VARIABLE SELECTION IN NONPARAMETRIC ADDITIVE MODELS

Jian Huang[1], Joel L. Horowitz[2] and Fengrong Wei[3]
[1]University of Iowa, [2]Northwestern University and [3]University of West Georgia

Abstract 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. 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: 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

AOS0908-014R2 (December 7, 2009)

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, Horowitz and Ma 2008); least absolute shrinkage and selection operator or Lasso (Tibshirani 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001; Fan and Peng 2004), and the minimum concave penalty (Zhang 2009). Much progress has been made in understanding the statistical properties of these methods. In particular, many authors have studied the variable selection, estimation and prediction properties of the Lasso in high-dimensional settings. See for example, 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 of economic development, the effects of covariates on the growth of gross domestic product can be nonlinear. Similarly, there is evidence of nonlinearity in the gene expression data used in the empirical example in Section 5.

There is a large body of literature on estimation in nonparametric additive models. For example, Stone (1985, 1986) showed that additive spline estimators achieve the same optimal rate of convergence for a general fixed p as for $p = 1$. Horowitz and Mammen (2004) and Horowitz,

Klemelä, 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 do not discuss variable selection in nonparametric additive models.

Zhang et al. (2004) and Lin and Zhang (2006) have investigated the use of penalization methods in smoothing spline ANOVA with a fixed number of covariates. Zhang et al. (2004) used a Lasso-type penalty but did not investigate model-selection consistency. 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 the special case of a tensor product design, the COSSO correctly selects the non-zero additive components with high probability. Zhang and Lin (2006) considered the COSSO for nonparametric regression in exponential families.

Meier, van de Geer, and Bühlmann (2008) treat variable selection in a nonparametric additive model in which the numbers of zero and non-zero f_j's may both be larger than n. They propose a penalized least-squares estimator for variable selection and estimation. They give conditions under which, with probability approaching 1, their procedure selects a set of f_j's containing all the additive components whose distance from zero in a certain metric exceeds a specified threshold. However, they do not establish model-selection consistency of their procedure. Even asymptotically, the selected set may be larger than the set of non-zero f_j's. Moreover, they impose a compatibility condition that relates the levels and smoothness of the f_j's. The compatibility condition does not have a straightforward, intuitive interpretation and, as they point out, cannot be checked empirically. Ravikumar, Liu, Lafferty and Wasserman (2009) proposed a penalized approach for variable selection in nonparametric additive models. In their approach, the penalty is imposed on the ℓ_2 norm of the nonparametric components, as well as the mean value of the components to ensure identifiability. In their theoretical results, they require that the eigenvalues of a 'design matrix' be bounded away from zero and infinity, where the 'design matrix' is formed from the basis functions for the nonzero components. It is not clear whether this condition holds in general, especially when the number of nonzero components diverges with n. Another critical condition required in the results of Ravikumar et al. (2009) is similar to the irrepresentable condition of Zhao and Yu (2007). It is not clear for what type of basis functions this condition is satisfied. We do not require such a condition in our results on selection consistency of the adaptive group Lasso.

Several other recent papers have also considered variable selection in nonparametric models. For example, Wang, Chen and Li (2007) and Wang and Xia (2008) considered the use of group Lasso and SCAD methods for model selection and estimation in varying coefficient models with a fixed number of coefficients and covariates. Bach (2007) applies what amounts to the group Lasso to a nonparametric additive model with a fixed number of covariates. He established model selection consistency under conditions that are considerably more complicated than the ones we require for a possibly diverging number of covariates.

In this paper, we propose to use the adaptive group Lasso 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 approximating model. To achieve model 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 nonparametric components. The adaptive group Lasso is a simple generalization of the adaptive Lasso (Zou 2006) to the method of the group Lasso (Yuan and Lin 2006). However, here we apply this approach to nonparametric additive modeling.

We assume that the number of non-zero f_j's is fixed. This enables us to achieve model selection consistency under simple assumptions that are easy to interpret. We do not have to impose compatibility or irrepresentable conditions, nor do we need to assume conditions on the eigenvalues of certain matrices formed from the spline basis functions. We show that the group Lasso selects a model whose number of components is bounded with probability approaching one by a constant that is independent of the sample size. Then, using the group Lasso result as the initial estimator, the adaptive group Lasso selects the correct model with probability approaching 1 and achieves the optimal rate of convergence for nonparametric estimation of an additive model.

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 includes concluding remarks. Proofs of the results stated in Section 3 are given in the Appendix.

2 Adaptive group Lasso in nonparametric additive models

We describe a two-step approach that uses 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 $\mathrm{E}f_j(X_j) = 0, 1 \le j \le p$. Let $a = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = b$ be a partition of $[a, b]$ into K subintervals $I_{Kt} = [\xi_t, \xi_{t+1}), t = 0, \ldots, K-1$ 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 $\max_{1 \le k \le K+1} |\xi_k - \xi_{k-1}| = O(n^{-v})$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \ge 1$ consisting of functions s satisfying: (i) the restriction of s to I_{Kt} is a polynomial of degree l for $1 \le t \le K$; (ii) for $l \ge 2$ and $0 \le l' \le l-2$, 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 $\{\phi_k, 1 \le k \le m_n\}$ 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

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} \phi_k(x), \quad 1 \le j \le p. \tag{2}$$

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 $\boldsymbol{\beta}_{nj} = (\beta_{j1}, \ldots, \beta_{jm_n})'$ and $\boldsymbol{\beta}_n = (\boldsymbol{\beta}'_{n1}, \ldots, \boldsymbol{\beta}'_{np})'$. Let $w_n = (w_{n1}, \ldots, w_{np})'$ be a given vector of weights, where $0 \le w_{nj} \le \infty, 1 \le j \le p$. Consider the penalized least squares criterion

$$L_n(\mu, \boldsymbol{\beta}_n) \;=\; \sum_{i=1}^n \Big[Y_i - \mu - \sum_{j=1}^p \sum_{k=1}^{m_n} \beta_{jk} \phi_k(X_{ij}) \Big]^2 + \lambda_n \sum_{j=1}^p w_{nj} \|\boldsymbol{\beta}_{nj}\|_2, \tag{3}$$

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

$$\sum_{i=1}^n \sum_{k=1}^{m_n} \beta_{jk} \phi_k(X_{ij}) = 0, 1 \le j \le p. \tag{4}$$

These centering constraints are sample analogs of the identifying restriction $Ef_j(X_j) = 0, 1 \leq j \leq p$. We can convert (3)-(4) to an unconstrained optimization problem by centering the response and the basis functions. Let

$$\bar{\phi}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \phi_k(X_{ij}), \quad \psi_{jk}(x) = \phi_k(x) - \bar{\phi}_{jk}. \tag{5}$$

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

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

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

$$L_n(\boldsymbol{\beta}_n; \lambda) = \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}_n\|_2^2 + \lambda_n \sum_{j=1}^{p} w_{nj} \|\boldsymbol{\beta}_{nj}\|_2. \tag{6}$$

Here we have dropped μ in the argument of L_n. With the centering, $\widehat{\mu} = \overline{Y}$. Then minimizing (3) subject to (4) is equivalent to minimizing (6) with respect to $\boldsymbol{\beta}_n$, but the centering constraints are not needed for (6).

