VARIABLE SELECTION IN NONPARAMETRIC ADDITIVE MODELS

Jian Huang[1], Joel L. Horowitz[2] and Fengrong Wei[1]
[1]University of Iowa, [2]Northwestern University

The University of Iowa
Department of Statistics and Actuarial Science
Technical Report No. 394
November 24, 2008

Summary. We consider a nonparametric additive model of a conditional mean function in which the number of variables and additive components may be larger than the sample size but the number of non-zero additive components is "small" relative to the sample size. The statistical problem is to determine which additive components are non-zero. The additive components are approximated by truncated series expansions with B-spline bases. With this approximation, the problem of component selection becomes that of selecting the groups of coefficients in the expansion. We apply the adaptive group Lasso to select nonzero components, using the group Lasso to obtain an initial estimator and reduce the dimension of the problem. We give conditions under which the group Lasso selects a model whose number of components is comparable with the underlying model and, the adaptive group Lasso selects the non-zero components correctly with probability approaching one as the sample size increases and achieves the optimal rate of convergence. Following model selection, oracle-efficient, asymptotically normal estimators of the non-zero components can be obtained by using existing methods. The results of Monte Carlo experiments show that the adaptive group Lasso procedure works well with samples of moderate size. A data example is used to illustrate the application of the proposed method.

Key words and phrases. Adaptive group Lasso; component selection; high-dimensional data; nonparametric regression; selection consistency.

Short title. Nonparametric component selection

AMS 2000 subject classification. Primary 62G08, 62G20; secondary 62G99

1 Introduction

Let $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$ be random vectors that are independently and identically distributed as $(Y, \mathbf{X})$, where Y is a response variable and $\mathbf{X} = (X_1, \ldots, X_p)'$ is a p-dimensional covariate vector. Consider the nonparametric additive model

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \tag{1}$$

where μ is an intercept term, X_{ij} is the jth component of X_i, the f_j's are unknown functions, and ε_i is an unobserved random variable with mean zero and finite variance σ^2. Suppose that some of the additive components f_j are zero. The problem addressed in this paper is to distinguish the nonzero components from the zero components and estimate the nonzero components. We allow the possibility that p is larger than the sample size n, which we represent by letting p increase as n increases. We propose a penalized method for variable selection in (1) and show that the proposed method can correctly select the nonzero components with high probability.

There has been much work on penalized methods for variable selection and estimation with high-dimensional data. Methods that have been proposed include the bridge estimator (Frank and Friedman 1993; Huang, Ma and Horowitz 2008), least absolute shrinkage and selection operator (Lasso, Tibshirani 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001; Fan and Peng 2004), the elastic net penalty (Zou and Hastie 2006), and the minimum concave penalty (Zhang 2007). Much progress has been made in understanding the statistical properties of these methods in both fixed p and $p \gg n$ settings. In particular, many authors have studied the variable selection, estimation and prediction properties of the Lasso in both low- and high-dimensional settings. See for example, Knight and Fu (2001); Greenshtein and Ritov (2004); Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Zou (2006); Bunea, Tsybakov and Wegkamp (2007); Meinshausen and Yu (2008); Huang, Ma and Zhang (2008); van de Geer (2008) and Zhang and Huang (2008), among others. All these authors assume a linear or other parametric model. In many applications, however, there is little a priori justification for assuming that the effects of covariates take a linear form or belong to any other known, finite-dimensional parametric family. For example, in studies investigating the relationship between a phenotype and genomic measurements, it is necessary to take account of environmental covariates whose effects on the outcome variable can be nonlinear. In studies of economic development, the effects of covariates on the growth of gross domestic product can be nonlinear.

In the context of smoothing spline ANOVA, Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) method for model selection and estimation in multivariate nonparametric regression models. For fixed p, they showed that the COSSO estimator in the additive model converges at the rate $n^{-d/(2d+1)}$, where d is the order of smoothness of the components. They also showed that, in a special case of tensor product design, the COSSO correctly selects the components with high probability. Zhang and Lin (2006) considered the COSSO for nonparametric regression in exponential families.

There is a large body of literature on estimation in nonparametric additive models. For example, Stone (1985, 1986) showed that the additive spline estimators achieve optimal rate of convergence for general and fixed p as for $p = 1$. Horowitz and Mammen (2004) and Horowitz, Klemelae, and Mammen (2006) showed that if p is fixed and mild regularity conditions hold, then oracle-efficient estimates of the f_j's can be obtained by a two-step procedure. Here, oracle efficiency means that the estimator of each f_j has the same asymptotic distribution that it would have if all the other f_j's were known. However, these papers did not discuss variable selection in nonparametric additive models.

In this paper, we propose to use the group Lasso method for variable selection in (1) based on a spline approximation to the nonparametric components. With this approximation, each nonparametric component is represented by a linear combination of spline basis functions. Consequently, the problem of component selection becomes that of selecting the groups of coefficients in the linear combinations. It is natural to apply the group Lasso method, since it is desirable to take into the grouping structure in the approximation model. To achieve selection consistency, we apply the group Lasso iteratively as follows. First, we use the group Lasso to obtain an initial estimator and reduce the dimension of the problem. Then we use the adaptive group Lasso to select the final set of groups of variables. This approach follows the idea of the adaptive Lasso (Zou 2006) and a proposal by Bühlmann and Meier (2008) in the context of variable selection in linear regression. They considered a combination of Lasso and adaptive Lasso steps, and more generally, a multi-step adaptive Lasso procedure.

We show that the group Lasso selects a model whose number of components is comparable with the underlying model. Then, using the group Lasso result as the initial estimator, the adaptive group Lasso selects the correct model with high probability and achieves the optimal rate of convergence. Following model selection, oracle efficient, asymptotically normal estimates of the non-zero additive components can be obtained by, for example, using the methods of Mammen, Linton, and Nielsen (1999) or Horowitz and Mammen (2004). An important aspect of our results is that p can be much larger than n.

The remainder of the paper is organized as follows. Section 2 describes the group Lasso and the adaptive group Lasso for variable selection in nonparametric additive models. Section 3 presents the asymptotic properties of these methods in "large p, small n" settings. Section 4 presents the results of simulation studies to evaluate the finite-sample performance of these methods. Section 5 provides an illustrative application, and Section 6 presents concluding remarks. Proofs of the results stated in Section 3 are given in Section 7.

2 The adaptive group Lasso in nonparametric additive models

We describe a two-step approach to using the group Lasso for variable selection based on a spline representation of each component in additive models. In the first step, we use the standard group Lasso to achieve an initial reduction of the dimension in the model and obtain an initial estimator of the nonparametric components. In the second step, we use the adaptive group Lasso to achieve consistent selection.

Suppose that each X_j takes values in $[a, b]$ where $a < b$ are finite numbers. To ensure unique identification of the f_j's, we assume that $\mathrm{E}f_j(X_j) = 0, 1 \leq j \leq p$. Let $a = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = b$ be a partition of $[a, b]$ into K subintervals $I_{Kt} = [\xi_t, \xi_{t+1}), t = 0, \dots, K-1$ and $I_{KK} = [\xi_K, \xi_{K+1}]$, where $K \equiv K_n = n^v$ with $0 < v < 0.5$ is a positive integer such that $\max_{1\leq k\leq K+1} |\xi_k - \xi_{k-1}| = O(n^{-v})$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ consisting of functions s satisfying: (i) the restriction of s to I_{Kt} is a polynomial of degree l for $1 \leq t \leq K$; (ii) for $l \geq 2$ and $0 \leq l' \leq l - 2$, s is l' times continuously differentiable on $[a, b]$. This definition is phrased after Stone (1985), which is a descriptive version of Schumaker (1981), page 108, Definition 4.1.

There exists a normalized B-spline basis $\{\phi_k, 1 \leq k \leq m_n\}$ for $\mathcal{S}_n$, where $m_n \equiv K_n + l$ (Schumaker 1981). Thus for any $f_{nj} \in \mathcal{S}_n$, we can write

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_k(x), \quad 1 \leq j \leq p.$$

Under suitable smoothness assumptions, the f_j's can be well approximated by functions in $\mathcal{S}_n$. Accordingly, the variable selection method described in this paper is based on the representation (2).

Let $\|\mathbf{a}\|_2 \equiv \left(\sum_{j=1}^m |a_j|^2\right)^{1/2}$ denote the ℓ_2 norm of any vector $\mathbf{a} \in \mathbb{R}^m$. Let $\boldsymbol{\beta}_{nj} = (\beta_{j1}, \dots, \beta_{jm_n})'$ and $\boldsymbol{\beta}_n = (\boldsymbol{\beta}'_{n1}, \dots, \boldsymbol{\beta}'_{np})'$. Let $w_n = (w_{n1}, \dots, w_{np})'$ be a given vector of weights, where $0 \leq w_{nj} \leq \infty, 1 \leq j \leq p$. Consider the penalized least squares criterion

$$L_n(\mu, \boldsymbol{\beta}_n) \;=\; \sum_{i=1}^n \Big[Y_i - \mu - \sum_{j=1}^p \sum_{k=1}^{m_n} \beta_{jk}\phi_k(X_{ij})\Big]^2 + \lambda_n \sum_{j=1}^p w_{nj}\|\boldsymbol{\beta}_{nj}\|_2, \tag{2}$$

where λ_n is a penalty parameter. We study the estimators that minimize $L_n(\mu, \boldsymbol{\beta}_n)$ subject to the constraints

$$\sum_{i=1}^n \sum_{k=1}^{m_n} \beta_{jk}\phi_k(X_{ij}) = 0, 1 \leq j \leq p. \tag{3}$$

These centering constraints are sample analogs of the identifying restriction $Ef_j(X_j) = 0, 1 \leq j \leq p$. We can convert (2)-(3) to an unconstrained optimization problem by centering the response and the basis functions. Let

$$\bar{\phi}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_k(X_{ij}), \quad \psi_{jk}(x) = \phi_k(x) - \bar{\phi}_{jk}. \tag{4}$$

For simplicity and without causing confusion, we simply write $\psi_k(x) = \psi_{jk}(x)$. Denote

$$Z_{ij} = \big(\psi_1(X_{ij}), \ldots, \psi_{m_n}(X_{ij})\big)'.$$

So Z_{ij} consists of values of the (centered) basis functions at the ith observation of the jth covariate. Let $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{nj})'$ be the $n \times m_n$ 'design' matrix corresponding to the jth covariate. The total 'design' matrix is $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$. Let $\mathbf{Y} = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})'$. With this notation, we can write

$$L_n(\boldsymbol{\beta}_n; \lambda) = \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_n \sum_{j=1}^{p} w_{nj} \|\boldsymbol{\beta}_{nj}\|_2. \tag{5}$$

Here we have dropped μ in the argument of L_n. With the centering, $\widehat{\mu} = \overline{Y}$. Then minimizing (2) subject to (3) is equivalent to minimizing (5) with respect to $\boldsymbol{\beta}_n$, but the centering constraints are not needed for (5).

We now describe the two-step approach to component selection in the nonparametric additive model (1).

Step 1. Compute the group Lasso estimator. Let

$$L_{n1}(\boldsymbol{\beta}_n, \lambda_{n1}) = \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_{n1} \sum_{j=1}^{p} \|\boldsymbol{\beta}_{nj}\|_2.$$

This objective function is the special case of (5) that is obtained by setting $w_{nj} = 1, 1 \leq j \leq p$. The group Lasso estimator is $\widetilde{\boldsymbol{\beta}}_n \equiv \widetilde{\boldsymbol{\beta}}_n(\lambda_{n1}) = \arg\min_{\boldsymbol{\beta}_n} L_{n1}(\boldsymbol{\beta}_n; \lambda_{n1})$.

