cemmap
centre for microdata methods and practice

Penalized estimation of high-dimensional models under a generalized sparsity condition

Joel Horowitz
Jian Huang

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP17/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

PENALIZED ESTIMATION OF HIGH-DIMENSIONAL MODELS UNDER A GENERALIZED SPARSITY CONDITION

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208

and

Jian Huang
Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA 52242

December 2011

ABSTRACT

We consider estimation of a linear or nonparametric additive model in which a few coefficients or additive components are "large" and may be objects of substantive interest, whereas others are "small" but not necessarily zero. The number of small coefficients or additive components may exceed the sample size. It is not known which coefficients or components are large and which are small. The large coefficients or additive components can be estimated with a smaller mean-square error or integrated mean-square error if the small ones can be identified and the covariates associated with them dropped from the model. We give conditions under which several penalized least squares procedures distinguish correctly between large and small coefficients or additive components with probability approaching 1 as the sample size increases. The results of Monte Carlo experiments and an empirical example illustrate the benefits of our methods.

Key words: Penalized regression, high-dimensional data, variable selection.

AMS 2010 subject classification: Primary 62J05, 62J07; secondary 62E20, 62F12

The research of Joel L. Horowitz was supported in part by NSF grant SES-0817552.

PENALIZED ESTIMATION OF HIGH-DIMENSIONAL MODELS UNDER A GENERALIZED SPARSITY CONDITION

1. Introduction

We consider the mean-regression models

$$(1) \qquad Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i;\ i = 1,...,n$$

and

$$(2) \qquad Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i;\ i = 1,...,n,$$

where $Y_i \in \mathbb{R}$ is a response variable, the X_{ij}'s are scalar covariates that are fixed in model (1) and random in model (2), and the ε_i's are unobserved mean-zero random variables. In model (1), the β_j's are unknown constant coefficients. In model (2), the f_j's are unknown functions. We assume without loss of generality that the data are centered and the f_j's are normalized so that there is no intercept in either model. In model (1), we also assume that $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$ for each $j = 1,..., p$. The number of covariates (p) may be large, possibly larger than the sample size (n).

Motivated by applications in economics and other social sciences, we assume that some β_j's or f_j's are "large" in a sense that will be defined and that one or more of the large β_j's or f_j's are the objects of substantive interest. The remaining β_j's or f_j's are small but not necessarily zero. They are not objects of substantive interest, but including them in the model reduces the bias of estimates of the large β_j's or f_j's. Our interest is in estimating the large β_j's or f_j's that are of substantive interest. It turns out that the mean-square estimation errors of the large β_j's or integrated mean-square estimation errors of the large f_j's can be reduced by identifying and dropping from the model the covariates associated with small β_j's or f_j's. We give conditions under which several penalized least squares procedures distinguish correctly between large and small β_j's or f_j's with probability approaching one as $n \to \infty$. We also show that these methods provide consistent estimators of the large β_j's and f_j's. The penalization methods we consider include the adaptive LASSO, bridge estimation, and estimation with the SCAD or minimax concave (MC) penalty functions.

In model (1), let $A_s \subset \{1,..., p\}$ denote the set of small coefficients. These are defined as coefficients satisfying the generalized sparsity condition (GSC)

(3) $$\sum_{j \subset A_s} |\beta_j| \leq \eta_n ,$$

where $\{\eta_n\}$ is a sequence of non-negative constants. In our most general approach, which is the adaptive LASSO, $\eta_n = o(n^{-1/2})$. This condition is weaker than the one commonly used in the literature, which is that

(4) $$|\beta_j| = 0 \text{ if } j \in A_s .$$

Note that (4) is a special case of the GSC. In practice, the GSC can be a more realistic formulation of sparse models than is (4). Let A_0 denote the complement of A_s. We define the elements of A_0 to be large coefficients. In the adaptive LASSO, we assume that the coefficients β_j in A_0 satisfy $|\beta_j| \gg \sqrt{(\log p)/n}$. We define a covariate to be important if its coefficient is in A_0 and unimportant if its coefficient is in A_s. The other penalization methods that we consider require more restrictive definitions of the large and/or small coefficients. These definitions depend on the penalization method and are given in Section 3.2 of this paper.