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 (6) 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

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

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:

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

In (1), without loss of generality, suppose that the first q components are nonzero, that is, $f_j(x) \ne 0, 1 \le j \le q$, but $f_j(x) \equiv 0, q + 1 \le j \le p$. 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.

We make the following assumptions.

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

(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 $P(|\varepsilon_i| > x) \le K \exp(-Cx^2), i = 1, \ldots, n$, for all $x \ge 0$ 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) The covariate vector X has a continuous density and there exist constants C_1 and C_2 such that the density function g_j of X_j satisfies $0 < C_1 \le g_j(x) \le C_2 < \infty$ on $[a,\ b]$ for every $1 \le j \le p$.

We note that (A1), (A3), and (A4) are standard conditions for nonparametric additive models. They would be needed to estimate the nonzero additive components at the optimal ℓ_2 rate of

convergence on $[a, b]$, even if q were fixed and known. Only (A2) strengthens the assumptions needed for nonparametric estimation of a nonparametric additive model. While condition (A1) is reasonable in most 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. The only technical reason that we assume this condition is related to Lemma 3 given in the appendix, which is concerned with the properties of the smallest and largest eigenvalues of the 'design matrix' formed from the spline basis functions. If this lemma can be extended to the case of a divergent number of components, then (A1) can be relaxed. However, it is clear that there needs to be restriction on the number of nonzero components to ensure model identification.

3.1 Estimation consistency of the group Lasso

In this section, we consider the selection and estimation properties of the group Lasso estimator. Define $\widetilde{A}_1 = \{j : \|\widetilde{\boldsymbol{\beta}}_{nj}\|_2 \neq 0, 1 \leq j \leq p\}$. Let $|A|$ denote the cardinality of any set $A \subseteq \{1, \ldots, p\}$.

Theorem 1 *Suppose that (A1) to (A4) hold and* $\lambda_{n1} \geq C\sqrt{n \log(pm_n)}$ *for a sufficiently large constant* C.

(i) With probability converging to 1, $|\widetilde{A}_1| \leq M_1|A_1| = M_1 q$ *for a finite constant* $M_1 > 1$.

(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* $\boldsymbol{\beta}_{nj}, 1 \leq j \leq q$, *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, with probability approaching 1, 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 $\widetilde{f}_{nj}(x) = \sum_{j=1}^{m_n} \widetilde{\beta}_{jk}\psi(x), 1 \leq j \leq p$. The following theorem is a consequence of Theorem 1.

Theorem 2 *Suppose that (A1) to (A4) hold and that* $\lambda_{n1} \geq C\sqrt{n\log(pm_n)}$ *for a sufficiently large constant* C*. Then,*

(i) Let $\widetilde{A}_f = \{j : \|\widetilde{f}_{nj}\|_2 > 0, 1 \leq j \leq p\}$*. There is a constant* $M_1 > 1$ *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.

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.

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 zero components can be as large as $\exp(o(n^{2d/(2d+1)}))$. For example, if each f_j has continuous second derivative ($d = 2$), then it is $\exp(o(n^{4/5}))$, 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. We make the following assumptions.

(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

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

We state condition (B1) for a general initial estimator, to highlight the point that the availability of an r_n-consistent estimator at zero is crucial for the adaptive group Lasso to be selection consistent. In other words, any initial estimator satisfying (B1) will ensure that the adaptive group Lasso (based on this initial estimator) is selection consistent, provided that certain regularity conditions are satisfied. We note that it follows immediately from Theorem 1 that the group Lasso estimator satisfies (B1). We will come back to this point below.

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 $\mathrm{sgn}_0(\|\widehat{\boldsymbol{\beta}}_{nj}\|) = \mathrm{sgn}_0(\|\boldsymbol{\beta}_{nj}\|), 1 \leq j \leq p$, where $\mathrm{sgn}_0(|x|) = 1$ if $|x| > 0$ and $= 0$ if $|x| = 0$.

Theorem 3 *Suppose that conditions (B1), (B2) and (A1)-(A4) hold. Then*

(i)

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

(ii)

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

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 (A1)-(A4) hold. Then*

(i)

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

(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 distinguish nonzero components from zero components. Part (ii) gives an upper bound on the rate of convergence of the estimator.

We now apply the above results to our proposed procedure described in Section 2, in which we first obtain the the group Lasso estimator and then use it as the initial estimator in the adaptive group Lasso.

By Theorem 1, if $\lambda_{n1} \asymp \sqrt{n\log(pm_n)}$ and $m_n \asymp n^{1/(2d+1)}$ for $d \geq 1$, then the group Lasso estimator satisfies (B1) with $r_n \asymp n^{d/(2d+1)}/\sqrt{\log(pm_n)}$. In this case, (B2) simplifies to

$$\frac{\lambda_{n2}}{n^{(8d+3)/(8d+4)}} = o(1) \quad \text{and} \quad \frac{n^{1/(4d+2)}\log^{1/2}(pm_n)}{\lambda_{n2}} = o(1). \tag{7}$$

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 $\lambda_{n1} \asymp \sqrt{n\log(pm_n)}$ and $m_n \asymp n^{1/(2d+1)}$ be the initial estimator in the adaptive group Lasso. Suppose that the conditions of Theorem 1 hold. If $\lambda_{n2} \leq O(n^{1/2})$ and satisfies (7), then the adaptive group Lasso consistently selects the nonzero components in (1), that is, part (i) of Theorem 4 holds. In addition,*

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

This corollary follows directly from Theorems 1 and 4. The largest λ_{n2} allowed is $\lambda_{n2} = O(n^{1/2})$. With this λ_{n2}, the first equation in (6) is satisfied. Substitute it into the second equation in (6), we obtain $p = \exp(o(n^{2d/(2d+1)}))$, which is the largest p permitted and can be larger than n. Thus, under the conditions of this corollary, our proposed adaptive group Lasso estimator using the group Lasso as the initial estimator is selection consistent and achieves optimal rate of convergence even when p is larger than n. Following model selection, oracle-efficient, asymptotically normal estimators of the non-zero components can be obtained by using existing methods.