Step 2. Use the group Lasso estimator $\widetilde{\boldsymbol{\beta}}_n$ to obtain the weights by setting

$$w_{nj} = \begin{cases} \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2^{-1}, & \text{if } \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 > 0, \\ \infty, & \text{if } \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 = 0. \end{cases}$$

The adaptive group Lasso objective function is

$$L_{n2}(\boldsymbol{\beta}_n; \lambda_{n2}) = \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_{n2} \sum_{j=1}^{p} w_{nj} \|\boldsymbol{\beta}_{nj}\|_2.$$

Here we define $0 \cdot \infty = 0$. Thus the components not selected by the group Lasso are not included in Step 2. The adaptive group Lasso estimator is $\widehat{\boldsymbol{\beta}}_n \equiv \widehat{\boldsymbol{\beta}}_n(\lambda_{n2}) = \arg\min_{\boldsymbol{\beta}_n} L_{n2}(\boldsymbol{\beta}_n; \lambda_{n2})$.

Finally, the adaptive group Lasso estimators of μ and f_j are

$$\widehat{\mu}_n = \overline{Y} \equiv n^{-1} \sum_{i=1}^{n} Y_i, \quad \widehat{f}_{nj}(x) = \sum_{k=1}^{m_n} \widehat{\beta}_{jk} \psi_k(x), \ 1 \leq j \leq p.$$

3 Main results

This section presents our results on the asymptotic properties of the estimators defined in Steps 1 and 2 of Section 2.

Let k be a non-negative integer, and let $\alpha \in (0, 1]$ be such that $d = k + \alpha > 0.5$. Let $\mathcal{F}$ be the class of functions f on $[0, 1]$ whose kth derivative $f^{(k)}$ exists and satisfies a Lipschitz condition of order α:

$$|f^{(k)}(s) - f^{(k)}(t)| \leq C|s - t|^{\alpha} \quad \text{for } s, t \in [a, b].$$

In (1), without loss of generality, suppose that the first q components are nonzero, that is, $f_j(x) \neq 0, 1 \leq j \leq q$, but $f_j(x) \equiv 0, q + 1 \leq j \leq p$. Let $A_1 = \{1, \ldots, q\}$ and $A_0 = \{q + 1, \ldots, p\}$. Define $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$ for any function f, whenever the integral exists.

Consider the following conditions.

(A1) The number of nonzero components q is fixed and there is a constant $c_f > 0$ such that $\min_{1 \leq j \leq q} \|f_j\|_2 \geq c_f$.

(A2) The random variables $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Furthermore, their tail probabilities satisfy $P(|\varepsilon_i| > x) \leq K \exp(-Cx^2), i = 1, \ldots, n$, for all $x \geq 0$ and for constants C and K.

(A3) $\mathrm{E}f_j(X_j) = 0$ and $f_j \in \mathcal{F}, j = 1, \ldots, q$.

(A4) There exist constants C_1 and C_2 such that the density function g_j of X_j satisfies $0 < C_1 \leq g_j(x) \leq C_2 < \infty$ on $[a, \ b]$ for every $1 \leq j \leq p$.

Let $|A|$ denote the cardinality of any set $A \subset \{1, \ldots, p\}$. Denote

$$\boldsymbol{\beta}_A = (\boldsymbol{\beta}_j', j \in A)' \quad \text{and} \quad \mathbf{Z}_A = (\mathbf{Z}_j, j \in A).$$

Here $\boldsymbol{\beta}_A$ is an $|A|m_n \times 1$ vector and $\mathbf{Z}_A$ is an $n \times |A|m_n$ matrix. Let $\mathbf{C}_A = \mathbf{Z}_A'\mathbf{Z}_A/n$. When $A = \{1, \ldots, p\}$, we simply write $\mathbf{C} = \mathbf{Z}'\mathbf{Z}/n$. Define

$$\rho_{\min}(m) = \min_{|A|=m} \min_{\|\mathbf{u}\|=1} \mathbf{u}'\mathbf{C}_A\mathbf{u}, \ \rho_{\max}(m) = \max_{|A|=m} \max_{\|\mathbf{u}\|=1} \mathbf{u}'\mathbf{C}_A\mathbf{u}.$$

So if we take $A = \{1, \ldots, p\}$, then $\rho_{\min}(p)$ and $\rho_{\max}(p)$ are the smallest and largest eigenvalues of $\mathbf{C}$, respectively. Note that if $n < pm_n$, then $\rho_{\min}(p) = 0$.

3.1 Estimation consistency of the group Lasso

Let q^* be a fixed integer such that, for $\rho_{n*} = \rho_{\min}(q^*)$, $\rho_n^* = \rho_{\max}(q^*)$ and

$$\bar{\rho} = \rho_n^*/\rho_{n*} \quad \text{and} \quad M_1 = 2 + 4\bar{\rho}, \tag{6}$$

we have

$$(2 + 4\bar{\rho})q + 1 \le q^*. \tag{7}$$

Below, for any two sequences $\{a_n, b_n, n = 1, 2, \ldots,\}$, we write $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 \le a_n/b_n \le c_2$ for all n sufficiently large, and write $a_n \asymp_p b_n$ if this inequality holds with probability converging to one.

By Lemma 3 in Section 6, we have $\rho_{n*} \asymp_p m_n^{-1}$, $\rho_n^* \asymp_p m_n^{-1}$ and $\bar{\rho} \asymp_p 1$. Define

$$\lambda_{n,p} = 2\sigma\sqrt{8(1 + c_0)m_n q^* \bar{\rho}\rho_n^* n \log(pm_n)},$$

where $c_0 \ge 0$. Note that for fixed q^*, $\lambda_{n,p} \asymp_p \sqrt{n\log(pm_n)}$. Let $A_1 = \{j : \|f_j\|_2 \ne 0, 1 \le j \le p\} = \{j : \|\boldsymbol{\beta}_{nj}\|_2 \ne 0, 1 \le j \le p\} = \{1, \ldots, q\}$ and $\widetilde{A}_1 = \{j : \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \ne 0, 1 \le j \le p\}$.

Theorem 1 *Suppose that (A1) to (A4) and (7) hold and that $\lambda_{n1} \ge \lambda_{n,p}$.*

(i) With probability converging to 1, $|\widetilde{A}_1| \le M_1|A_1| = M_1 q$ for M_1 defined in (6).

(ii) If $m_n^2 \log(pm_n)/n \to 0$ and $(\lambda_{n1}^2 m_n)/n^2 \to 0$ as $n \to \infty$, then all the nonzero $\boldsymbol{\beta}_{nj}, 1 \le j \le q$, are selected with probability converging to one.

(iii)

$$\sum_{j=1}^{p} \|\widetilde{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2 = O_p\Big(\frac{m_n^2 \log(pm_n)}{n}\Big) + O_p\Big(\frac{m_n}{n}\Big) + O\Big(\frac{1}{m_n^{2d-1}}\Big) + O\left(\frac{4m_n^2\lambda_{n1}^2}{n^2}\right).$$

Part (i) of Theorem 1 says that the group Lasso selects a model whose dimension is a constant multiple of the number of non-zero additive components f_j, regardless of the number of additive components that are zero.. Part (ii) implies that every nonzero coefficient will be selected with high probability. Part (iii) shows that the difference between the coefficients in the spline representation of the nonparametric functions in (1) and their estimators converges to zero in probability. The rate of convergence is determined by four terms: the stochastic error in estimating the nonparametric components (the first term) and the intercept μ (the second term), the spline approximation error (the third term) and the bias due to penalization (the fourth term).

Let $\widetilde{f}_{nj}(x) = \sum_{j=1}^{m_n} \widetilde{\beta}_{jk}\psi(x), 1 \le j \le p$. The following theorem is a consequence of Theorem 1.

Theorem 2 *Suppose that (A1) to (A4) hold and that $\lambda_{n1} \ge \lambda_{n,p}$. Then,* *(i) Let $\widetilde{A}_f = \{j : \|\widetilde{f}_{nj}\|_2 > 0, 1 \leq j \leq p\}$. There is a constant $M_1 > 0$ such that, with probability converging to 1, $|\widetilde{A}_f| \leq M_1 q$.*

(ii) If $(m_n \log(pm_n))/n \to 0$ and $(\lambda_{n1}^2 m_n)/n^2 \to 0$ as $n \to \infty$, then all the nonzero additive components $f_j, 1 \leq j \leq q$, are selected with probability converging to one.

(iii)

$$\|\widetilde{f}_{nj} - f_j\|_2^2 = O_p\Big(\frac{m_n \log(pm_n)}{n}\Big) + O_p\left(\frac{1}{n}\right) + O\Big(\frac{1}{m_n^{2d}}\Big) + O\left(\frac{4m_n\lambda_{n1}^2}{n^2}\right), \; j \in \widetilde{A}_2,$$

where $\widetilde{A}_2 = A_1 \cup \widetilde{A}_1$.

Thus under the conditions of Theorem 2, the group Lasso selects all the nonzero additive components with high probability. Part (iii) of the theorem gives the rate of convergence of the group Lasso estimator of the nonparametric components. Below, for any two sequences $\{a_n, b_n, n = 1, 2, \ldots, \}$, we write $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 \leq a_n/b_n \leq c_2$ for all n sufficiently large, and write $a_n \asymp_p b_n$ if this inequality holds with probability converging to one.

We now state a useful corollary of Theorem 2.

Corollary 1 *Suppose that (A1) to (A4) hold. If $\lambda_{n1} \asymp \sqrt{n\log(pm_n)}$ and $m_n \asymp n^{1/(2d+1)}$. Then,*

(i) If $n^{-2d/(2d+1)}\log(p) \to 0$ as $n \to \infty$, then with probability converging to one, all the nonzero components $f_j, 1 \leq j \leq q$, are selected and the number of selected components is no more than $M_1 q$.

(ii)

$$\|\widetilde{f}_{nj} - f_j\|_2^2 = O_p(n^{-2d/(2d+1)}\log(pm_n)), \; j \in \widetilde{A}_2.$$

For the λ_{n1} and m_n given in Corollary 1, the number of nonzero components can be as large as $p_n = \exp(o(n^{2d/(2d+1)}))$. For example, if each f_j has continuous second derivative ($d = 2$), then $p_n = \exp(o(n^{4/5}))$, which can be much larger than n.

3.2 Selection consistency of the adaptive group Lasso

We now consider the properties of the adaptive group Lasso. We first state a general result concerning the selection consistency of the adaptive group Lasso, assuming an initial consistent estimator is available. We then apply to the case when the group Lasso is used as the initial estimator.

Denote $\boldsymbol{\beta}_{nA_1} = (\boldsymbol{\beta}'_{nj}, j \in A_1)'$, $\mathbf{Z}_{A_1} = (\mathbf{Z}_j, j \in A_1)$ and $\mathbf{C}_{A_1} = n^{-1}\mathbf{Z}'_{A_1}\mathbf{Z}_{A_1}$. Let ρ_{n1} and ρ_{n2} be the smallest and largest eigenvalues of $\mathbf{C}_{A_1}$, respectively. Denote $b_{n1} = \min\{\|\boldsymbol{\beta}_{nj}\|_2 : j \in A_1\}$. Consider the following conditions.