In model (2), let $A_0 \subset \{1,..., p\}$ again denote the set of large additive components. We define these to be components that are non-zero in the sense that $Ef_j(X_{ij})^2 > 0$ and assume that the number of such components, q, is fixed as $n \to \infty$. Specifically,

$$A_0 = \{j: Ef_j(X_{ij})^2 > 0\}$$

and $|A_0| = q$ is fixed. We assume that the remaining additive components are small or zero in the sense that

(5) $$p \max_{x:\, j \notin A_0} f_j(x)^2 = o(n^{-2d/(2d+1)}) ,$$

where d is measures the smoothness of the additive components. Let A_s denote the set of small additive components. Condition (5) is weaker than but includes as a special case the condition used by Huang, Horowitz, and Wei (2010), which is that $f_j(v) = 0$ for all v if $j \in A_s$.

In model (1), we assume that the number of large coefficients is fixed as $n \to \infty$. Thus, for example, if p is fixed, the small coefficients are smaller than $O(n^{-1/2})$ and the large coefficients are larger than $O(n^{-1/2})$ as $n \to \infty$. In this case, the mean-square estimation errors of the large coefficients are smaller if all the unimportant covariates are excluded from the model

than if any of the unimportant covariates is included. Thus, when the objective is to estimate one or more large coefficients, it is better to drop the unimportant covariates from the model.

The assumption that the number of large coefficients is fixed is motivated by applications in the social sciences. In these applications, it is not unusual for survey data to contain hundreds or thousands of variables that are arguably related to the dependent variable of interest in the sense of having non-zero β_j coefficients in (1). However, in typical applications, most of these coefficients are thought to be small in the sense of having magnitudes and effects on the dependent variable that are smaller than the random sampling errors of their estimates. The "large" coefficients are typically few in number. For example, in an economic wage equation, the dependent variable is the logarithm of an individual's weekly wage, and the objects of interest are the coefficients of a few covariates such as an individual's years of education, years of labor-force experience, and labor union membership. However, widely available data sets for estimating wage equations can contain hundreds or even thousands of variables that may be weakly related to wages. It is not clear *a priori* which of these variables should be included in a wage equation, though it is clear that including all of them will result in very imprecise estimates of the coefficients of interest. This illustrates the usefulness of a systematic method for discriminating between covariates with large and small coefficients. We give conditions under which certain penalized least squares estimators do this with probability approaching 1 as $n \to \infty$.

In model (2), the asymptotic distributions and, therefore, integrated mean-square errors of the estimators of the large f_j's is independent of the number of small f_j's, provided that this number is also fixed as $n \to \infty$ (Horowitz and Mammen 2004). We give conditions under which a penalized least-squares estimation procedure reduces the number of small f_j's to a fixed value when the number of covariates associated with small f_j's is an increasing function of n.

Our objectives in this paper differ from those of most of the literature on estimation of high-dimensional mean-regression models. In most of the literature, the large β_j's or f_j's are assumed to be bounded away from zero, and the small ones are assumed to be exactly zero. Interest centers on identifying and estimating the large β_j's or f_j's (model selection) or selecting covariates that yield good predictions of Y. In this paper, the large β_j's are not necessarily bounded away from zero as $n \to \infty$ and the small β_j's or f_j's are not necessarily

zero. Moreover, our concern is with estimating a few large β_j's or f_j's , not with model selection or predicting Y .

The remainder of this paper is organized as follows. Section 2 presents a literature review. Section 3 describes penalized least-squares methods for selecting and estimating the large coefficients of model (1). These include the adaptive LASSO (Zou 2006) and a class of penalization methods that includes the bridge, SCAD, and MC penalties as special cases. Section 4 deals with model (2). Section 5 presents the results of a Monte Carlo investigation of the numerical performance of the adaptive LASSO. Section 6 presents an empirical example, and Section 7 presents concluding comments. The proofs of theorems are in the appendix, which is Section 8.