4 Simulation studies

We use simulation to evaluate the performance of the adaptive group Lasso with regard to variable selection. The generating model is

$$y_i = f(x_i) + \varepsilon_i \equiv \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, i = 1, \cdots, n. \tag{8}$$

Since p can be larger than n, we consider two ways to select the penalty parameter, the BIC (Schwarz 1978) and the EBIC (Chen and Chen 2008, 2009). The BIC is defined as

$$BIC(\lambda) = \log(\text{RSS}_\lambda) + df_\lambda \cdot \frac{\log n}{n}.$$

Here RSS_λ is the residual sum of squares for a given λ, and the degrees of freedom $df_\lambda = \hat{q}_\lambda m_n$, where $\hat{q}_\lambda$ is the number of nonzero estimated components for the given λ. The EBIC is defined as

$$EBIC(\lambda) = \log(RSS_\lambda) + df_\lambda \cdot \frac{\log n}{n} + \nu \cdot df_\lambda \cdot \frac{\log p}{n},$$

where $0 \leq \nu \leq 1$ is a constant. We use $\nu = 0.5$.

We have also considered two other possible ways of defining df: (a) using the trace of a linear smoother based on a quadratic approximation; (b) using the number of estimated nonzero components. We have decided to use the definition given above based on the results from our simulations. We note that the df for the group Lasso of Yuan and Lin (2006) requires an initial (least squares) estimator, which is not available when $p > n$. Thus their df is not applicable to our problem.

In our simulation example, we compare the adaptive group Lasso with the group Lasso and ordinary Lasso. Here the ordinary Lasso estimator is defined as the value that minimizes

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

This simple application of the Lasso does not take into account the grouping structure in the spline expansions of the components. The group Lasso and the adaptive group Lasso estimates are computed using the algorithm proposed by Yuan and Lin (2006). The ordinary Lasso estimates are computed using the Lars algorithms (Efron et al. 2004). The group Lasso is used as the initial estimate for the adaptive group Lasso.

We also compare the results from the nonparametric additive modeling with those from the

standard linear regression model with Lasso. We note that this is not a fair comparison because the generating model is highly nonlinear. Our purpose is to illustrate that it is necessary to use nonparametric models when the underlying model deviates substantially from linear models in the context of variable selection with high-dimensional data and that model misspecification can lead to bad selection results.

Example 1. We generate data from the model

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

where $f_1(t) = 5t$, $f_2(t) = 3(2t-1)^2, f_3(t) = 4\sin(2\pi t)/(2-\sin(2\pi t))$, $f_4(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3)$, and $f_5(t) = \cdots = f_p(t) = 0$. Thus the number of nonzero functions is $q = 4$. This generating model is the same as Example 1 of Lin and Zhang (2006). However, here we use this model in high-dimensional settings. We consider the cases where $p = 1000$ and three different sample sizes: $n = 50, 100$ and 200. We use the cubic B-spline with six evenly distributed knots for all the functions f_k. The number of replications in all the simulations is 400.

The covariates are simulated as follows. First, we generate $w_{i1}, \cdots, w_{ip}, u_i, u_i^{'}, v_i$ 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 the parameter t controls the amount of correlation among predictors. We have $Corr(x_{ik}, x_{ij}) = t^2/(1+t^2)$, $1 \le j \le 4$, $1 \le k \le 4$ and $Corr(x_{ik}, x_{ij}) = t^2/(1+t^2)$, $4 \le j \le p$, $4 \le k \le p$, but the covariates of the nonzero components and zero components are independent. We consider $t = 0, 1$ in our simulation. The signal to noise ratio is defined to be $sd(f)/sd(\epsilon)$. The error term is chosen to be $\epsilon_i \sim N(0, 1.27^2)$ to give a signal-to-noise ratio (SNR) 3.11 : 1. This value is the same as the estimated SNR in the real data example below, which is the square root of the ratio of the sum of estimated components squared divided by the sum of residual squared.

The results of 400 Monte Carlo replications are summarized in Table 1. The columns are the mean number of variables selected (NV), model error (ER), the percentage of replications in which all the correct additive components are included in the selected model (IN), and the percentage of replications in which precisely the correct components are selected (CS). The corresponding standard errors are in parentheses. The model error is computed as the average of $n^{-1}\sum_{i=1}^{n}[\hat{f}(x_i) - f(x_i)]^2$ over the 400 Monte Carlo replications, where f is the true conditional mean function.

Table 1 shows that the adaptive group Lasso selects all the non-zero components (IN) and selects exactly the correct model (CS) more frequently than the other methods do. For example,

with the BIC and $n = 200$, the percentage of correct selections (CS) by the adaptive group Lasso ranges from 65.25% to 81%, which is much higher than the ranges 30-57.75% for the group Lasso and 12-15.75% for the ordinary Lasso. The adaptive group Lasso and group Lasso perform better than the ordinary Lasso in all of the experiments, which illustrates the importance of taking account of the group structure of the coefficients of the spline expansion. Correlation among covariates increases the difficulty of component selection, so it is not surprising that all methods perform better with independent covariates than with correlated ones. The percentage of correct selections increases as the sample size increases. The linear model with Lasso never selects the correct model. This illustrates the poor results that can be produced by a linear model when the true conditional mean function is nonlinear.

Table 1 also shows that the model error (ME) of the group Lasso is only slightly larger than that of the adaptive group Lasso. The models selected by the group Lasso nest and, therefore, have more estimated coefficients than the models selected by the adaptive group Lasso. Therefore, the group Lasso estimators of the conditional mean function have a larger variance and larger ME. The differences between the MEs of the two methods are small, however, because as can be seen from the NV column, the models selected by the group Lasso in our experiments have only slightly more estimated coefficients than the models selected by the adaptive group Lasso.