(B1) The initial estimators $\widetilde{\boldsymbol{\beta}}_{nj}$ are r_n-consistent at zero:

$$r_n \max_{j\in A_0} \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 = O_P(1), \quad r_n \to \infty,$$

and there exists a constant $c_b > 0$ such that

$$\mathrm{P}(\min_{j\in A_1} \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \geq c_b b_{n1}) \to 1,$$

where $b_{n1} = \min_{j\in A_1} \|\boldsymbol{\beta}_{nj}\|_2$.
(B2) Let q be the number of nonzero components and $s_n = p - q$ be the number of zero components. Suppose that

$$\begin{array}{lrcl} (a) & \dfrac{m_n}{n^{1/2}} + \dfrac{\lambda_{n2} m_n^{1/4}}{n} & = & o(1), \\ (b) & \dfrac{n^{1/2}\big\{\log(s_n m_n)\big\}^{1/2}}{\lambda_{n2} r_n} + \dfrac{n}{\lambda_{n2} r_n m_n^{(2d+1)/2}} & = & o(1). \end{array}$$

For $\widehat{\boldsymbol{\beta}}_n \equiv (\widehat{\boldsymbol{\beta}}'_{n1}, \ldots, \widehat{\boldsymbol{\beta}}'_{np})'$ and $\boldsymbol{\beta}_n \equiv (\boldsymbol{\beta}'_{n1}, \ldots, \boldsymbol{\beta}'_{np})'$, we say $\widehat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n$ if $\mathrm{sgn}_0(\|\widehat{\boldsymbol{\beta}}_{nj}\|) = \mathrm{sgn}_0(\|\boldsymbol{\beta}_{nj}\|), 1 \leq j \leq p_n$, where $\mathrm{sgn}_0(|x|) = 1$ if $|x| > 0$ and $= 0$ if $|x| = 0$.

Theorem 3 *Suppose that conditions (B1), (B2) and (A2)-(A4) hold. Then*
(i)

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n) \to 1.$$

(ii)

$$\sum_{j=1}^{q} \|\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2 = O_p\Big(\frac{m_n^2}{n}\Big) + O_p\Big(\frac{m_n}{n}\Big) + O\Big(\frac{1}{m_n^{2d-1}}\Big) + O\left(\frac{4m_n^2\lambda_{n2}^2}{n^2}\right).$$

This theorem is concerned with the selection and estimation properties of the adaptive group Lasso in terms of $\widehat{\boldsymbol{\beta}}_n$. The following theorem states the results in terms of the estimators of the nonparametric components.

Theorem 4 *Suppose that conditions (B1), (B2) and (A2)-(A4) hold. Then*
(i)

$$\mathrm{P}\Big(\|\widehat{f}_{nj}\|_2 > 0, j \in A_1 \textit{ and } \|\widehat{f}_{nj}\|_2 = 0, j \in A_0\Big) \to 1.$$

(ii)

$$\sum_{j=1}^{q} \|\widehat{f}_{nj} - f_j\|_2^2 = O_p\Big(\frac{m_n}{n}\Big) + O_p\Big(\frac{1}{n}\Big) + O\Big(\frac{1}{m_n^{2d}}\Big) + O\left(\frac{4m_n\lambda_{n2}^2}{n^2}\right).$$

Part (i) of this theorem states that the adaptive group Lasso can consistently distinguishing nonzero components from zero components. Part (ii) gives an upper bound on the rate of convergence of the estimator.

Condition (B2) can be further simplified if we have $r_n \asymp n^{d/(2d+1)}/\sqrt{\log(p_n m_n)}$ in the initial estimator, e.g., the group Lasso with $\lambda_{n1} \asymp \sqrt{n \log(p_n m_n)}$ and use $m_n \asymp n^{1/(2d+1)}$ for $d \geq 1$. In this case, (B2) becomes

$$\frac{\lambda_{n2}}{n^{(8d+3)/(8d+4)}} = o(1) \quad \text{and} \quad \frac{n^{1/(4d+2)} \log(p_n m_n)}{\lambda_{n2}} = o(1). \tag{8}$$

We summarize the above discussion in the following corollary.

Corollary 2 *Let the group Lasso estimator* $\widetilde{\boldsymbol{\beta}}_n \equiv \widetilde{\boldsymbol{\beta}}_n(\lambda_{n1})$ *with* $\lambda_{n1} \asymp \sqrt{n \log(p_n m_n)}$ *be the initial estimator in the adaptive group Lasso. Suppose that the conditions of Theorem 1 hold. If* $\lambda_{n2} \geq O(n^{1/2})$ *and satisfies (8), then the adaptive group Lasso consistently selects the nonzero components in (1), that is, part (i) of Theorem 4 holds. In addition,*

$$\sum_{j=1}^{q} \|\widehat{f}_{nj} - f_j\|_2^2 = O_p\big(n^{-2d/(2d+1)}\big).$$

4 Simulation studies

We use simulation to evaluate the performance of the adaptive group Lasso with regard to variable selection. We compare it with the group Lasso and Lasso. Here the Lasso estimator is defined as the value that minimizes

$$\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_n \sum_{j=1}^{p} \sum_{k=1}^{m_n} |\beta_{jk}|.$$

The Lasso estimator does not take into account the grouping structure in the spline expansions of the components.

The generating model is

$$y_i = f(x_i) + \varepsilon_i \equiv \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, i = 1, \cdots, n.$$

We consider two cases, $p = 21$ and $p = 100$. In each case, the number of nonzero functions $q = 4$. Thus $f_j(x) \equiv 0$ for $j = 5, \cdots, p$. The sample size $n = 100$. We use the cubic B-spline with six evenly distributed knots for all the functions $f_k(x)$, The group Lasso and the adaptive group Lasso estimates are computed using the algorithm proposed by Yuan and Lin (2006). The Lasso estimates are computed using the Lars algorithm (Efron et al. 2004).

The group Lasso is used as the initial estimate for the adaptive group lasso. The penalty parameters in all the methods are chosen using the BIC (Schwarz 1978). The number of replications in all the simulations is 400.

Example 1.

$$f(x_i) = -7 + 8x_{i1} - 3x_{i2} + 10x_{i3}^2 - 6x_{i4}(x_{i4} - 1) + \epsilon_i,$$

The covariates are simulated in the following way. First, $w_{i1}, \cdots, w_{ip}, u_i, v_i$ are generated independently from $N(0,1)$ truncated to the interval $(0,1)$, $i = 1, \cdots, n$. Then we set $x_{ik} = (w_{ik} + tu_i)/(1+t)$ for $k = 1, \cdots, 4$ and $x_{ik} = (w_{ik} + tv_i)/(1+t)$ for $k = 5, \cdots, p$, where t controls the amount of correlation among predictors. We use $Cor(x_{ik}, x_{ij}) = t^2/(1+t^2), 1 \leq j \leq 4, 1 \leq k \leq 4$ and $Cor(x_{ik}, x_{ij}) = t^2/(1+t^2), 5 \leq j \leq p, 5 \leq k \leq p$, but the covariates of the nonzero components and zero components are independent. We consider $t = 0, 1$ in our simulation. The error term $\epsilon_i \sim N(0, 1.5^2)$.

Example 2.

$$f(x_i) = -5 + 8x_{i1}^3 + 10x_{i2}(1 - x_{i2}) - 10x_{i3}^5 - 8x_{i4}^2 + \epsilon_i,$$

where $x_{i1}, \cdots, x_{ip}$ are simulated in the same way as the covariates in Example 1 and the error term $\epsilon_i \sim N(0, 1.5^2)$.

Example 3.

$$f(x) = -4 + 4x_1 + \cos(2\pi x_2) - 8x_3^3 + \sqrt{x_4(1 - x_4)}\sin\left(\frac{2\pi(1 + 2^{(9-4s)/5})}{x + 2^{(9-4s)/5}}\right)$$

where $s = 3$. where $x_{i1}, \cdots, x_{ip}$ are simulated similar as Example 1 and $s = 3$ which controls the oscillation of the function $f_4(x_4)$ (Donoho and Johnstone (1994)). The covariates are simulated similarly to those in Example 1, except that $w_{i1}, \cdots, w_{ip}, u_i, v_i$ are now simulated independently from $U[0,1]$, $i = 1, \cdots, n$. The error term $\varepsilon_i \sim N(0, 1)$.

The results are summarized in Tables 1 and 2, based on 400 runs. The columns in these tables are: model error (ER), percentage of occasions on which the correct components are included in the selected model (IN%) and percentage of occasions on which correct components are selected (CS%), averaged over 400 replications. Enclosed in parentheses are the corresponding standard errors. The model error is computed as $n^{-1}\sum_{i=1}^{n}(\widehat{f}(x_i) - f(x_i))^2$, where f is the function given above in each example.

Several observations can be made from Tables 1 and 2. The adaptive group Lasso has higher percentage of occasions on which correct models are selected than the group Lasso and the Lasso. For each method, the results for $p = 21$ are better than those for $p = 100$ in terms of model error and variable selection. The is to be expected since variable selection is more difficult in a larger model than in a small one. The results for covariates independently simulated are better than those when covariates are correlated. Finally, the models selected by the group Lasso and adaptive group Lasso have similar model error to those selected by the Lasso, but have higher percentages of correct selections. This shows it is important to take into account the natural group structure in the spline based approach considered here.

These simulation results suggest that both the group Lasso and adaptive group Lasso are effective for component selection in sparse, high-dimensional nonparametric additive models, and that the adaptive group Lasso can considerably improve the selection results over the group Lasso.

5 Application to economic growth data

Sala-i-Martin (1997) investigated the relation between economic growth and a variety of covariates in 99 countries. His model is

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i,$$

where Y_i is the average rate of growth of the gross domestic product of country i from 1960-1992, X_{ij} is the jth covariate in country i, β_0 and the β_js are intercept and regression coefficients, and ε_i is an unobserved random variable with mean 0. The data are available at *www.columbia.edu/∼xs23/data/millions.htm*. They were compiled from a variety of sources and include 59 potential covariates that describe economic, political, social, and geographical characteristics of the countries. Sala-i-Martin (1997) used a heuristic method to select covariates for inclusion in the model. This method required repeatedly estimating model (10) by ordinary least squares using some 2 million different subsets of the covariates. Sala-i-Martin (1997) found that 22 of the 59 variables were significant. He noted that although his model is linear, other investigators have found nonlinear relations between the covariates and economic growth. Many of the covariates in Sala-i-Martins data are binary. We used the methods of Sections 2-4 to model the relation between the economic growth rate and 21 continuous covariates in the data. The covariates are listed in Table 3. We estimated model (1) using the ordinary (not group) Lasso, group Lasso, and adaptive group Lasso. We scaled the covariates so that their values are between 0 and 1 and used cubic splines with 7 knots to estimate the additive components. The check signs in Table 3 indicate the variables that were selected by the three methods. With Lasso, a variable is considered to be selected if any of the estimated coefficients of the spline approximation to its additive component are non-zero. Table 4 shows the residual sums of squares obtained with each of the estimation methods. The group Lasso and adaptive group Lasso select the same 13 variables out of the total of 21. The ordinary Lasso selects all of the variables, so it does not reduce the dimension of the model. Figures 1 and 2 shows the estimated additive components. Many are highly nonlinear, confirming the need for a model selection method that takes account of nonlinearity.

6 Concluding remarks

In this paper, we have studied the asymptotic properties of the group Lasso and adaptive group Lasso for variable selection in nonparametric additive models when p is large. An important condition required in our results is that the number of nonzero components is fixed. While this condition is reasonable is many applications, it would be interesting to relax this condition and investigate the case when the number of nonzero components can also increase with the sample size. Clearly, there needs to be restriction on the number of nonzero components so that the model is identifiable.

