}+n\lambda_{02}\sum_j|\delta_{\beta j}|)$. Assumption (A7) implies $\sum_i q_2(z_i^*,Y_i)\delta^{\mathrm{T}}U_iU_i^{\mathrm{T}}\delta\ge nk_1\|\delta\|^2$, and thus equation (A.20) becomes

$$\|\delta\|^2\le\left(1+o_P(1)\right)\frac{\lambda_{01}\sum_j\|\delta_{aj}\|_{D_j}+\lambda_{02}\sum_j|\delta_{\beta j}|}{k_1}.$$

Using Cauchy–Schwarz inequality we have $\lambda_{01}\sum_j\|\delta_{aj}\|_{D_j}+\lambda_{02}\sum_j|\delta_{\beta j}|=O((\sqrt{s_1}\lambda_{01}+\sqrt{s_2}\lambda_{02})\|\delta\|)$ and the convergence rate is obtained. ■

***A.4. Discussions of Assumption* (A7).** We will first discuss how the restricted eigenvalued condition (A7) can be implied by the sparse Riesz condition and then consider how the sparse Riesz condition can be satisfied in the semiparametric case. We do not aim to conduct a comprehensive study on these eigenvalue assumptions or provide very general sufficient conditions for (A7). The main goal is just to show that (A7) can be satisfied in some situations, and we make some further simplifying assumptions to facilitate this discussion.

We will relate (A7) to sparse Riesz condition as in Bickel, Ritov, and Chernozhukov (2009). Let $A\subseteq\{1,\dots,p\}$ and denote by U_{iA} the subvector of U_i containing only components associated with predictors in A. Define $c^*(m)=\sup_{|A|\le m,\|v\|=1}\sum_i v^TU_{iA}U_{iA}^Tv/n$

and $c_*(m) = \inf_{|A|\le m, \|v\|=1} \sum_i v^T U_{iA} U_{iA}^T v/n$. Conditions on the magnitudes of $c^*(m)$ and $c_*(m)$ are usually referred to as sparse Riesz conditions.

The following discussion is mainly adapted from Bickel et al. (2009), in particular their proof of Lemma 4.1, and we only focus on the modifications required. The paper Bickel et al. (2009) contains other sufficient conditions for restricted eigenvalue assumption, but here we only focus on part (ii) of their Lemma 4.1.

To ease notation, we first assume we take $\lambda_{01} = \lambda_{02}$ in the LASSO estimator. This is a theoretically plausible constraint such as we use in the discussion following Theorem 4. In this discussion only, we suppose the covariates are rearranged such that $S = \{1, \ldots, s = s_1 + s_2\}$ are the indexes for all the nonzero components and let $S^c = \{s+1, \ldots, p = p_1 + p_2\}$. Also, for the $(p_1 K + p_2)$-dimensional vector v as defined in (A7), we write $v = (v_1, \ldots, v_p)$ where each v_j is either K-dimensional or a scalar (i.e., we do not distinguish v_j associated with the nonparametric and the parametric parts in notation) and the constraint is now written as

$$\sum_{j\in S^c} \|v_j\| \le (1+\gamma) \sum_{j\in S} \|v_j\|,$$

where $\|v_j\|$ actually denotes $\|v_j\|_{D_j}$ when v_j is a vector or $|v_j|$ when v_j is a scalar. Also we let $\|v\| = \sqrt{\sum_{j=1}^p \|v_j\|^2}$, and similarly $\|v_S\| = \sqrt{\sum_{j\in S} \|v_j\|^2}$, for example. Note that $\|v\|$ defined here has the same order as the usual euclidean norm of v. The purpose of using this notation is that now it is clearer that everything becomes similar to the parametric case.