Example 2. We now compare the adaptive group Lasso with the COSSO (Lin and Zhang 2006). This comparison is suggested to us by the Associate Editor. Because the COSSO algorithm only works for the case when p is smaller than n, we use the same set-up as in Example 1 of Lin and Zhang (2006). In this example, the generating model is as in (8) with 4 nonzero components. Let $X_j = (W_j + tU)/(1 + t)$, $j = 1, \cdots, p$, where $W_1, \cdots, W_p$ and U are i.i.d. from $N(0, 1)$, truncated to the interval $[0, 1]$. Therefore $\text{corr}(X_j, X_k) = t^2/(1 + t^2)$ for $j \neq k$. The random error term $\epsilon \sim N(0, 1.32^2)$. The SNR is 3:1. We consider three different sample sizes $n = 50, 100$ or 200 and three different number of predictors $p = 10, 20$ or 50. The COSSO estimator is computed using the Matlab software which is publicly available at *http://www4.stat.ncsu.edu/∼hzhang/cosso.html*.

The COSSO procedure uses either generalized cross validation or 5-fold cross validation. Based the simulation results of Lin and Zhang (2006) and our own simulations, the COSSO with 5-fold cross validation has better selection performance. Thus we compare the adaptive group Lasso with BIC or EBIC with the COSSO with 5-fold cross validation. The results are given in Table 2. For independent predictors, when $n = 200$ and $p = 10, 20$ or 50, the adaptive group Lasso and COSSO have similar performance in terms of selection accuracy and model error. However, for smaller n and larger p, the adaptive group Lasso does significantly better. For example, for $n = 100$ and $p = 50$, the percentage of correct selection for the adaptive group Lasso is 81-83%,

but it is only 11% for the COSSO. The model error of the adaptive group Lasso is similar to or smaller than that of the COSSO. In several experiments, the model error of the COSSO is 2 to more than 7 times larger than that of the adaptive group Lasso. It is interesting to note that when $n = 50$ and $p = 20$ or 50, the adaptive group Lasso still does a descent job in selecting the correct model, but the COSSO does poorly in these two cases. In particular, for $n = 50$ and $p = 50$, the COSSO did not select the exact correct model in all the simulation runs. For dependent predictors, the comparison is even mode favorable to the adaptive group Lasso, which performs significantly better than COSSO in terms of both model error and selection accuracy in all the cases.

5 Data example

We use the data set reported in Scheetz et al. (2006) to illustrate the application of the proposed method in high-dimensional settings. For this data set, 120 twelve-week-old male rats were selected for tissue harvesting from the eyes and for microarray analysis. The microarrays used to analyze the RNA from the eyes of these animals contain over 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multi-chip averaging method (Irizzary et al. 2003) method to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale.

We are interested in finding the genes that are related to the gene TRIM32. This gene was recently found to cause Bardet-Biedl syndrome (Chiang et al. (2006)), which is a genetically heterogeneous disease of multiple organ systems including the retina. Although over 30,000 probe sets are represented on the Rat Genome 230 2.0 Array, many of them are not expressed in the eye tissue and initial screening using correlation shows that most probe sets have very low correlation with TRIM32. In addition, we are expecting only a small number of genes to be related to TRIM32. Therefore, we use 500 probe sets that are expressed in the eye and have highest marginal correlation in the analysis. Thus the sample size is $n = 120$ (i.e., there are 120 arrays from 120 rats) and $p = 500$. It is expected that only a few genes are related to TRIM32. Therefore this is a sparse, high-dimensional regression problem.

We use the nonparametric additive model to model the relation between the expression of TRIM32 and those of the 500 genes. We estimate model (1) using the ordinary Lasso, group Lasso, and adaptive group Lasso for the nonparametric additive model. To compare the results of the nonparametric additive model with that of the linear regression model, we also analyzed the data using the linear regression model with Lasso. We scale the covariates so that their values are

between 0 and 1 and use cubic splines with six evenly distributed knots to estimate the additive components. The penalty parameters in all the methods are chosen using the BIC or EBIC as in the simulation study. Table 3 lists the probes selected by the group Lasso and the adaptive group Lasso, indicated by the check signs. Table 4 shows the number of variables, the residual sums of squares obtained with each estimation method. For the ordinary Lasso with the spline expansion, a variable is considered to be selected if any of the estimated coefficients of the spline approximation to its additive component are non-zero. Depending on whether BIC or EBIC is used, the group Lasso selects 16-17 variables, the adaptive group Lasso selects 15 variables and the ordinary Lasso with the spline expansion selects 94-97 variables, the linear model selects 8-14 variables. Table 4 shows that the adaptive group Lasso does better than the other methods in terms of residual sum of squares (RSS). We have also examined the plots (not shown) of the estimated additive components obtained with the group Lasso and the adaptive group Lasso, respectively. Most are highly nonlinear, confirming the need for taking into account nonlinearity.

In order to evaluate the performance of the methods, we use cross-validation and compare the prediction mean square errors (PEs). We randomly partition the data into 6 subsets, each set consisting of 20 observations. We then fit the model with 5 subsets as training set and calculate the PE for the remaining set which we consider as test set. We repeat this process 6 times, considering one of the 6 subsets as test set every time. We compute the average of the numbers of probes selected and the prediction errors of these 6 calculations. Then we replicate this process 400 times (this is suggested to us by the Associate Editor). Table 5 gives the average values over 400 replications. The adaptive group Lasso has smaller average prediction error than the group Lasso, the ordinary Lasso and the linear regression with Lasso. The ordinary Lasso selects far more probe sets than the other approaches, but this does not lead to better prediction performance. Therefore, in this example, the adaptive group Lasso provides the investigator a more targeted list of probe sets, which can serve as a starting point for further study.

It is of interest to compare the selection results from the adaptive group Lasso and the linear regression model with Lasso. The adaptive group Lasso and the linear model with Lasso select different sets of genes. When the penalty parameter is chosen with the BIC, the adaptive group Lasso selects 5 genes that are not selected by the linear model with Lasso. In addition, the linear model with Lasso selects 5 genes that are not selected by the adaptive group Lasso. When the penalty parameter is selected with the EBIC, the adaptive group Lasso selects 10 genes that are not selected by the linear model with Lasso. The estimated effects of many of the genes are nonlinear, and the Monte Carlo results of Section 4 show that the performance of the linear model with Lasso can be very poor in the presence of nonlinearity. Therefore, we interpret the differences between

the gene selections of the adaptive group Lasso and the linear model with Lasso as evidence that the selections produced by the linear model are misleading.