We have only considered the linear nonparametric additive models. The adaptive group Lasso can be applied to generalized nonparametric additive models, such as the generalized logistic nonparametric additive model and other nonparametric models with high-dimensional data. However, more work is needed to understand the properties of this approach in those more complicated models.

ACKNOWLEDGEMENTS. The work of Huang is supported in part by NSF grants DMS-0706108 and DMS-0805670 and NIH grant R01CA120988. The work of Horowitz is supported in part by NSF grants DMS 0706348 and SES 0817552.

7 Proofs

We first prove the following lemmas. Denote the centered versions of $\mathcal{S}_n$ by

$$\mathcal{S}_{nj}^0 = \Big\{ f_{nj} : f_{nj}(x) = \sum_{k=1}^{m_n} b_{jk}\psi_k(x),\ (\beta_{j1}, \ldots, \beta_{jm_n}) \in \mathbb{R}^{m_n} \Big\},\ 1 \le j \le p,$$

where ψ_k's are the centered spline bases defined in (4).

Lemma 1 *Suppose that $f \in \mathcal{F}$ and $\mathrm{E}f(X_j) = 0$. Then, under (A3) and (A4), there exists an $f_n \in \mathcal{S}_{nj}^0$ satisfying*

$$\|f_n - f\|_2 = O_p(m_n^{-d} + m_n^{1/2}n^{-1/2}).$$

In particular, if we choose $m_n = O(n^{1/(2d+1)})$, then

$$\|f_n - f\|_2 = O_p(m_n^{-d}) = O_p(n^{-d/(2d+1)}).$$

Proof of Lemma 1. By (A4), for $f \in \mathcal{F}$, there is an $f_n^* \in \mathcal{S}_n$ such that $\|f - f_n^*\|_2 = O(m_n^{-d})$. Let $f_n = f_n^* - n^{-1}\sum_{i=1}^n f_n^*(X_{ij})$. Then $f_n \in \mathcal{S}_{nj}^0$ and $|f_n - f| \le |f_n^* - f| + |P_n f_n^*|$, where P_n is the empirical measure of iid random variables $X_{1j}, \ldots, X_{nj}$. Consider

$$P_n f_n^* = (P_n - P) f_n^* + P(f_n^* - f).$$

Here we use the linear functional notation, i.e., $Pf = \int f dP$, where P is the probability measure of X_{1j}. Now $(P_n - P) f_n^* = O_p(n^{-1/2} m_n^{1/2})$, and by (A4), $|P(f_n^* - f)| \le C_2 \|f_n^* - f\|_2 = O(m_n^{-d})$ for some constant $C_2 > 0$. The lemma follows from the triangle inequality. □

Lemma 2 *Suppose that conditions (A2) and (A4) hold. Let* $T_{jk} = n^{-1/2} m_n^{1/2} \sum_{i=1}^n \psi_k(X_{ij})\varepsilon_i, 1 \le j \le p, 1 \le k \le m_n$. *Let* $T_n = \max_{1\le j\le p, 1\le k\le m_n} |T_{jk}|$. *Then*

$$\mathrm{E}(T_n) \le C_1 n^{-1/2} m_n^{1/2} \sqrt{\log(pm_n)} \Big(\sqrt{2C_2 m_n^{-1} n \log(pm_n)} + 4\log(2pm_n) + C_2 n m_n^{-1} \Big)^{1/2}.$$

where C_1 *and* C_2 *are two positive constants. In particular, when* $m_n \log(pm_n)/n \to 0$,

$$\mathrm{E}(T_n) = O(1)\sqrt{\log(pm_n)}.$$

Proof of Lemma 2. Let $s_{njk}^2 = \sum_{i=1}^n \psi_k^2(X_{ij})$. Conditional on X_{ij}'s, $T_{jk} \sim N(0, s_{njk}^2)$. Let $s_n^2 = \max_{1\le j\le p, 1\le k\le m_n} s_{njk}^2$. By (A2) and the maximal inequality for subgaussian random variables (Van der Vaart and Wellner 1996, Lemmas 2.2.1 and 2.2.2),

$$\mathrm{E}\Big(\max_{1\le j\le p, 1\le k\le m_n} |T_{jk}| \,\Big|\{X_{ij}, 1\le i\le n, 1\le j\le p\}\Big) \le C_1 n^{-1/2} m_n^{1/2} s_n \sqrt{\log(pm_n)}.$$

Therefore,

$$\mathrm{E}\Big(\max_{1\le j\le p, 1\le k\le m_n} |T_{jk}|\Big) \le C_1 n^{-1/2} m_n^{1/2} \sqrt{\log(pm_n)}\, \mathrm{E}(s_n), \tag{9}$$

where $C_1 > 0$ is a constant. By (A4) and the properties of B-splines,

$$|\psi_k(X_{ij})| \le |\phi_k(X_{ij})| + |\bar{\phi}_{jk}| \ \le 2 \text{ and } \mathrm{E}(\psi_k(X_{ij}))^2 \le C_2 m_n^{-1}, \tag{10}$$

for a constant $C_2 > 0$, for every $1 \le j \le p$ and $1 \le k \le m_n$. By (10),

$$\sum_{i=1}^n \mathrm{E}[\psi_k^2(X_{ij}) - \mathrm{E}\psi_k^2(X_{ij})]^2 \le 4C_2 n m_n^{-1}, \tag{11}$$

and

$$\max_{1\le j\le p, 1\le k\le m_n} \sum_{i=1}^n \mathrm{E}\psi_k^2(X_{ij}) \le C_2 n m_n^{-1}. \tag{12}$$

By Lemma 4.2 of Van de Geer (2008), (10) and (11) imply

$$\mathrm{E}\Big(\max_{1\le j\le p, 1\le k\le m_n}\Big|\sum_{i=1}^{n}\{\psi_k^2(X_{ij})-\mathrm{E}\psi_k^2(X_{ij})\}\Big|\Big)\le\sqrt{2C_2m_n^{-1}n\log(pm_n)}+4\log(2pm_n).$$

Therefore, by (12) and the triangle inequality,

$$\mathrm{E}s_n^2\le\sqrt{2C_2m_n^{-1}n\log(pm_n)}+4\log(2pm_n)+C_2nm_n^{-1}.$$

Now since $\mathrm{E}s_n\le(\mathrm{E}s_n^2)^{1/2}$, we have

$$\mathrm{E}s_n\le\Big(\sqrt{2C_2m_n^{-1}n\log(pm_n)}+4\log(2pm_n)+C_2nm_n^{-1}\Big)^{1/2}. \tag{13}$$

The lemma follows from (9) and (13). □

Recall $\mathbf{C}_A=n^{-1}\mathbf{Z}_A'\mathbf{Z}_A$, where $A\subset\{1,\dots,p\}$. Let $\rho_{\min}(\mathbf{C}_A)$ and $\rho_{\max}(\mathbf{C}_A)$ be the minimum and maximum eigenvalues of $\mathbf{C}_A$, respectively.

Lemma 3 *Let $m_n=O(n^\gamma)$ where $0<\gamma<0.5$. Suppose that $|A|$ is bounded by a fixed constant independent of n and p. Let $h\equiv h_n\asymp m_n^{-1}$. Then, under (A3) and (A4), with probability converging to one,*

$$c_1h_n\le\rho_{\min}(\mathbf{C}_A)\le\rho_{\max}(\mathbf{C}_A)\le c_2h_n,$$

where c_1 and c_2 are two positive constants.

Proof of Lemma 3. Without loss of generality, suppose $A=\{1,\dots,k\}$. Then $\mathbf{Z}_A=(\mathbf{Z}_1,\dots,\mathbf{Z}_q)$. Let $\mathbf{b}=(\mathbf{b}_1',\dots,\mathbf{b}_q')'$, where $\mathbf{b}_j\in R^{m_n}$. By Lemma 3 of Stone (1985),

$$\|\mathbf{Z}_1\mathbf{b}_1+\cdots+\mathbf{Z}_q\mathbf{b}_q\|_2\ge c_3(\|\mathbf{Z}_1\mathbf{b}_1\|_2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2)$$

for a certain constant $c_3>0$. By the triangle inequality,

$$\|\mathbf{Z}_1\mathbf{b}_1+\cdots+\mathbf{Z}_q\mathbf{b}_q\|_2\le\|\mathbf{Z}_1\mathbf{b}_1\|_2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2.$$

Since $\mathbf{Z}_A\mathbf{b}=\mathbf{Z}_1\mathbf{b}_1+\cdots+\mathbf{Z}_q\mathbf{b}_q$, the above two inequalities imply that

$$c_3(\|\mathbf{Z}_1\mathbf{b}_1\|_2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2)\le\|\mathbf{Z}_A\mathbf{b}\|_2\le\|\mathbf{Z}_1\mathbf{b}_1\|_2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2.$$

Therefore,

$$c_3^2(\|\mathbf{Z}_1\mathbf{b}_1\|_2^2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2^2)\le\|\mathbf{Z}_A\mathbf{b}\|_2^2\le2(\|\mathbf{Z}_1\mathbf{b}_1\|_2^2+\cdots+\|\mathbf{Z}_q\mathbf{b}_q\|_2^2). \tag{14}$$

Let $\mathbf{C}_j = n^{-1}\mathbf{Z}_j'\mathbf{Z}_j$. By Lemma 6.2 of Shen, Wolf and Zhou (1998),

$$c_4 h \le \rho_{\min}(\mathbf{C}_j) \le \rho_{\max}(\mathbf{C}_j) \le c_5 h, j \in A. \tag{15}$$

Since $\mathbf{C}_A = n^{-1}\mathbf{Z}_A'\mathbf{Z}_A$, it follows from (14) that

$$c_3^2\left(\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1 + \cdots + \mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q\right) \le \mathbf{b}'\mathbf{C}_A\mathbf{b} \le 2\left(\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1 + \cdots + \mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q\right).$$

Therefore, by (15),

$$\begin{aligned}\frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}\|_2^2} + \cdots + \frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}\|_2^2} &= \frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}_1\|_2^2}\frac{\|\mathbf{b}_1\|_2^2}{\|\mathbf{b}\|_2^2} + \cdots + \frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}_q\|_2^2}\frac{\|\mathbf{b}_q\|_2^2}{\|\mathbf{b}\|_2^2}\\ &\ge \rho_{\min}(\mathbf{C}_1)\frac{\|\mathbf{b}_1\|_2^2}{\|\mathbf{b}\|_2^2} + \cdots + \rho_{\min}(\mathbf{C}_q)\frac{\|\mathbf{b}_q\|_2^2}{\|\mathbf{b}\|_2^2}\\ &\ge c_4 h.\end{aligned}$$

Similarly,

$$\frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}\|_2^2} + \cdots + \frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}\|_2^2} \le c_5 h.$$

Thus we have

$$c_3^2 c_4 h \le \frac{\mathbf{b}'\mathbf{C}_A\mathbf{b}}{\mathbf{b}'\mathbf{b}} \le 2c_5 h.$$

Thus the lemma follows. □

Proof of Theorem 1. The proof of parts (i) and (ii) essentially follows the proof of Theorem 1 of Wei and Huang (2008). The only change that must be made here is that we need to consider the approximation error of the regression functions by splines. Specifically, let $\boldsymbol{\xi}_n = \boldsymbol{\varepsilon}_n + \boldsymbol{\delta}_n$, where $\boldsymbol{\delta}_n = (\delta_{n1}, \ldots, \delta_{nn})'$ with $\delta_{ni} = \sum_{j=1}^{q_n}(f_{0j}(X_{ij}) - f_{nj}(X_{ij}))$. Since $\|f_{0j} - f_{nj}\|_2 = O(m_n^{-d}) = O(n^{-d/(2d+1)})$ for $m_n = n^{1/(2d+1)}$, we have

$$\|\boldsymbol{\delta}_n\|_2 \le C_1\sqrt{nqm_n^{-2d}} = C_1 q n^{1/(4d+2)},$$

for some constant $C_1 > 0$. For any integer t, let

$$\chi_t = \max_{|A|=t}\max_{\|U_{A_k}\|_2=1, 1\le k\le t}\frac{|\boldsymbol{\xi}_n' V_A(\mathbf{s})|}{\|V_A(\mathbf{s})\|_2} \text{ and } \chi_t^* = \max_{|A|=t}\max_{\|U_{A_k}\|_2=1, 1\le k\le t}\frac{|\boldsymbol{\varepsilon}_n' V_A(\mathbf{s})|}{\|V_A(\mathbf{s})\|_2}$$

where $V_A(S_A) = \boldsymbol{\xi}_n'(\mathbf{Z}_A(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\bar{S}_A - (I - P_A)X\boldsymbol{\beta}$ for $N(A) = q_1 = m \ge 0$, $S_A = (S_{A_1}', \cdots, S_{A_m}')'$, $S_{A_k} = \lambda\sqrt{d_{A_k}}U_{A_k}$ and $\|U_{A_k}\|_2 = 1$.