2. Review of the Literature

LASSO-type penalization methods for model selection (Tibshirani 1996) have attracted much attention in recent years. There is also a large literature on the use of LASSO for the related problem of prediction (see, e.g., Greenshtein and Ritov (2004) and Bickel, Ritov, and Tsybakov 2009). Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) showed that, under a strong irrepresentable condition on the design matrix, the LASSO for model (1) is model-selection consistent in high-dimensional settings. Zhang (2009) gave conditions under which the LASSO combined with a thresholding procedure consistently distinguishes between coefficients that are zero and coefficients whose magnitudes as $n \to \infty$ exceed n^{-s} for some $s < 1/2$. Zou (2006) proposed the adaptive LASSO and gave conditions under which it is model-selection consistent when the number of covariates is fixed. Huang, Ma, and Zhang (2008) provided conditions under which the adaptive LASSO is model-selection consistent even when the number of covariates is as large as $\exp(n^a)$ for some $a \in (0,1)$. Huang, Horowitz, and Wei (2010) considered model (2) and showed that a form of adaptive group LASSO provides consistent model selection in a high-dimensional setting.

Non-LASSO penalization approaches have also been considered. Knight and Fu (2001) and Huang, Horowitz, and Ma (2008) established model-selection consistency of bridge estimators. Antoniadis and Fan (2001) proposed the SCAD penalty. Fan and Li (2001); Fan and Peng (2004) further investigated the properties of least-squares and penalized likelihood estimators with the SCAD penalty. Zhang (2010) investigated penalized least squares estimation with the MC penalty. Other penalization methods have been investigated by Fan, Peng, and Huang (2005); Lv and Fan (2009); and Zou and Zhang (2009).

The foregoing model-selection procedures assume that the large β_j's in model (1) are non-zero and that the small β_js are exactly zero. In a recent paper, Zhang and Huang (2008) studied the selection properties of the LASSO under the GSC when $p > n$. They showed that the LASSO selects a model that includes all the covariates with large coefficients and has the right order of dimensionality. However, in general, the LASSO also includes some covariates with small coefficients. Thus, for example, the LASSO tends to select a model that is too large when the large coefficients are larger and the small coefficients are smaller than $O(n^{-1/2})$. Zhang (2009) gave conditions under which the LASSO combined with a thresholding procedure correctly selects coefficients that are sufficiently far from zero. However, Zhang's procedure requires a user-selected threshold, and it is not clear how to choose this threshold in applications.

In this paper, we give conditions under which with probability approaching one as $n \to \infty$, several penalized least-squares procedures correctly distinguish between large and small coefficients or additive components under the GSC. No user-selected thresholds are needed.

3. The Linear Model

This section describes methods for selecting and estimating the large β_j coefficients in model (1). Section 3.1 gives conditions under which the adaptive LASSO procedure of Zou (2006) distinguishes correctly between large and small β_j's as $n \to \infty$. Section 3.2 gives conditions under which penalized least-squares estimation with a SCAD, MC, or bridge penalty function does this.

3.1 The Adaptive LASSO

Define $\boldsymbol{y} = (Y_1, ..., Y_n)'$. Let $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})'$ denote the vector of values of the j'th covariate, and let $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ denote the design matrix. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, and let $\boldsymbol{\beta}_0$ denote the true but unknown value of $\boldsymbol{\beta}$. Let $\|\cdot\|_2$ denote the ℓ_2 norm. The ordinary LASSO objective function is

$$L_1(\boldsymbol{\beta}; \lambda_1) = 0.5\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|, \tag{6}$$

where $\lambda_1 \geq 0$ is the penalty parameter. The LASSO estimator is defined as $\tilde{\beta}_n(\lambda_1) = \arg\min_{\boldsymbol{\beta}} L_1(\boldsymbol{\beta}; \lambda_1)$.