6 Concluding remarks

In this paper, we propose to use the adaptive group Lasso for variable selection in nonparametric additive models in sparse, high-dimensional settings. A key requirement for the adaptive group Lasso to be selection consistent is that the initial estimator is estimation consistent and selects all the important components with high probability. In low-dimensional settings, finding an initial consistent estimator is relatively easy and can be achieved by many well established approaches such as the additive spline estimators. However, in high-dimensional settings, finding an initial consistent estimator is difficult. Under the conditions stated in Theorem 1, the group Lasso is shown to be consistent and selects all the important components. Thus the group Lasso can be used as the initial estimator in the adaptive Lasso to achieve selection consistency. Following model selection, oracle-efficient, asymptotically normal estimators of the non-zero components can be obtained by using existing methods. Our simulation results indicate that our procedure works well for variable selection in the models considered. Therefore, the adaptive group Lasso is a useful approach for variable selection and estimation in sparse, high-dimensional nonparametric additive models.

Our theoretical results are concerned with a fixed sequence of penalty parameters, which are not applicable to the case where the penalty parameters are selected based on data driven procedures such as the BIC. This is an important and challenging problem that deserves further investigation, but is beyond the scope of this paper. We have only considered 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 authors wish to thank the Editor, Associate Editor and two anonymous referees for their helpful comments. Jian Huang is supported in part by NIH grant CA120988 and NSF grant DMS 0805670. Joel L. Horowitz was supported in part by NSF Grant grant SES-0817552.

7 Appendix: 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 (5).

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}. For any $\varepsilon > 0$, the bracketing number $N_{[]}(\varepsilon, \mathcal{S}_{nj}^0, L_2(P))$ of $\mathcal{S}_{nj}^0$ satisfies $\log N_{[]}(\varepsilon, \mathcal{S}_{nj}^0, L_2(P)) \le c_1 m_n \log(1/\varepsilon)$ for some constant $c_1 > 0$ (Shen and Wong 1994, page 597). Thus by the maximal inequality, see e.g. Van der Vaart (1998, page 288), $(P_n - P)f_n^* = O_p(n^{-1/2}m_n^{1/2})$. 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*

$$T_{jk} = n^{-1/2} m_n^{1/2} \sum_{i=1}^n \psi_k(X_{ij})\varepsilon_i, 1 \le j \le p, 1 \le k \le m_n,$$

and $T_n = \max_{1\le j\le p, 1\le k\le m_n} |T_{jk}|$. 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}'s are subgaussian. 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 A.1 of Van de Geer (2008), (10) and (11) imply

$$\mathrm{E}\Big(\max_{1\le j\le p, 1\le k\le m_n} \Big| \sum_{i=1}^n \{\psi_k^2(X_{ij}) - \mathrm{E}\psi_k^2(X_{ij})\}\Big|\Big) \le \sqrt{2C_2 m_n^{-1} n \log(pm_n)} + 4\log(2pm_n).$$

Therefore, by (12) and the triangle inequality,

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

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

$$\mathrm{E}s_n \le \Big(\sqrt{2C_2 m_n^{-1} n \log(pm_n)} + 4\log(2pm_n) + C_2 n m_n^{-1}\Big)^{1/2}. \tag{13}$$

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

Denote

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

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 $\mathbf{C} = \mathbf{Z}'\mathbf{Z}/n$. 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,*

$$c_1 h_n \le \rho_{\min}(\mathbf{C}_A) \le \rho_{\max}(\mathbf{C}_A) \le c_2 h_n,$$

where c_1 and c_2 are two positive constants.

Proof of Lemma 3. Without loss of generality, suppose $A = \{1, \ldots, k\}$. Then $\mathbf{Z}_A = (\mathbf{Z}_1, \ldots, \mathbf{Z}_q)$. Let $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_q')'$, 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,

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

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

$$\begin{aligned}\frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}\|_2^2}+\cdots+\frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}\|_2^2} &= \frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}_1\|_2^2}\frac{\|\mathbf{b}_1\|_2^2}{\|\mathbf{b}\|_2^2}+\cdots+\frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}_q\|_2^2}\frac{\|\mathbf{b}_q\|_2^2}{\|\mathbf{b}\|_2^2}\\ &\geq \rho_{\min}(\mathbf{C}_1)\frac{\|\mathbf{b}_1\|_2^2}{\|\mathbf{b}\|_2^2}+\cdots+\rho_{\min}(\mathbf{C}_q)\frac{\|\mathbf{b}_q\|_2^2}{\|\mathbf{b}\|_2^2}\\ &\geq c_4h.\end{aligned}$$

Similarly,

$$\frac{\mathbf{b}_1'\mathbf{C}_1\mathbf{b}_1}{\|\mathbf{b}\|_2^2}+\cdots+\frac{\mathbf{b}_q'\mathbf{C}_q\mathbf{b}_q}{\|\mathbf{b}\|_2^2}\leq c_5h.$$

Thus we have

$$c_3^2c_4h\leq\frac{\mathbf{b}'\mathbf{C}_A\mathbf{b}}{\mathbf{b}'\mathbf{b}}\leq 2c_5h.$$

The lemma follows. □

Proof of Theorem 1. The proof of parts (i) and (ii) essentially follows the proof of Theorem 2.1 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

$$\|\boldsymbol{\delta}_n\|_2\leq C_1\sqrt{nqm_n^{-2d}}=C_1qn^{1/(4d+2)},$$

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

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

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 $N(A)=q_1=m\geq 0$, $S_A=(S_{A_1}^{'},\cdots,S_{A_m}^{'})'$, $S_{A_k}=\lambda\sqrt{d_{A_k}}U_{A_k}$ and $\|U_{A_k}\|_2=1$.

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

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

and

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

where $t_0\geq 0$.

As in the proof of Theorem 2.1 of Wei and Huang (2008),

$$(\mathbf{Z}, \boldsymbol{\varepsilon}_n) \in \Omega_q \Rightarrow |\widetilde{A}_1| \leq M_1 q,$$

for a constant $M_1 > 1$. By the triangle and Cauchy-Schwarz inequalities,

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

In the proof of Theorem 2.1 of Wei and Huang (2008), it is shown that

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

Since

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

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

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

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 $\widetilde{\boldsymbol{\beta}}_n \equiv (\widetilde{\boldsymbol{\beta}}_{n1}', \ldots, \widetilde{\boldsymbol{\beta}}_{np}')'$,

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

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

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

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

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

Now

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

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

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

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,

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

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

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

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

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

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 $\widetilde{f}_j, 1 \le j \le p$, 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)),

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

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 $\|\mathbf{Hx}\| \le \|\mathbf{H}\|\|\mathbf{x}\|$ for a column vector $\mathbf{x}$ whose dimension is the same as the number of the columns of $\mathbf{H}$.