For a sufficiently large constant $C_2 > 0$, define

$$\Omega_{t_0} = \{(\mathbf{Z}, \boldsymbol{\varepsilon}_n) : x_t \le \sigma C_2\sqrt{((t \vee 1)m_n \log(pm_n)}, \forall t \ge t_0\},$$

and

$$\Omega_{t_0}^* = \{(\mathbf{Z}, \boldsymbol{\varepsilon}_n) : x_t^* \le \sigma C_2 \sqrt{(t \vee 1) m_n \log(p m_n)}, \forall t \ge t_0\},$$

where $t_0 \ge 0$.

As in the proof of Theorem 1 of Wei and Huang (2008) and in the proof of Theorem 1 of Zhang and Huang (2008),

$$(\mathbf{Z}, \boldsymbol{\varepsilon}_n) \in \Omega_q \Rightarrow |\widetilde{A}_1| \le M_1^*(\lambda_{n1}) q.$$

By the triangle and Cauchy-Schwarz inequalities,

$$\frac{|\boldsymbol{\xi}_n' V_A(\mathbf{s})|}{\|V_A(\mathbf{s})\|_2} = \frac{|\boldsymbol{\varepsilon}_n' V_A(\mathbf{s}) + \boldsymbol{\delta}_n' V_A(\mathbf{s})|}{\|V_A(\mathbf{s})\|_2} \le \frac{|\boldsymbol{\varepsilon}_n' V_A(\mathbf{s})|}{\|V_A\|_2} + \|\boldsymbol{\delta}_n\|. \tag{16}$$

In the proof of Theorem 1 of Wei and Huang (2008), it is shown that

$$\mathrm{P}(\Omega_0^*) \ge 2 - \frac{2}{p^{1+c_0}} - \exp\left(\frac{2p}{p^{1+c_0}}\right) \to 1. \tag{17}$$

Since

$$\frac{|\boldsymbol{\delta}_n' V_A(\mathbf{s})|}{\|V_A(\mathbf{s})\|_2} \le \|\boldsymbol{\delta}_n\|_2 \le= C_1 q n^{\frac{1}{2(2d+1)}}$$

and $m_n = O(n^{1/(2d+1)})$, we have for all $t \ge 0$ and n sufficiently large,

$$\|\boldsymbol{\delta}_n\|_2 \le C_1 q n^{\frac{1}{2(2d+1)}} \le \sigma C_2 \sqrt{(t \vee 1) m_n \log(p)}\,. \tag{18}$$

It follows from (16), (17) and (18) that $\mathrm{P}(\Omega_0) \to 1$. This completes the proof of part (i) of Theorem 1.

Before proving part (ii), we first prove part (iii) of Theorem 1. By the definition of $\widetilde{\boldsymbol{\beta}}_n \equiv (\widetilde{\boldsymbol{\beta}}_{n1}', \ldots, \widetilde{\boldsymbol{\beta}}_{np}')'$,

$$\|\mathbf{Y} - \mathbf{Z}\widetilde{\boldsymbol{\beta}}_n\|_2^2 + \lambda_{n1} \sum_{j=1}^p \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \le \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_{n1} \sum_{j=1}^p \|\boldsymbol{\beta}_{nj}\|_2. \tag{19}$$

Let $A_2 = \{j : \|\boldsymbol{\beta}_{nj}\|_2 \ne 0 \text{ or } \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \ne 0\}$ and $d_{n2} = |A_2|$. By part (i), $d_{n2} = O_p(q)$. By (19) and the definition of A_2,

$$\|\mathbf{Y} - \mathbf{Z}_{A_2}\widetilde{\boldsymbol{\beta}}_{nA_2}\|_2^2 + \lambda_{n1} \sum_{j \in A_2} \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \le \|\mathbf{Y} - \mathbf{Z}_{A_2}\boldsymbol{\beta}_{nA_2}\|_2^2 + \lambda_{n1} \sum_{j \in A_2} \|\boldsymbol{\beta}_{nj}\|_2. \tag{20}$$

Let $\boldsymbol{\eta}_n = \mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n$. Write

$$\mathbf{Y} - \mathbf{Z}_{A_2}\widetilde{\boldsymbol{\beta}}_{nA_2} = \mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n - \mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}) = \boldsymbol{\eta}_n - \mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}).$$

We have

$$\|\mathbf{Y} - \mathbf{Z}_{A_2}\widetilde{\boldsymbol{\beta}}_{nA_2}\|_2^2 = \|\mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2})\|_2^2 - 2\boldsymbol{\eta}_n'\mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}) + \boldsymbol{\eta}_n'\boldsymbol{\eta}_n.$$

We can rewrite (20) as

$$\|\mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2})\|_2^2 - 2\boldsymbol{\eta}_n'\mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}) \le \lambda_{n1}\sum_{j\in A_1}\|\boldsymbol{\beta}_{nj}\|_2 - \lambda_{n1}\sum_{j\in A_1}\|\widetilde{\boldsymbol{\beta}}_{nj}\|_2. \quad (21)$$

Now

$$\Big|\sum_{j\in A_1}\|\boldsymbol{\beta}_{nj}\|_2 - \sum_{j\in A_1}\|\widetilde{\boldsymbol{\beta}}_{nj}\|_2\Big| \le \sqrt{|A_1|}\cdot\|\widetilde{\boldsymbol{\beta}}_{nA_1} - \boldsymbol{\beta}_{nA_1}\|_2 \le \sqrt{|A_1|}\cdot\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2. \quad (22)$$

Let $\boldsymbol{\nu}_n = \mathbf{Z}_{A_2}(\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2})$. Combining (20), (21) and (22) to get

$$\|\boldsymbol{\nu}_n\|_2^2 - 2\boldsymbol{\eta}_n'\boldsymbol{\nu}_n \le \lambda_{n1}\sqrt{|A_1|}\cdot\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2. \quad (23)$$

Let $\boldsymbol{\eta}_n^*$ be the projection of $\boldsymbol{\eta}_n$ to the span of $\mathbf{Z}_{A_2}$, that is, $\boldsymbol{\eta}_n^* = \mathbf{Z}_{A_2}(\mathbf{Z}_{A_2}'\mathbf{Z}_{A_2})^{-1}\mathbf{Z}_{A_2}'\boldsymbol{\eta}_n$. By the Cauchy-Schwartz inequality,

$$2|\boldsymbol{\eta}_n'\boldsymbol{\nu}_n| \le 2\|\boldsymbol{\eta}_n^*\|_2\cdot\|\boldsymbol{\nu}_n\|_2 \le 2\|\boldsymbol{\eta}_n^*\|_2^2 + \frac{1}{2}\|\boldsymbol{\nu}_n\|_2^2. \quad (24)$$

From (23) and (24), we have

$$\|\boldsymbol{\nu}_n\|_2^2 \le 4\|\boldsymbol{\eta}_n^*\|_2^2 + 2\lambda_{n1}\sqrt{|A_1|}\cdot\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2.$$

Let c_{n*} be the smallest eigenvalue of $\mathbf{Z}_{A_2}'\mathbf{Z}_{A_2}/n$. By Lemma 3 and part (i), $c_{n*} \asymp_p m_n^{-1}$. Since $\|\boldsymbol{\nu}_n\|_2^2 \ge nc_{n*}\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2^2$ and $2ab \le a^2 + b^2$,

$$nc_{n*}\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2^2 \le 4\|\boldsymbol{\eta}_n^*\|_2^2 + \frac{(2\lambda_{n1}\sqrt{|A_1|})^2}{2nc_{n*}} + \frac{1}{2}nc_{n*}\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2^2.$$

It follows that

$$\|\widetilde{\boldsymbol{\beta}}_{nA_2} - \boldsymbol{\beta}_{nA_2}\|_2^2 \le \frac{8\|\boldsymbol{\eta}_n^*\|_2^2}{nc_{n*}} + \frac{4\lambda_{n1}^2|A_1|}{n^2c_{n*}^2}. \quad (25)$$

Let $f_0(\mathbf{X}_i) = \sum_{j=1}^p f_{0j}(X_{ij})$ and $f_{0A}(\mathbf{X}_i) = \sum_{j\in A} f_{0j}(X_{ij})$. Write

$$\eta_i = Y_i - \mu - f_0(\mathbf{X}_i) + (\mu - \overline{Y}) + f_0(\mathbf{X}_i) - \sum_{j\in A_2} Z'_{ij}\boldsymbol{\beta}_{nj} = \varepsilon_i + (\mu - \overline{Y}) + f_{A_2}(\mathbf{X}_i) - f_{nA_2}(\mathbf{X}_i).$$

Since $|\mu - \overline{Y}|^2 = O_p(n^{-1})$ and $\|f_{0j} - f_{nj}\|_\infty = O(m_n^{-d})$, we have

$$\|\boldsymbol{\eta}_n^*\|_2^2 \le 2\|\boldsymbol{\varepsilon}_n^*\|_2^2 + O_p(1) + O(nd_{n2}m_n^{-2d}), \tag{26}$$

where $\boldsymbol{\varepsilon}_n^*$ is the projection of $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ to the span of $\mathbf{Z}_{A_2}$. We have

$$\|\boldsymbol{\varepsilon}_n^*\|_2^2 = \|(\mathbf{Z}'_{A_2}\mathbf{Z}_{A_2})^{-1/2}\mathbf{Z}'_{A_2}\boldsymbol{\varepsilon}_n\|_2^2 \le \frac{1}{nc_{n*}}\|\mathbf{Z}'_{A_2}\boldsymbol{\varepsilon}_n\|_2^2.$$

Now

$$\max_{A:|A|\le d_{n2}} \|\mathbf{Z}'_A\boldsymbol{\varepsilon}_n\|_2^2 = \max_{A:|A|\le d_{n2}} \sum_{j\in A} \|\mathbf{Z}'_j\boldsymbol{\varepsilon}_n\|_2^2 \le d_{n2}m_n \max_{1\le j\le p, 1\le k\le m_n} |\mathcal{Z}'_{jk}\boldsymbol{\varepsilon}|^2,$$

where $\mathcal{Z}_{jk} = (\psi_k(X_{1j}), \ldots, \psi_k(X_{nj}))'$. By Lemma 2,

$$\max_{1\le j\le p, 1\le k\le m_n} |\mathcal{Z}'_{jk}\boldsymbol{\varepsilon}_n|^2 = nm_n^{-1} \max_{1\le j\le p, 1\le k\le m_n} |(m_n/n)^{1/2}\mathcal{Z}'_{jk}\boldsymbol{\varepsilon}_n|^2 = O_p(1)nm_n^{-1}\log(pm_n).$$