The adaptive LASSO objective function is

(7) $$L_2(\boldsymbol{\beta};\lambda_1) = 0.5\left\|\boldsymbol{y} - \boldsymbol{X\beta}\right\|_2^2 + \lambda_2 \sum_{j=1}^{p} w_j \,|\, \beta_j \,|,$$

where $\lambda_2 \geq 0$ is the penalty parameter. The weights w_j are

$$w_j = |\, \tilde{\beta}_{nj} \,|^{-1},$$

where $\tilde{\beta}_{nj}$ is the j'th component of $\tilde{\beta}_n(\lambda_1)$. The adaptive LASSO estimator is defined as $\hat{\beta}_n(\lambda_2) = \arg\min_{\boldsymbol{\beta}} L_2(\boldsymbol{\beta};\lambda_1)$. We define $w_j = \infty$ when $\tilde{\beta}_{nj} = 0$, and we set $0 \times \infty = 0$. Minimization of (7) results in $\hat{\beta}_{nj} = 0$ if $w_j = \infty$. Thus, if a variable is not selected by the LASSO, it is not selected by the adaptive LASSO. Coefficients that are known to be large *a priori* can be omitted from the penalty term.

Under conditions (A1)-(A3) below, the LASSO selects (asymptotically) all coefficients that exceed a certain threshold. However, the LASSO also tends to select coefficients that are below the threshold. The adaptive LASSO is a way to correct LASSO's over-selection problem.

We use the following notation. For any $A \subseteq \{1,..., p\}$, let $X_A = \{\boldsymbol{x}_j : j \in A\}$ and $C_A = X_A' X_A / n$. Define

$$c_{\min}(m) = \min_{|A|=m,\ \|\nu\|_2=1} \min \nu' C_A \nu, \qquad c_{\min}(m) = \max_{|A|=m,\ \|\nu\|_2=1} \max \nu' C_A \nu,$$

where $|A|$ is the number of elements of A. We say that the covariate matrix X satisfies the sparse Riesz condition (SRC) with rank q and spectrum bounds $0 < c_* < c^* < \infty$ if

(8) $$c_* \leq c_{\min}(q) \leq c_{\max}(q) \leq c^* \quad \forall\ A \text{ with } |A| = q \text{ and } \nu \in \mathbb{R}^q.$$

Under (8), all the eigenvalues of C_A are contained in the interval $[c_*, c^*]$ when $|A| \leq q$.

We make the following assumptions.

(A1) The random variables $\varepsilon_1, \varepsilon_2, ...$ are independently and identically distributed with mean 0. There are constants $C > 0$ and $K > 0$ such that $P(|\varepsilon_i| > z) \leq K \exp(-Cz^2)$ for all $z \geq 0$ and $i = 1, 2, ...$

(A2) There is a finite constant $c_1 > 0$ such that $\eta_n \leq c_1 \sqrt{q}\, \lambda_1 / n$. Moreover, q is fixed, and $|A_0| \equiv k \leq q$.

(A3) The SRC holds.

Condition (A1) requires the ε_is to have subgaussian tails. Condition (A2) defines the class of small coefficients and states our assumption that the number of large coefficients is fixed.

Assumption (A3) holds if the restricted eigenvalue assumption $RE(s, c_0)$ of Bickel, Ritov, and Tsybakov (2009) holds for some $s \geq q/2$.

Let $\tilde{A}_1 = \{j: \tilde{\beta}_{nj}(\lambda_1) \neq 0\}$ be the set of coefficients estimated to be non-zero by the LASSO. The following lemma, which is proved in Zhang and Huang (2008), summarizes important properties of $\tilde{A}_1$ and $\tilde{\beta}_{nj}$.

Lemma 1: *Let (A1)-(A3) hold, and let* $\lambda_1 = O(\sqrt{n \log p})$. *Then there are finite constants* M_1 *and* M_2 *such that*

(i) $|\tilde{A}_1| \leq M_1 q$ *with probability approaching 1 as* $n \to \infty$.

(ii) All covariates with $\beta_{0j}^2 > M_2 q \lambda_1^2 / (c_* c^* n^2)$ *are selected with probability approaching 1 as* $n \to \infty$.

(iii) $\left\| \tilde{\beta}_n - \boldsymbol{\beta}_0 \right\|_2^2 = O_p(h_n^2)$, *where* $h_n = \sqrt{(\log p)/n}$.