More specifically, given v satisfying the preceding displayed constraint, we partition S^c into subsets of size m with last subset of size $\le m$ (we will set $m = s\log n - s$ later). Thus we write $S^c = \cup_{h=1}^H S_h$ where S_h contains the indexes j corresponding to m largest $\|v_j\|$ outside of $\cup_{k=1}^{h-1} S_k$. Let $S_{01} = S \cup S_1$. Using the same arguments as in the proof of Lemma 4.1 of Bickel et al. (2009), we get $\left(\sum_i v^T U_i U_i^T v\right)/n \ge \left(\sqrt{c_*(s+m)} - (1+\gamma)\sqrt{c^*(m)}\sqrt{s/m}\right) \|v_{S_{01}}\|^2$. Furthermore, because the kth largest value among $\|v_j\|, s+1 \le j \le p$ satisfies $\|v_j\| \le \sum_{s+1\le j\le p} \|v_j\|/k$, we have $\|v_{S_{01}^c}\|^2 \le \left(\sum_{s+1\le j\le p} \|v_j\|\right)^2 \sum_{k\ge m+1} (1/k^2) \le \left(\sum_{s+1\le j\le p} \|v_j\|\right)^2/m$, and thus

$$\begin{aligned}\|v\| &\le \|v_{S_{01}}\| + \|v_{S_{01}^c}\| \le \|v_{S_{01}}\| + \frac{\sum_{s+1\le j\le p} \|v_j\|}{\sqrt{m}}\\ &\le \|v_{S_{01}}\| + (1+\gamma)\frac{\sum_{1\le j\le s} \|v_j\|}{\sqrt{m}} \le \|v_{S_{01}}\| + (1+\gamma)\sqrt{s/m}\|v_S\|\\ &\le \left(1 + (1+\gamma)\sqrt{s/m}\right) \|v_{S_{01}}\|.\end{aligned}$$

Then we have

$$\frac{\sum_i v^T U_i U_i^T v}{(n)\|v\|^2} \ge C\left(1 + (1+\gamma)\sqrt{s/m}\right)^{-2} \left(\sqrt{c_*(s+m)} - (1+\gamma)\sqrt{c^*(m)}\sqrt{s/m}\right).$$

As a result, the preceding expression is bounded away from zero if $s + m = s\log n$ and $c_*(s\log n)$ is bounded away from zero, and $c^*(m)$ is bounded.

We now consider how $c^*(m)$ and $c_*(m)$ can be bounded and bounded away from zero for suitable m. For simplicity, we assume that only the nonparametric components are present because the parametric case has been well studied in the literature (Belloni and Chernozhukov, 2012; Zhang and Huang, 2008).

The lemma that follows gives sufficient conditions for $c^*(m)$ and $c_*(m)$ to be bounded and bounded away from zero. Because we are concerned only with the nonparametric components, we write p in place of p_1 to simplify notation.

LEMMA A.1. *For any vector* $v=(v_1^{\mathrm{T}},\ldots,v_p^{\mathrm{T}})^{\mathrm{T}}$, *denote* $g_j(x)=\sum_k v_{jk}B_{jk}(x)$. *Suppose* $E\left[\left(\sum_{j\in A} g_j(W_j)\right)^2\right]/\sum_{j\in A}\|g_j\|^2$ *is bounded and bounded away from zero uniformly over A with* $|A|\le m$. *If* $Km\log K/n\to 0$ *and* $Km\log p/n\to 0$, *then*

$$C_*\|v\|^2\le \frac{v^T\sum_i Z_{iA}Z_{iA}^Tv}{n}\le C^*\|v\|^2,\quad \forall |A|\le m,$$

for some constants $C_*,C^*>0$ *with probability tending to 1.*

Proof. $v^T\sum_i Z_{iA}Z_{iA}^Tv$ can be written as

$$\sum_i\sum_{j,j'\in A,k,k'\in\{1,\ldots,K\}} v_{jk}B_{jk}(W_{ij})B_{j'k'}(W_{ij'})v_{j'k'}=\sum_i\sum_{j\in A,j'\in A} g_j(W_{ij})g_{j'}(W_{ij'}). \tag{A.21}$$