It follows that,

$$\|\boldsymbol{\varepsilon}_n^*\|_2^2 = O_p(1)\frac{d_{n2}\log(pm_n)}{c_{n*}}. \tag{27}$$

Combining (25), (26), and (27), we get

$$\|\widetilde{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}_{A_2}\|_2^2 \le O_p\Big(\frac{d_{n2}\log(pm_n)}{nc_{n*}^2}\Big) + O_p\Big(\frac{1}{nc_{n*}}\Big) + O\Big(\frac{d_{n2}m_n^{-2d}}{c_{n*}}\Big) + \frac{4\lambda_{n1}^2|A_1|}{n^2c_{n*}^2}.$$

Since $d_{n2} = O_p(q)$, $c_{n*} \asymp_p m_n^{-1}$ and $c_n^* \asymp_p m_n^{-1}$, we have

$$\|\widetilde{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}_{A_2}\|_2^2 \le O_p\Big(\frac{m_n^2\log(pm_n)}{n}\Big) + O_p\Big(\frac{m_n}{n}\Big) + O\Big(\frac{1}{m_n^{2d-1}}\Big) + O\left(\frac{4m_n^2\lambda_{n1}^2}{n^2}\right).$$

This completes the proof of part (iii).

We now prove part (ii). Since $\|f_j\|_2 \ge c_f > 0, 1 \le j \le q$, $\|f_j - f_{nj}\|_2 = O(m_n^{-d})$ and $\|f_{nj}\|_2 \ge \|f_j\|_2 - \|f_j - f_{nj}\|_2$, we have $\|f_{nj}\|_2 \ge 0.5c_f$ for n sufficiently large. By a result of de Boor (2001), see also (12) of Stone (1986), there are positive constants c_6 and c_7 such that

$$c_6m_n^{-1}\|\boldsymbol{\beta}_n\|_2^2 \le \|f_{nj}\|_2^2 \le c_7m_n^{-1}\|\boldsymbol{\beta}_{nj}\|_2^2.$$

It follows that,

$$\|\boldsymbol{\beta}_{nj}\|_2^2 \ge c_7^{-1}m_n\|f_{nj}\|_2^2 \ge 0.25c_7^{-1}c_f^2m_n.$$

Therefore, if $\|\boldsymbol{\beta}_{nj}\|_2 \neq 0$ but $\|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 = 0$, then

$$\|\widetilde{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2 \geq 0.25 c_7^{-1} c_f^2 m_n. \tag{28}$$

However, since $(m_n \log(pm_n))/n \to 0$ and $(\lambda_{n1}^2 m_n)/n^2 \to$, (28) contradicts part (iii). □

Proof of Theorem 2. By the definition of $\widetilde{f}_j, 1 \leq j \leq p$, parts (i) and (ii) follow from parts (i) and (ii) of Theorem 1 directly.

Now consider part (iii). By the properties of spline (de Boor (2001)),

$$c_6 m_n^{-1} \|\widetilde{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2 \leq \|\widetilde{f}_{nj} - f_{nj}\|_2^2 \leq c_7 m_n^{-1} \|\widetilde{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2.$$

Thus

$$\|\widetilde{f}_{nj} - f_{nj}\|_2^2 = O_p\Big(\frac{m_n \log(pm_n)}{n}\Big) + O_p\left(\frac{1}{n}\right) + O\Big(\frac{1}{m_n^{2d}}\Big) + O\left(\frac{4m_n\lambda_{n1}^2}{n^2}\right). \tag{29}$$

By (A3),

$$\|f_j - f_{nj}\|_2^2 = O(m_n^{-2d}). \tag{30}$$

Part (iii) follows from (29) and (30). □

In the proofs below, for any matrix $\mathbf{H}$, denote its 2-norm by $\|\mathbf{H}\|$, which is equal to its largest eigenvalue. This norm satisfies the inequality $\|\mathbf{H}\mathbf{x}\| \leq \|\mathbf{H}\|\|\mathbf{x}\|$ for a column vector $\mathbf{x}$ whose dimension is the same as the number of the columns of $\mathbf{H}$.

Proof of Theorem 3. By the KKT, a necessary and sufficient condition for $\widehat{\boldsymbol{\beta}}_n$ is

$$\begin{cases} 2\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\beta}}_n) = \lambda_{n2} w_{nj} \frac{\widehat{\boldsymbol{\beta}}_{nj}}{\|\widehat{\boldsymbol{\beta}}_{nj}\|}, & \|\widehat{\boldsymbol{\beta}}_j\|_2 \neq 0, j \geq 1, \\ 2\|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\beta}}_n)\|_2 < \lambda_{n2} w_{nj}, & \|\widehat{\boldsymbol{\beta}}_{nj}\| = 0, j \geq 1. \end{cases} \tag{31}$$

Let $\boldsymbol{\nu}_n = (w_{nj}\widehat{\boldsymbol{\beta}}_j/(2\|\widehat{\boldsymbol{\beta}}_{nj}\|), j \in A_1)'$. Define

$$\widehat{\boldsymbol{\beta}}_{nA_1} = (\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1})^{-1}(\mathbf{Z}_{A_1}'\mathbf{Y} - \lambda_{n2}\boldsymbol{\nu}_n). \tag{32}$$

If $\widehat{\boldsymbol{\beta}}_{nA_1} =_0 \boldsymbol{\beta}_{nA_1}$, then the equation in (31) holds for $\widehat{\boldsymbol{\beta}}_n \equiv (\widehat{\boldsymbol{\beta}}_{nA_1}', \mathbf{0}')'$. Thus, since $\mathbf{Z}\widehat{\boldsymbol{\beta}}_n = \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1}$ for this $\widehat{\boldsymbol{\beta}}_n$ and $\{\mathbf{Z}_j, j \in A_1\}$ are linearly independent,

$$\widehat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n \quad \text{if} \quad \begin{cases} \widehat{\boldsymbol{\beta}}_{nA_1} =_0 \boldsymbol{\beta}_{nA_1} \\ \|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1})\|_2 \leq \lambda_{n2} w_{nj}/2, \ \forall j \notin A_1. \end{cases}$$

This is true if

$$\widehat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n \quad \text{if} \quad \begin{cases} \|\boldsymbol{\beta}_{nj}\|_2 - \|\widehat{\boldsymbol{\beta}}_{nj}\|_2 < \|\boldsymbol{\beta}_{nj}\|_2, & \forall j \in A_1, \\ \|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1})\|_2 \le \lambda_{n2} w_{nj}/2, & \forall j \notin A_1. \end{cases}$$

Therefore,

$$\begin{aligned} \mathrm{P}(\widehat{\boldsymbol{\beta}}_n \neq_0 \boldsymbol{\beta}_n) \;\; &\le \;\; \mathrm{P}\Big(\|\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2 \ge \|\boldsymbol{\beta}_{nj}\|_2, \exists j \in A_1\Big) \\ &\quad + \mathrm{P}\Big(\|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1})\|_2 > \lambda_{n2} w_{nj}/2, \exists j \notin A_1\Big). \end{aligned}$$

Let $f_{0j}(\mathbf{X}_j) = (f_{0j}(X_{1j}), \ldots, f_{0j}(X_{nj}))'$ and $\boldsymbol{\delta}_n = \sum_{j \in A_1} f_{0j}(\mathbf{X}_j) - \mathbf{Z}_{A_1}\boldsymbol{\beta}_{nA_1}$. By Lemma 1, we have

$$n^{-1}\|\boldsymbol{\delta}_n\|^2 = O_p(q m_n^{-2d}). \tag{33}$$

Let $\mathbf{H}_n = \mathbf{I}_n - \mathbf{Z}_{A_1}(\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1})^{-1}\mathbf{Z}_{A_1}'$. By (32),

$$\widehat{\boldsymbol{\beta}}_{nA_1} - \boldsymbol{\beta}_{nA_1} = n^{-1}\mathbf{C}_{A_1}^{-1}\big(\mathbf{Z}_{A_1}'(\boldsymbol{\varepsilon}_n + \boldsymbol{\delta}_n) - \lambda_{n2}\boldsymbol{\nu}_n\big), \tag{34}$$

and

$$\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1} = \mathbf{H}_n\boldsymbol{\varepsilon}_n + \mathbf{H}_n\boldsymbol{\delta}_n + \lambda_{n2}\mathbf{Z}_{A_1}\mathbf{C}_{A_1}^{-1}\boldsymbol{\nu}_n/n. \tag{35}$$

Based on these two equations, Lemma 5 below shows that

$$\mathrm{P}\Big(\|\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2 \ge \|\boldsymbol{\beta}_{nj}\|_2, \exists j \in A_1\Big) \to 0,$$

and Lemma 6 below shows that

$$\mathrm{P}\Big(\|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1})\|_2 > \lambda_{n2} w_{nj}/2, \exists j \notin A_1\Big) \to 0.$$

These two equations lead to part (i) of the theorem.

We now prove part (ii) of Theorem 3. As in (26), for $\boldsymbol{\eta}_n = \mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n$ and $\boldsymbol{\eta}_{n1}^* = \mathbf{Z}_{A_1}(\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1})^{-1}\mathbf{Z}_{A_1}'\boldsymbol{\eta}_n$, we have

$$\|\boldsymbol{\eta}_{n1}^*\|_2^2 \le 2\|\boldsymbol{\varepsilon}_{n1}^*\|_2^2 + O_p(1) + O(q n m_n^{-2d}), \tag{36}$$

where $\boldsymbol{\varepsilon}_{n1}^*$ is the projection of $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ to the span of $\mathbf{Z}_{A_1}$. We have

$$\|\boldsymbol{\varepsilon}_{n1}^*\|_2^2 = \|(\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1})^{-1/2}\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2^2 \le \frac{1}{n\rho_{n1}}\|\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2^2 = O_p(1)\frac{|A_1|}{\rho_{n1}}. \tag{37}$$

Now similarly to the proof of (25), we can show that

$$\|\widehat{\boldsymbol{\beta}}_{nA_1} - \boldsymbol{\beta}_{nA_1}\|_2^2 \leq \frac{8\|\boldsymbol{\eta}_{n1}^*\|_2^2}{n\rho_{n1}} + \frac{4\lambda_{n2}^2|A_1|}{n^2\rho_{n1}^2}. \tag{38}$$

Combining (36), (37) and (38), we get

$$\|\widehat{\boldsymbol{\beta}}_{nA_1} - \boldsymbol{\beta}_{nA_1}\|_2^2 = O_p\Big(\frac{8}{n\rho_{n1}^2}\Big) + O_p\Big(\frac{1}{n\rho_{n1}}\Big) + O\Big(\frac{1}{m_n^{2d-1}}\Big) + O\left(\frac{4\lambda_{n2}^2}{n^2\rho_{n1}^2}\right).$$

Since $\rho_{n1} \asymp_p m_n^{-1}$, the result follows.

The following lemmas are needed in the proof of Theorem 3.