Lemma 1 shows that with high probability, the number of covariates selected by the LASSO is a finite multiple of the number of covariates in A_0 (and, therefore, of the number of covariates with large coefficients). Moreover, all covariates exceeding the threshold in (ii) are selected with probability approaching 1 as $n \to \infty$. In particular, all of the covariates with large coefficients are selected with probability approaching 1 if $\eta_n = o(\sqrt{(\log p)/n})$ and the large β_j's are larger than $O(\sqrt{(\log p)/n})$. In addition, the LASSO estimator is estimation consistent. However, estimation consistency does not imply model-selection consistency.

We now give conditions under which the adaptive LASSO achieves model-selection consistency. Denote the smallest and largest eigenvalues of $C_{A_0} = X'_{A_0} X_{A_0}/n$ by τ_{n1} and τ_{n2}, respectively. Make the following additional assumptions.

(A4) There are constants $0 < \tau_1 \leq \tau_2 < \infty$ such that $\tau_1 \leq \tau_{n1} \leq \tau_{n2} \leq \tau_2$ for all sufficiently large n.

(A5) Let $b_{n1} = \min_{j \in A_0} |\beta_{0j}|$. As $n \to \infty$, the nonstochastic quantities η_n, h_n, λ_2, and b_{n1} satisfy

$$\frac{\lambda_2}{n b_{n1}^2} + \frac{(h_n + \eta_n)}{b_{n1}} + \frac{n\eta(h_n + \eta_n)}{\lambda_2} \to 0.$$

In our model, $|A_0|$ is fixed as $n \to \infty$, so it is reasonable to assume in (A4) that the eigenvalues of C_{A_0} are bounded away from 0 and ∞. (A5) restricts η_n, λ_2, and b_{n1}. It requires b_{n1}, the

smallest of the large coefficients, to be not too small and the ℓ_1 norm of the small coefficients to be not too large. In particular, it requires $b_{n1} \gg \eta_n$. In other words, there must be enough separation between the large and small coefficients for the adaptive LASSO to distinguish between them.

Now define $\hat{\beta}_{nA_0} = \{\hat{\beta}_{nj} : j \in A_0\}$ and $\beta_{0A_0} = \{\beta_{0j} : j \in A_0\}$. For any vector $u = (u_1, u_2, ...)'$, define $\text{sgn}(u) = (\text{sgn}(u_1), \text{sgn}(u_2), ...)'$, where $\text{sgn}(u_1) = -1$, 0, or 1 according to whether $u_1 < 0$, $u_1 = 0$, or $u_1 > 0$.

Theorem 1: *Let (A1)-(A5) hold. Then as* $n \to \infty$,

$$P(\hat{\beta}_{nj} = 0 \ \forall \ j \in A_s) \to 1 \quad \textit{and} \quad P(\text{sgn}(\hat{\beta}_{nA_0}) = \text{sgn}(\beta_{0A_0})) \to 1.$$

Thus, with probability approaching 1 as $n \to \infty$, the adaptive LASSO selects all the covariates with large coefficients and drops the covariates with small coefficients in the sense that it sets the coefficients of those covariates equal to zero.

If, as often happens in social science applications, the total number of covariates is less than the sample size, then we can consider a model in which p is fixed as $n \to \infty$, the small coefficients satisfy $\eta_n = o(n^{-1/2})$, and the large coefficients satisfy $b_{n1} \geq \kappa\sqrt{(\log n)/n}$ as $n \to \infty$ for some constant $\kappa > 0$. It follows from Theorem 1 with $\lambda_2 \propto \sqrt{\log n}$ that as $n \to \infty$, the adaptive LASSO estimates of the large coefficients are non-zero and the estimates of the small coefficients are zero. Moreover, a straightforward calculation shows that the mean-square error (MSE) of the adaptive LASSO estimator of each large coefficient is never larger and, except in special cases, is strictly smaller than the MSE of the ordinary least squares (OLS) estimator that is obtained when all covariates are included in (1). Thus, the adaptive LASSO improves the precision of the estimates of the large coefficients.