Denote $\boldsymbol{\beta}_{nA_1} = (\boldsymbol{\beta}_{nj}', j \in A_1)'$, $\widehat{\boldsymbol{\beta}}_{nA_1} = (\widehat{\boldsymbol{\beta}}_{nj}', j \in A_1)'$, and $\mathbf{Z}_{A_1} = (\mathbf{Z}_j, j \in A_1)$. Define $\mathbf{C}_{A_1} = n^{-1}\mathbf{Z}_{A_1}'\mathbf{Z}_{A_1}$. Let ρ_{n1} and ρ_{n2} be the smallest and largest eigenvalues of $\mathbf{C}_{A_1}$, respectively.

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

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

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,

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

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,

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

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

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

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

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

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),

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

and $\sum_{j\in A_1}\|\boldsymbol{\beta}_{nj}\|^{-2} \le qb_{n1}^{-2}$. 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),*

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

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,

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

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

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

By (33), the second term

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

By Lemma 4, the third term

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \max_{j \notin A_1} \lambda_{n2} n^{-1} \|\mathbf{Z}_j \mathbf{Z}_{A_1} \mathbf{C}_{A_1}^{-1} \boldsymbol{\nu}_n\|_2 &\leq \lambda_{n2} \max_{j \in A_1} \|n^{-1/2} \mathbf{Z}_j\|_2 \cdot \|n^{-1/2} \mathbf{Z}_{A_1} \mathbf{C}_{A_1}^{-1/2}\|_2 \cdot \|\mathbf{C}_{A_1}^{-1/2}\|_2 \cdot \|\boldsymbol{\nu}_n\|_2 \\ &= \lambda_{n2} \rho_{n3}^{1/2} \rho_{n1}^{-1/2} O_p(q b_{n1}^{-1}). \end{aligned} \tag{50}$$

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] Bach, F. R. (2007). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9, 1179-1225.

[2] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statististic*, 169 - 194.

[3] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model space. *Biometrika*, 95, 759-771

[4] Chen, J. and Chen, Z. (2009). Extended BIC for small-n-large-P sparse GLM. Available from *http://www.stat.nus.sg/∼ stachenz/ChenChen.pdf*.

[5] Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M., and Sheffield, V. C. (2006). Homozygosity mapping with SNP arrays identifies a novel Gene for Bardet-Biedl Syndrome (BBS10). *Proceedings of the National Academy of Sciences*, 103, 6287-6292.

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

[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with Discussion). *Annals of Statistics*, 32, 407-499.

[8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96, 1348-1360.

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

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

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

[12] Horowitz, J.L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of American Statistical Association*, 100, 1238-1249.

[13] Huang, J., Horowitz, J. L. and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36, 587-613.

[14] Huang, J., Ma, S. and Zhang, C.-H.(2008). Adaptive Lasso for high-dimensional regression models. *Statistica Sinica*, 18, 1603-1618.

[15] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D,, Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249-264.

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

[17] Meier, L., van de Geer, S. and Bühlmann (2009). High-dimensional additive modeling. *Annals of Statistics*, 37, 3779-3821.

[18] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34, 1436-1462.

[19] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37, 246-270.

[20] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society, Ser. B*, 71, 1009-1030.

[21] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103, 14429-14434.

[22] Schwarz, G.(1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

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

[24] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13, 689-705.

[25] Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, 14, 590-606.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society. Ser. B*, 58, 267-288.

[27] van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36, 614-645.

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

[29] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23, 1486-1494.

[30] Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association.* 104, 747-757.

[31] Wei, F. and Huang, J. (2008). Consistent group selection in high-dimensional linear regression. *Technical report #387, Department of Statistics and Actuarial Science, University of Iowa.* Available from *http://www.stat.uiowa.edu/techrep/tr387.pdf.*

[32] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society. Ser. B*, 68, 49-67.

[33] Zhang, C.-H. (2007). Nearly unbiased variable selection under minimax concave penalty. To appear in the *Annals of Statistics.*

[34] Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of American Statistical Association*, 99, 659-672.

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

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

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

[38] Zhou, S., Shen, X. and Wolf, D. A. (1998). Local asymptotics for regression splines and confidence regions *Annals of Statistics*, 26, 1760-1782.

[39] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of American Statistical Association*, 101, 1418-1429.

Jian Huang,
Department of Statistics and Actuarial Science, 241 SH
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 West Georgia
Carrollton, Georgia 30118
E-mail: fwei@westga.edu