Lemma 4 *For $\boldsymbol{\nu}_n = (w_{nj}\widetilde{\boldsymbol{\beta}}_j/(2\|\widetilde{\boldsymbol{\beta}}_{nj}\|), j \in A_1)'$, under condition (B1),*

$$\|\boldsymbol{\nu}_n\|^2 = O_p(h_n^2) = O_p\big((b_{n1}^2c_b)^{-2}r_n^{-1} + qb_{n1}^{-1}\big).$$

Proof of Lemma 4. Write

$$\|\boldsymbol{\nu}_n\|^2 = \sum_{j\in A_1} w_j^2 = \sum_{j\in A_1} \|\widetilde{\boldsymbol{\beta}}_{nj}\|^{-2} = \sum_{j\in A_1} \frac{\|\boldsymbol{\beta}_{nj}\|^2 - \|\widetilde{\boldsymbol{\beta}}_{nj}\|^2}{\|\boldsymbol{\beta}_{nj}\|^2 \cdot \|\widetilde{\boldsymbol{\beta}}_{nj}\|^2} + \sum_{j\in A_1} \|\boldsymbol{\beta}_{nj}\|^{-1}.$$

Under (B2),

$$\sum_{j\in A_1} \frac{\big|\|\boldsymbol{\beta}_{nj}\|^2 - \|\widetilde{\boldsymbol{\beta}}_{nj}\|^2\big|}{\|\boldsymbol{\beta}_{nj}\|^2 \cdot \|\widetilde{\boldsymbol{\beta}}_{nj}\|^2} \leq Mc_b^{-2}b_{n1}^{-4}\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\|,$$

and

$$\sum_{j\in A_1} \|\boldsymbol{\beta}_{nj}\|^{-2} \leq qb_{n1}^{-2}.$$

The claim follows. □

Let ρ_{n3} be the maximum of the largest eigenvalues of $n^{-1}\mathbf{Z}_j'\mathbf{Z}_j, j \in A_0$, that is, $\rho_{n3} = \max_{j\in A_0} \|n^{-1}\mathbf{Z}_j'\mathbf{Z}_j\|_2$. By Lemma 3,

$$b_{n1} \asymp O(m_n^{1/2}),\ \rho_{n1} \asymp_p m_n^{-1},\ \rho_{n2} \asymp_p m_n^{-1} \ \text{ and } \ \rho_{n3} \asymp_p m_n^{-1}. \tag{39}$$

Lemma 5 *Under conditions (B1), (B2), (A3) and (A4),*

$$\mathrm{P}\Big(\|\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2 \geq \|\boldsymbol{\beta}_{nj}\|_2, \exists j \in A_1\Big) \to 0. \tag{40}$$

Proof of Lemma 5. Let $\mathbf{T}_{nj}$ be an $m_n \times qm_n$ matrix with the form

$$\mathbf{T}_{nj} = (\mathbf{0}_{m_n}, \ldots, \mathbf{0}_{m_n}, \mathbf{I}_{m_n}, \mathbf{0}_{m_n}, \ldots, \mathbf{0}_{m_n}),$$

where $\mathbf{O}_{m_n}$ is an $m_n \times m_n$ matrix of zeros and $\mathbf{I}_{m_n}$ is an $m_n \times m_n$ identity matrix, and $\mathbf{I}_{m_n}$ is at the jth block. By (34),

$$\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj} = n^{-1}\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}(\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n + \mathbf{Z}_{A_1}'\boldsymbol{\delta}_n - \lambda_{n2}\boldsymbol{\nu}_n).$$

By the triangle inequality,

$$\|\widehat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2 \leq n^{-1}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2 + n^{-1}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\mathbf{Z}_{A_1}'\delta_n\|_2 + n^{-1}\lambda_{n2}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\boldsymbol{\nu}_n\|_2. \quad (41)$$

Let C be a generic constant independent of n. The first term on the right-hand side

$$\begin{aligned}\max_{j\in A_1} n^{-1}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2 &\leq n^{-1}\rho_{n1}^{-1}\|\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2 \\ &= n^{-1/2}\rho_{n1}^{-1}\|n^{-1/2}\mathbf{Z}_{A_1}'\boldsymbol{\varepsilon}_n\|_2 \\ &= O_p(1)n^{-1/2}\rho_{n1}^{-1}m_n^{-1/2}(qm_n)^{1/2} \qquad (42)\end{aligned}$$

By (33), the second term

$$\begin{aligned}\max_{j\in A_1} n^{-1}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\mathbf{Z}_{A_1}'\boldsymbol{\delta}_n\|_2 &\leq \|\mathbf{C}_{A_1}^{-1}\|_2 \cdot \|n^{-1}\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1}\|_2^{1/2} \cdot \|n^{-1}\boldsymbol{\delta}_n\|_2 \\ &= O_p(1)\rho_{n1}^{-1}\rho_{n2}^{1/2}q^{1/2}m_n^{-d}. \qquad (43)\end{aligned}$$

By Lemma 4, the third term

$$\max_{j\in A_1} n^{-1}\lambda_{n2}\|\mathbf{T}_{nj}\mathbf{C}_{A_1}^{-1}\boldsymbol{\nu}_n\|_2 \leq n\lambda_{n2}\rho_{n1}^{-1}\|\boldsymbol{\nu}_n\|_2 = O_p(1)\rho_{n1}^{-1}n^{-1}\lambda_{n2}h_n. \quad (44)$$

Thus (40) follows from (39), (42), (43), (44) and condition (B2a). □

Lemma 6 *Under conditions (B1), (B2), (A3) and (A4),*

$$\mathrm{P}\Big(\|\mathbf{Z}_j'(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1})\|_2 > \lambda_{n2}w_{nj}/2, \exists j \notin A_1\Big) \to 0. \quad (45)$$

Proof of Lemma 6. By (35), we have

$$\mathbf{Z}_j'\big(\mathbf{Y} - \mathbf{Z}_{A_1}\widehat{\boldsymbol{\beta}}_{nA_1}\big) = \mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\varepsilon}_n + \mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\delta}_n + \lambda n^{-1}\mathbf{Z}_j'\mathbf{Z}_{A_1}\mathbf{C}_{A_1}^{-1}\boldsymbol{\nu}_n. \quad (46)$$

Recall $s_n = p - q$ is the number of zero components in the model. By Lemma 2,

$$\mathrm{E}\Big(\max_{j\notin A_1}\|n^{-1/2}\mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\varepsilon}_n\|_2\Big) \leq O(1)\big\{\log(s_n m_n)\big\}^{1/2}. \quad (47)$$

Since $w_{nj} = \|\widehat{\boldsymbol{\beta}}_{nj}\|^{-1} = O_p(r_n)$ for $j \notin A_1$ and by (47), for the first term on the right hand side of (46) we have

$$\begin{aligned}
&\mathrm{P}\Big(\|\mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\varepsilon}_n\|_2 > \lambda_{n2}w_{nj}/6, \exists j \notin A_1\Big) \\
&\leq \mathrm{P}\Big(\|\mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\varepsilon}_n\|_2 > C\lambda_{n2}r_n, \exists j \notin A_1\Big) + o(1) \\
&= \mathrm{P}\Big(\max_{j\notin A_1}\|n^{-1/2}\mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\varepsilon}_n\|_2 > Cn^{-1/2}\lambda_{n2}r_n\Big) + o(1) \\
&\leq O(1)\frac{n^{1/2}\{\log(s_n m_n)\}^{1/2}}{C\lambda_{n2}r_n} + o(1). \qquad (48)
\end{aligned}$$

By (33), the second term on the right hand side of (46)

$$\max_{j\notin A_1}\|\mathbf{Z}_j'\mathbf{H}_n\boldsymbol{\delta}_n\|_2 \leq n^{1/2}\max_{j\notin A_1}\|n^{-1}\mathbf{Z}_j'\mathbf{Z}_j\|_2^{1/2}\cdot\|\mathbf{H}_n\|_2\cdot\|\boldsymbol{\delta}_n\|_2 = O(1)n\rho_{n3}^{1/2}q^{1/2}m_n^{-d}. \qquad (49)$$

By Lemma 4, the third term on the right hand side of (46)

$$\begin{aligned}
\max_{j\not n A_1}\lambda_{n2}n^{-1}\|\mathbf{Z}_j\mathbf{Z}_{A_1}\mathbf{C}_{A_1}^{-1}\boldsymbol{\nu}_n\|_2 &\leq \lambda_{n2}\max_{j\in A_1}\|n^{-1/2}\mathbf{Z}_j\|_2\cdot\|n^{-1/2}\mathbf{Z}_{A_1}\mathbf{C}_{A_1}^{-1/2}\|_2\cdot\|\mathbf{C}_{A_1}^{-1/2}\|_2\cdot\|\boldsymbol{\nu}_n\|_2 \\
&= \lambda_{n2}\rho_{n3}^{1/2}\rho_{n1}^{-1/2}O_p(qb_{n1}^{-1}). \qquad (50)
\end{aligned}$$

Therefore, (45) follows from (39), (48), (49), (50) and condition (B2b). □

Proof of Theorem 4. The proof is similar to that of Theorem 2 and is omitted.

References

[1] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic J. Statist.* 169 - 194.

[2] Bühlmann, P. and Meier, L. (2008). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" by H. Zou and R. Li. To appear in the *Ann. Statist.* **36**, 1534-1541.

[3] de Boor, C. (2001). *A Practical Guide to Splines*. Revised Edition. Springer, New York.

[4] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407-499.

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

[6] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

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

[8] Greenshtein E. and Ritov Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971988.

[9] Horowitz, J.L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function, *Ann. Statist.* **32**, 2412-2443.

[10] Horowitz, J.L., Klemela, J. and Mammen, E. (2006). Optimal estimation in additive regression models. *Bernoulli* **12**, 271-298.

[11] Horowitz, J.L. and Lee, S. (2005). Nonparametric Estimation of an Additive Quantile Regression Model. *J. Am. Statist. Assoc.* **100**, 1238-1249.

[12] Huang, J., Horowitz, J. L. and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-13.

[13] Huang, J., Ma, S. and Zhang, C.-H. (2006). Adaptive lasso for sparse high-dimensional regression models. Technical Report No. 374. The University of Iowa. To appear in *Statist. Sinica.*

[14] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356-1378.

[15] Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16**, 375-90.

[16] Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, **34**, 2272-2297.

[17] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[18] Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. To appear in *Ann. Statist.*

[19] Sala-I-Martin, Xavier X. (1997). I just ran two million regressions. *The American Economic Review* **87**, 178-183.

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

[21] Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

[22] Shen, X., Wolf, D. A. and Zhou, S. (1998). Local Asymptotics for Regression Splines and Confidence Regions *Ann. Statist.*, **26**, 1760-1782.

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

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

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. Ser. B* **58** 267-288.

[26] van de Geer, S. (2006). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36**, 614-645.

[27] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York.

[28] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49-67.

[29] Zhang, C.-H. (2007). Penalized linear unbiased selection. Technical report # 2007-003. Department of Statistics, Rutgers University. *www.stat.rutgers.edu/resources/technical_reports07.html*

[30] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

[31] Zhang, H. H. and Lin, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, **16** 1021-1041.

[32] Zhao, P., Rocha, G. and Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. To appear in the *Ann. Statist.*

[33] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Res.*, **7**, 2541 - 2563

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

[35] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

[36] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with Discussion). *Ann. Statist.* **36**, 1509-1533.