Using the property of splines, $\int g_j^2(t)dt\asymp\|v_j\|^2$. Let $g=(g_1,\ldots,g_p)$ and let $g_A=(g_j, j\in A)$ be the subvector of g. Defining $\|g_A\|_n^2=(1/n)\sum_i\left(\sum_{j\in A}g_j(W_{ij})\right)^2$ and the population version $\|g_A\|_W^2=E\left[\left(\sum_{j\in A}g_j(W_j)\right)^2\right]$ we will show that

$$P\left(\sup_{g_j\in\mathcal{S},|A|\le m}\left|\frac{\|g_A\|_n^2-\|g_A\|_W^2}{\|g_A\|_W^2}\right|>\epsilon\right)\le C_1p^2K^2\exp\left\{C_2\frac{n}{K}\frac{(\epsilon/m)^2}{1+(\epsilon/m)}\right\}, \tag{A.22}$$

where $\mathcal{S}$ is the set of spline functions. The preceding result can be shown following similar arguments as in Lemma A.2 of Huang, Wu, and Zhou (2004) with slight modifications due to the diverging dimensionality. First, as shown there we have, for any $1\le j,j'\le p, 1\le k,k'\le K$,

$$\begin{aligned}P&\left(|\sum_i\left(B_{jk}(W_{ij})B_{j'k'}(W_{ij'})\right)/n-E\left[B_{jk}(W_j)B_{j'k'}(W_{j'})\right]|>\epsilon\right)\\&\le C_1exp\left\{-\frac{(n\epsilon/K)^2}{C_2(n/K)+C_3n\epsilon/K}\right\}.\end{aligned}$$

Thus by a simple union bound, there is an event Ω_n with $P(\Omega_n)\ge 1-C_4p^2K^2\exp\left\{-C_5n/K(\epsilon/m)^2/(1+(\epsilon/m))\right\}$ such that on Ω_n, $|\sum_i\left(B_{jk}(W_{ij})B_{j'k'}W_{ij'}\right)/n-E\left[B_{jk}(W_j)B_{j'k'}(W_{j'})\right]|\le\epsilon/m$ for all $1\le j,j'\le p, 1\le k,k'\le K$. Then similar again to Lemma A.2 in Huang et al. (2004), we have that on Ω_n, $|\|g_A\|_n-\|g_A\|_W|\le C\epsilon\|g_A\|_W^2$, using that $|A(j,k)|\le Cm$ where $A(j,k)$ is the set of indexes (j',k') such that $j'\in A$ and supports of B_{jk} and $B_{j'k'}$ overlap. Thus, based on (A.22), if $Km\log K/n\to 0$ and $Km\log p/n\to 0$, we have $\sup_{v,|A|\le m}|\|g_A\|_n^2/\|g_A\|_W^2-1|=o_P(1)$.

Applying this result to the right-hand side of (A.21), using the assumption in the statement of the lemma, we know that $v^T \sum_i Z_{iA} Z_{iA} v / \big(n\|v\|^2\big)$ is bounded and bounded away from zero, uniformly in the choice of $|A|$ with $|A| \leq m$. ■

We now comment whether the assumption that $E[(\sum_{j\in A} g_j(W_j))^2]/\sum_{j\in A}\|g_j\|^2$ is bounded and bounded away from zero can be satisfied. This is similar to our discussion of (A2). If m is fixed, then this assumption is naturally true under mild assumptions. In particular if the covariates W_j are independent of each other, we obviously have $E[(\sum_{j\in A} g_j(W_j))^2]/\sum_{j\in A}\|g_j\|^2$ bounded and bounded away from zero as long as the densities of W_j are bounded and bounded away from zero. If on the other hand $m \to \infty$, whether this assumption holds true roughly depends on the correlation between different covariates, but it is still natural to assume it holds. If $E[(\sum_{j\in A} g_j(W_j))^2]/\sum_{j\in A}\|g_j\|^2$ converges to zero at a certain rate, this rate will also be reflected in the convergence rate of Theorem 4. As discussed in the text for (A2), the theory can be extended accordingly but it is generally hard to know at what rate this quantity shrinks to zero.