		Adaptive group Lasso				Group Lasso				Ordinary Lasso				Linear mode with Lasso			
		NV	ME	IN	CS	NV	ME	IN	CS	NV	ME	IN	CS	NV	ME	IN	CS
		Independent predistors															
$n = 200$	BIC	4.15	26.72	90.00	80.00	4.20	27.54	90.00	58.25	9.73	28.44	95.00	18.00	3.35	31.89	0.00	0.00
		(0.43)	(4.13)	(0.30)	(0.41)	(0.43)	(4.45)	(0.30)	(0.54)	(6.72)	(5.55)	(0.22)	(0.40)	(1.75)	(5.65)	(0.00)	(0.00)
	EBIC	4.09	26.64	92.00	81.75	4.18	27.40	92.00	60.00	9.58	28.15	95.00	32.50	3.30	32.08	0.00	0.00
		(0.38)	(4.06)	(0.24)	(0.39)	(0.40)	(4.33)	(0.24)	(0.50)	(6.81)	(5.25)	(0.22)	(0.47)	(1.86)	(5.69)	(0.00)	(0.00)
$n = 100$	BIC	4.73	28.26	85.00	70.00	5.03	29.07	85.00	35.00	17.25	29.50	82.50	12.00	6.35	31.57	5.00	0.00
		(1.18)	(5.71)	(0.36)	(0.46)	(1.22)	(6.01)	(0.36)	(0.48)	(8.72)	(5.89)	(0.38)	(0.44)	(2.91)	(7.22)	(0.22)	(0.00)
	EBIC	4.62	28.07	84.25	74.00	4.90	28.87	84.25	38.00	15.93	29.35	84.00	27.75	5.90	31.53	5.00	0.00
		(0.89)	(5.02)	(0.36)	(0.42)	(1.20)	(5.72)	(0.36)	(0.50)	(9.06)	(5.25)	(0.36)	(0.45)	(2.97)	(6.40)	(0.22)	(0.00)
$n = 50$	BIC	4.75	28.86	80.00	65.00	5.12	29.97	80.00	32.00	18.53	30.05	75.00	11.00	12.53	32.52	22.50	0.00
		(1.22)	(5.72)	(0.41)	(0.48)	(1.29)	(6.15)	(0.41)	(0.48)	(12.67)	(6.26)	(0.41)	(0.31)	(3.80)	(8.37)	(0.43)	(0.00)
	EBIC	4.69	28.94	78.00	65.00	5.01	29.82	78.00	36.00	17.27	30.50	77.50	26.00	10.33	31.64	20.00	0.00
		(1.98)	(6.48)	(0.40)	(0.48)	(1.21)	(6.11)	(0.40)	(0.49)	(15.32)	(7.89)	(0.39)	(0.44)	(3.19)	(8.17)	(0.41)	(0.00)
		Correlated predictors															
$n = 200$	BIC	3.20	27.76	66.00	60.00	3.85	28.12	66.00	30.00	9.13	28.80	56.00	11.00	1.08	32.18	0.00	0.00
		(1.27)	(4.74)	(0.46)	(0.50)	(1.49)	(4.76)	(0.46)	(0.46)	(7.02)	(5.36)	(0.51)	(0.31)	(0.33)	(8.99)	(0.00)	(0.00)
	EBIC	3.23	27.60	68.00	63.00	3.92	27.85	68.00	31.00	9.24	28.22	58.00	13.75	1.30	32.00	0.00	0.00
		(1.24)	(4.34)	(0.45)	(0.49)	(1.68)	(4.50)	(0.45)	(0.48)	(7.18)	(5.30)	(0.52)	(0.44)	(1.60)	(8.92)	(0.00)	(0.00)
$n = 100$	BIC	2.88	27.88	60.00	56.00	3.28	28.33	60.00	22.00	8.80	28.97	52.00	8.00	1.00	32.24	0.00	0.00
		(1.91)	(4.88)	(0.50)	(0.56)	(1.96)	(4.92)	(0.50)	(0.42)	(10.22)	(5.45)	(0.44)	(0.26)	(0.00)	(9.20)	(0.00)	(0.00)
	EBIC	3.04	27.78	61.75	58.00	3.44	28.16	61.75	24.00	9.06	28.55	54.00	10.00	1.00	32.09	0.00	0.00
		(1.46)	(4.85)	(0.49)	(0.54)	(1.52)	(4.90)	(0.49)	(0.43)	(11.24)	(5.42)	(0.46)	(0.28)	(0.00)	(8.98)	(0.00)	(0.00)
$n = 50$	BIC	2.50	28.36	48.50	38.00	3.10	29.37	48.50	20.00	8.01	30.48	30.00	5.00	1.00	33.28	0.00	0.00
		(1.64)	(5.32)	(0.50)	(0.55)	(1.78)	(5.98)	(0.50)	(0.41)	(11.42)	(6.77)	(0.46)	(0.23)	(0.00)	(9.42)	(0.00)	(0.00)
	EBIC	2.48	28.57	48.00	38.00	3.07	30.13	48.00	18.00	8.24	30.89	32.00	6.00	1.00	33.25	0.00	0.00
		(1.62)	(5.51)	(0.51)	(0.55)	(1.76)	(7.60)	(0.51)	(0.40)	(11.46)	(6.40)	(0.48)	(0.24)	(0.00)	(9.38)	(0.00)	(0.00)

Table 1: Example 1. Simulation results for the adaptive group Lasso, group Lasso, ordinary Lasso and linear model with Lasso, $n = 50, 100$ or 200, $p = 1000$. NV, average number of the variables being selected; ME, 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. Top panel, independent predictors; bottom panel, correlated predictors.