Jian Huang, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242.
E-mail: jian-huang@uiowa.edu

Joel L. Horowitz, Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208.
E-mail: joel-horowitz@northwestern.edu

Fengrong Wei, Department of Mathematics, University of Iowa, Iowa City, Iowa 52242.
E-mail: fengrong-wei@uiowa.edu

	Adaptive Group Lasso			Group Lasso			Lasso		
Example	ER	IN%	CS%	ER	IN%	CS%	ER	IN%	CS%
$t = 0$	38.5	84	61	38.6	85	42	38.7	80	40
	(4.5)	(1.8)	(2.5)	(4.4)	(1.8)	(2.5)	(4.4)	(2.0)	(2.5)
$t = 1$	32.1	79	60	32.2	82	42	32.5	75.8	39
	(4.1)	(2.0)	(2.4)	(4.2)	(2.0)	(2.5)	(4.2)	(2.1)	(2.5)
$t = 0$	28.4	59	34	28.9	67	30	28.9	62	30
	(4.1)	(2.5)	(2.4)	(4.1)	(2.4)	(2.3)	(4.1)	(2.5)	(2.3)
$t = 1$	25.4	59	31	25.9	67	23	25.9	62	20
	(3.8)	(2.4)	(2.3)	(3.8)	(2.4)	(2.1)	(3.8)	(2.4)	(2.0)
$t = 0$	62.8	98	76	62.8	98	48	62.8	97	27
	(3.7)	(0.7)	(2.1)	(3.7)	(0.7)	(2.5)	(3.8)	(0.7)	(2.2)
$t = 1$	60.9	99	60	60.9	99	27	60.9	97	18
	(2.0)	(0.5)	(2.4)	(2.0)	(0.5)	(0.5)	(2.1)	(0.7)	(1.9)

Table 1: $n = 100, p = 21$. ER, model error; IN%, percentage of occasions on which the correct components are included in the selected model; CS%, percentage of occasions on which correct components are selected, averaged over 400 replications. Enclosed in parentheses are the corresponding standard errors.

	Adaptive Group Lasso			Group Lasso			Lasso		
Example	ER	IN%	CS%	ER	IN%	CS%	ER	IN%	CS%
$t = 0$	38.9	49	42	39.3	51	26	40.5	30.1	5
	(4.7)	(2.5)	(2.5)	(4.7)	(2.5)	(2.2)	(4.8)	(2.3)	(1.0)
$t = 1$	32.0	47	39	32.3	49	27	33.3	38	8
	(4.1)	(2.5)	(2.4)	(4.1)	(2.5)	(2.2)	(4.3)	(2.4)	(1.4)
$t = 0$	28.1	28	17	29.5	30	14	30.2	34	9
	(4.4)	(2.1)	(1.9)	(4.4)	(0.4)	(1.8)	(4.4)	(2.4)	(1.4)
$t = 1$	29.3	20	18	31.0	20	9	31.51	40	10
	(4.6)	(2.4)	(2.2)	(4.4)	(2.2)	(1.5)	(4.4)	(2.4)	(1.5)
$t = 0$	62.6	87	71	62.6	87	42	62.9	32	11
	(4.4)	(1.7)	(2.3)	(4.4)	(1.7)	(2.5)	(4.4)	(2.3)	(1.5)
$t = 1$	60.7	94	53	60.8	95	21	61.1	51	12
	(2.5)	(1.2)	(2.5)	(2.4)	(1.1)	(0.4)	(2.5)	(2.5)	(1.6)

Table 2: $n = 100, p = 100$. ER, model error; IN%, percentage of occasions on which the correct components are included in the selected model; CS%, percentage of occasions on which correct components are selected, averaged over 400 replications. Enclosed in parentheses are the corresponding standard errors.

Variables	AGL	GL	Lasso
log GDP per capita in 1960	$\surd$	$\surd$	$\surd$
life expectancy in 1960	$\surd$	$\surd$	$\surd$
primary school enrollment rate in 1960	$\surd$	$\surd$	$\surd$
average rate of growth of population between 1960 and 1990			$\surd$
higher education enrollment rate in 1960			$\surd$
number of years on open economy	$\surd$	$\surd$	$\surd$
number of revolutions and coups	$\surd$	$\surd$	$\surd$
political rights	$\surd$	$\surd$	$\surd$
index of civil liberties			$\surd$
absolute latitude	$\surd$	$\surd$	$\surd$
fraction of primary exports in total exports in 1970	$\surd$	$\surd$	$\surd$
urbanization rate (fraction in cities)			$\surd$
fraction Buddhist	$\surd$	$\surd$	$\surd$
fraction Catholic	$\surd$	$\surd$	$\surd$
fraction Confucianist	$\surd$	$\surd$	$\surd$
fraction Hindu			$\surd$
fraction Muslim			$\surd$
fraction Protestant	$\surd$	$\surd$	$\surd$
fraction GDP in mining	$\surd$	$\surd$	$\surd$
fraction of the population that speaks foreign language			$\surd$
fraction of the population that speaks English			$\surd$

Table 3: Variables selected by the adaptive group Lasso, group Lasso and Lasso.

	No. of variables	RSS
Adaptive Group Lasso	13	41.73
Group Lasso	13	71.28
Lasso	21	14.93

Table 4: No. of variables means how many variables are selected, RSS is the residual sum of squares.

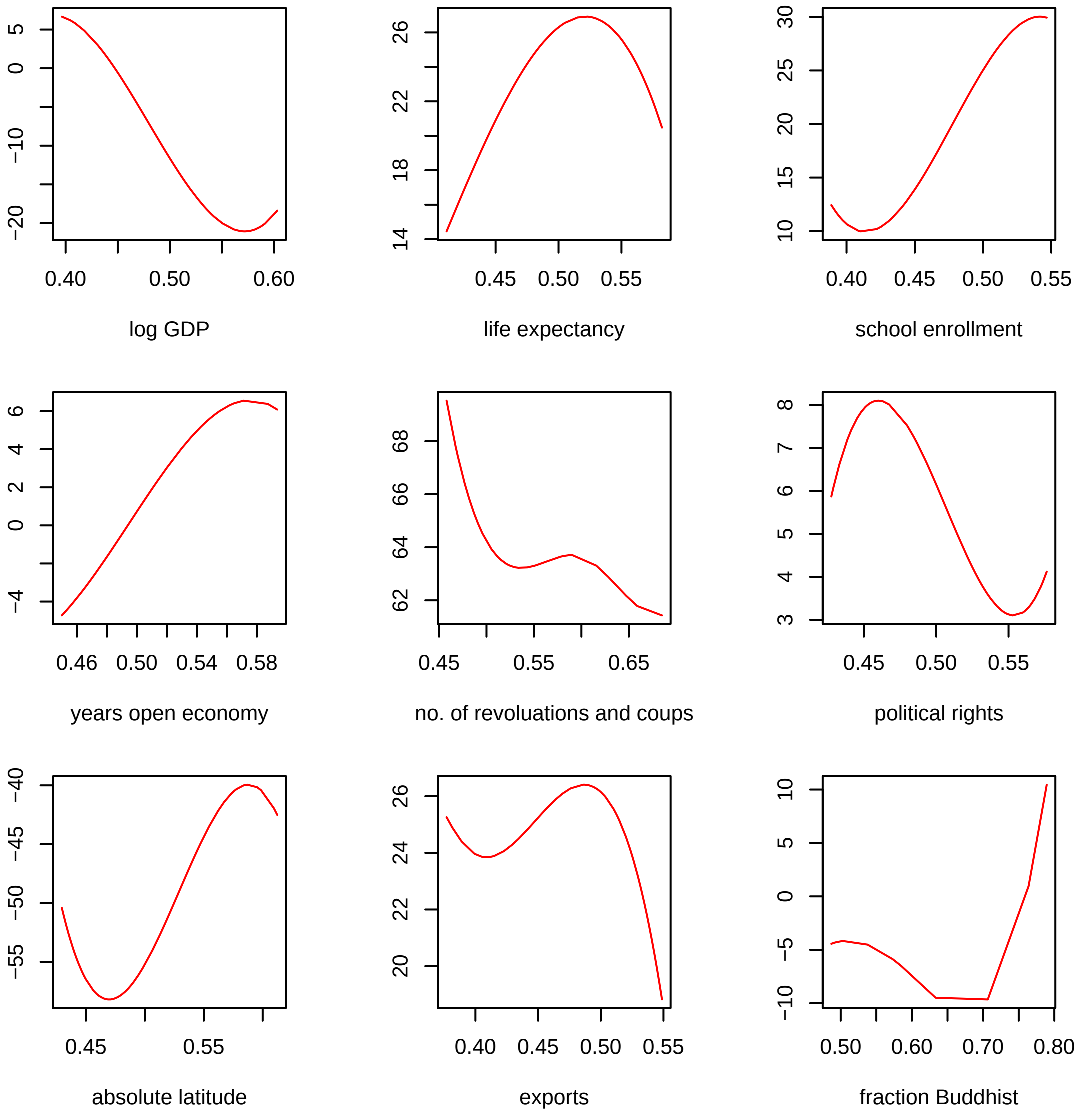

Figure 1: Plots of the estimated nonzero components by the adaptive group Lasso.

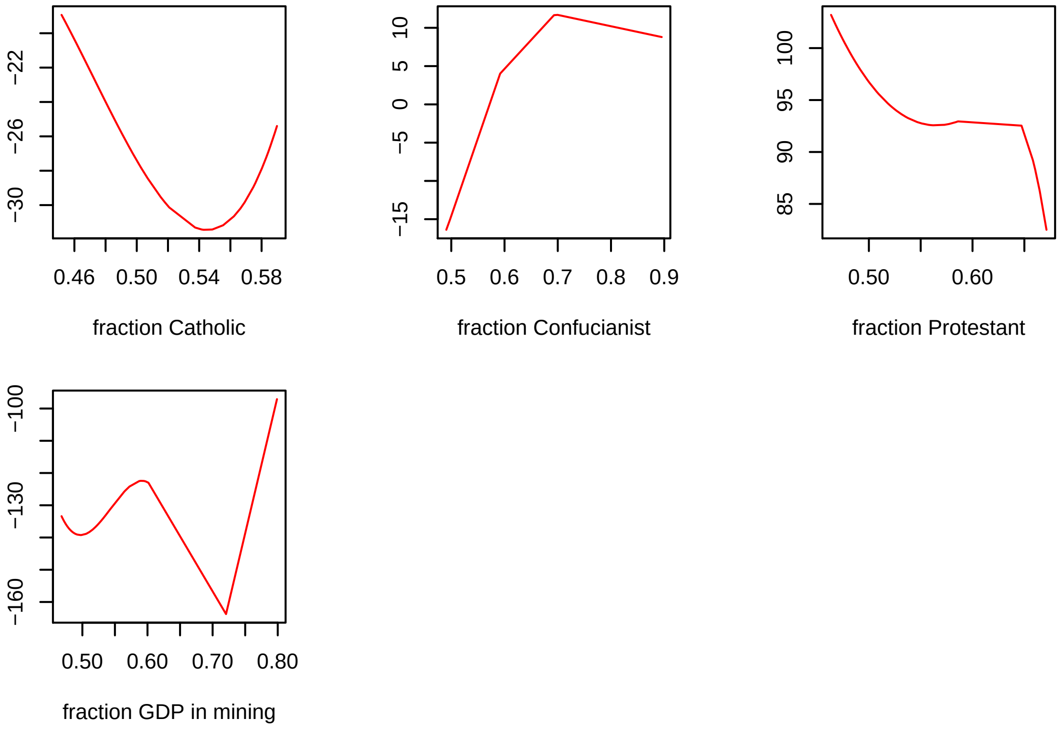

Figure 2: Plots of the estimated nonzero components by the adaptive group Lasso.