		$p = 10$				$p = 20$				$p = 50$			
		NV	ME	IN	CS	NV	ME	IN	CS	NV	ME	IN	CS
		Independent predictors											
$n = 200$	AGLasso(BIC)	4.02	0.27	100.00	98.00	4.01	0.34	96.00	92.00	4.10	0.88	98.00	90.00
		(0.14)	(0.10)	(0.00)	(0.14)	(0.40)	(0.10)	(0.20)	(0.27)	(0.39)	(0.19)	(0.14)	(0.30)
	AGLasso(EBIC)	4.02	0.27	100.00	99.00	4.05	0.32	100.00	94.00	4.08	0.87	98.00	90.00
		(0.14)	(0.09)	(0.00)	(0.10)	(0.22)	(0.09)	(0.00)	(0.24)	(0.30)	(0.16)	(0.14)	(0.30)
	COSSO(5CV)	4.06	0.29	100.00	98.00	4.10	0.37	100.00	92.00	4.49	1.53	94.00	84.00
		(0.24)	(0.07)	(0.00)	(0.14)	(0.39)	(0.11)	(0.00)	(0.27)	(1.10)	(0.86)	(0.24)	(0.37)
$n = 100$	AGLasso(BIC)	4.06	0.56	99.00	90.00	4.11	0.63	98.00	87.00	4.27	1.04	93.00	81.00
		(0.24)	(0.19)	(0.10)	(0.30)	(0.42)	(0.26)	(0.14)	(0.34)	(0.58)	(0.64)	(0.26)	(0.39)
	AGLasso(EBIC)	4.06	0.54	99.00	91.00	4.10	0.59	98.00	89.00	4.22	1.01	93.00	83.00
		(0.24)	(0.21)	(0.10)	(0.31)	(0.39)	(0.22)	(0.14)	(0.31)	(0.56)	(0.60)	(0.26)	(0.38)
	COSSO(5CV)	4.17	0.53	96.00	89.00	4.18	1.04	83.00	63.00	4.89	6.63	30.00	11.00
		(0.62)	(0.19)	(0.20)	(0.31)	(0.96)	(0.64)	(0.38)	(0.49)	(1.50)	(1.29)	(0.46)	(0.31)
$n = 50$	AGLasso(BIC)	4.18	0.72	98.00	84.00	4.25	0.99	96.00	79.00	4.30	1.06	90.00	71.00
		(0.66)	(0.56)	(0.14)	(0.36)	(0.72)	(0.60)	(0.20)	(0.41)	(0.89)	(0.68)	(0.30)	(0.46)
	AGLasso(EBIC)	4.16	0.70	98.00	84.00	4.24	1.02	94.00	78.00	4.27	1.04	92.00	73.00
		(0.64)	(0.52)	(0.14)	(0.36)	(0.70)	(0.62)	(0.20)	(0.42)	(0.86)	(0.64)	(0.27)	(0.45)
	COSSO(5CV)	4.41	1.77	61.00	58.00	5.06	5.53	33.00	20.00	5.96	7.60	8.00	0.00
		(1.08)	(1.35)	(0.46)	(0.42)	(1.54)	(1.88)	(0.47)	(0.40)	(2.20)	(2.07)	(0.27)	(0.00)
		Correlated predictors											
$n = 200$	AGLasso(BIC)	3.75	0.49	82.00	70.00	3.71	1.20	75.00	66.00	3.50	1.68	68.00	62.00
		(0.61)	(0.14)	(0.39)	(0.46)	(0.68)	(0.89)	(0.41)	(0.46)	(0.92)	(1.29)	(0.45)	(0.49)
	AGLasso(EBIC)	3.75	0.49	82.00	70.00	3.73	1.18	75.00	68.00	3.58	1.60	70.00	65.00
		(0.61)	(0.14)	(0.39)	(0.46)	(0.65)	(0.88)	(0.41)	(0.45)	(0.84)	(1.27)	(0.46)	(0.46)
	COSSO(5CV)	3.70	0.53	69.00	41.00	3.89	1.24	57.00	36.00	4.11	1.76	41.00	16.00
		(0.58)	(0.17)	(0.46)	(0.49)	(0.60)	(0.90)	(0.50)	(0.48)	(0.86)	(1.33)	(0.49)	(0.37)
$n = 100$	AGLasso(BIC)	3.72	1.40	78.00	68.00	3.68	1.78	70.00	64.00	3.02	3.07	63.00	59.00
		(0.66)	(0.70)	(0.40)	(0.45)	(0.74)	(1.15)	(0.46)	(0.48)	(1.58)	(2.37)	(0.49)	(0.51)
	AGLasso(EBIC)	3.70	1.46	75.00	66.00	3.71	1.74	72.00	64.00	3.20	2.98	65.00	60.00
		(0.72)	(0.78)	(0.41)	(0.46)	(0.68)	(1.06)	(0.42)	(0.48)	(1.42)	(1.96)	(0.46)	(0.50)
	COSSO(5CV)	3.98	1.42	41.00	26.00	4.14	1.76	30.00	6.00	4.24	6.88	8.00	0.00
		(0.64)	(0.74)	(0.49)	(0.42)	(2.27)	(1.11)	(0.46)	(0.24)	(2.96)	(2.91)	(0.27)	(0.00)
$n = 50$	AGLasso(BIC)	3.30	2.26	70.00	62.00	3.06	3.02	65.00	60.00	2.87	4.01	52.00	42.00
		(1.16)	(1.09)	(0.46)	(0.49)	(1.52)	(2.14)	(0.46)	(0.50)	(1.56)	(3.69)	(0.44)	(0.52)
	AGLasso(EBIC)	3.32	2.20	70.00	64.00	3.10	3.01	68.00	62.00	2.90	3.88	50.00	42.00
		(1.14)	(1.06)	(0.46)	(0.48)	(1.51)	(2.12)	(0.45)	(0.49)	(1.54)	(3.62)	(0.42)	(0.52)
	COSSO(5CV)	4.14	3.77	25.00	6.00	4.20	6.98	5.00	0.00	4.90	9.93	1.00	0.00
		(2.25)	(2.02)	(0.44)	(0.24)	(2.88)	(2.82)	(0.22)	(0.00)	(3.30)	(4.08)	(0.10)	(0.00)

Table 2: Example 2. Simulation results comparing the adaptive group Lasso and COSSO. $n = 50, 100$ or 200, $p = 10, 20$ or 50. NV, average number of the variables being selected; ME, model error; IN, percentage of occasions on which all 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.

Probes	GL(BIC)	AGL(BIC)	Linear(BIC)	GL(EBIC)	AGL(EBIC)	Linear(EBIC)
1389584_*at*	√	√	√	√	√	√
1383673_*at*	√	√	√	√	√	√
1379971_*at*	√	√	√	√	√	√
1374106_*at*	√		√	√		√
1393817_*at*	√	√	√	√	√	
1373776_*at*	√	√	√	√	√	
1377187_*at*	√	√	√	√	√	
1393955_*at*	√	√	√	√	√	
1393684_*at*	√	√		√	√	
1381515_*at*	√	√		√	√	
1382835_*at*	√	√	√	√	√	
1385944_*at*	√	√	√	√	√	
1382263_*at*	√	√	√	√	√	√
1380033_*at*	√	√		√	√	
1398594_*at*	√		√			√
1376744_*at*	√	√		√	√	
1382633_*at*	√	√		√	√	
1383110_*at*			√			√
1386683_*at*			√			√

Table 3: Probe sets selected by the group Lasso and the adaptive group Lasso in the data example using BIC or EBIC for penalty parameter selection. GL, group Lasso; AGL, adaptive group Lasso; Linear, linear model with Lasso.

	BIC		EBIC	
	No. of probe sets	RSS	No. of probe sets	RSS
Adaptive group Lasso	15	1.52e-03	15	1.52e-03
Group Lasso	17	3.24e-03	16	3.40e-03
Ordinary Lasso	97	2.96e-07	94	8.10e-08
Linear regression with Lasso	14	2.62e-03	8	3.75e-03

Table 4: Analysis results for the data example. No. of probes, the number of probe sets selected; RSS, the residual sum of squares of the fitted model.

	Adaptive group Lasso		Group Lasso		Ordinary Lasso		Linear model with Lasso	
	ANP	PE	ANP	PE	ANP	PE	ANP	PE
BIC	15.75	1.86e-02	16.45	2.89e-02	78.48	1.40e-02	9.25	2.26e-02
	(0.85)	(0.47e-02)	(0.88)	(0.49e-02)	(3.62)	(0.90e-02)	(0.88)	(1.41e-2)
EBIC	15.55	1.78e-02	16.75	1.99e-02	80.00	1.23e-02	9.15	2.03e-02
	(0.82)	(0.42e-02)	(0.84)	(0.47e-02)	(3.50)	(0.89e-02)	(0.86)	(1.39e-02)

Table 5: Comparison of adaptive group Lasso, group Lasso, ordinary Lasso, and linear regression model with Lasso for the data example. ANP, the average number of probe sets selected averaged across 400 replications; PE, the average of prediction mean square errors for the test set.