If $p > n$, we can consider a model in which the large coefficients satisfy $b_{n1} \geq \kappa(\log p)/n^{1/2}$ for some constant $\kappa > 0$, and $\lambda_2 \propto \log p$. Then it follows again from Theorem 1 that as $n \to \infty$, the adaptive LASSO estimates of the large coefficients are non-zero and the estimates of the small coefficients are zero. Moreover, the MSE of the adaptive LASSO estimator of each large coefficient is no larger and, except in special cases, is strictly smaller than MSE of the OLS estimator that is obtained by including in the model any group of up to $n - q - 1$ unimportant covariates or linear combinations of unimportant covariates. In summary, the adaptive LASSO estimator reduces the MSE of the estimator of any large coefficient if there is sufficient separation between the magnitudes of the large and small coefficients.

3.2 Penalized Least-Squares Estimation with Other Penalty Functions

We now investigate penalized least-squares estimation of model (1) with a class of penalty functions that includes the bridge, SCAD, and MC penalties. As in Section 3.1, we consider a two-step estimation procedure. The first step is the same as that in Section 3.1. It consists of solving the problem $\tilde{\beta}_n(\lambda_1) = \arg\min_{\boldsymbol{\beta}} L_1(\boldsymbol{\beta};\lambda_1)$, where L_1 is defined in (6). Under the assumptions of Lemma 1, the number of non-zero components of $\tilde{\beta}_n$ is fixed as $n \to \infty$ and includes all the large β_j's. Let $\tilde{\boldsymbol{X}}$ denote the design submatrix consisting of the columns of $\boldsymbol{X}$ corresponding to non-zero components of $\tilde{\beta}_n$. Let $\|\cdot\|_2$ denote the ℓ_2 norm. The second estimation step consists of minimizing

$$L_3(\boldsymbol{\beta};\lambda_1) = \left\|\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right\|_2^2 + \sum_{j:\tilde{\beta}_j \neq 0} p_{\lambda_n}(|\beta_j|),$$

where p_λ is a penalty function and λ_n is the penalty parameter. Denote the resulting estimator by $\hat{\beta}_n(\lambda_n)$.

We assume that the penalty function satisfies the following condition.

(A6) The penalty function has the form $p_\lambda(v) = \lambda f(v)$, where f is a bounded, non-decreasing function that may depend on n and λ and satisfies

(i) $f(0) = 0$.

(ii) One of the following holds:

(a) There are constants $C < \infty$ and τ that may depend on n and λ such that $0 \le f'(v) \le C$ for all v, and $f'(v) = 0$ if $v \ge \tau$. Moreover there are constants $b > 0$ and $\delta > 0$ such that $f'(v) \ge \delta$ if $v \le b\lambda/n$.

(b) There is a $C < \infty$ such that $0 < f'(v) \le C$ for all $v \ge \varepsilon$ and some $\varepsilon > 0$. Moreover $0 \le f(v) \le Cv^\gamma$ for all $v > 0$ and some γ such that $0 < \gamma < 1$. Also, $\lim_{\delta/v \to 0}[f(|v+\delta|) - f(|v|)] \ge c|\delta|^\gamma$.

In addition, we adopt the following more restrictive definitions of large and small coefficients.

(A7) If (A6)(ii)(a) holds, then the large coefficients satisfy $|\beta_j| \gg [\lambda_n(\log p)/n]^{1/2} \gg \tau$ for all $j \in A_0$, where $\{\lambda_n\}$ is a sequence of positive constants such that $n^{-1/2}\lambda_n \to \infty$ and $n^{-\theta}\lambda_n \to 0$ for some $\theta > 1/2$ as $n \to \infty$. The small coefficients satisfy $\sum_{j \in A_0} |\beta_j| = o(n^{-\theta})$. If (A6)(ii)(b) holds, then the large coefficients satisfy $|\beta_j| \ge \varepsilon$ for

all $j \in A_0$ and some $\varepsilon > 0$. The small coefficients satisfy $\sum_{j \in A_s} |\beta_j|^\gamma = o(n^{-1/2})$ for the γ in (A6)(ii)(b).

The SCAD and MC penalty functions satisfy (A6)(ii)(a). We write the SCAD penalty function as

$$p'_{\lambda_n}(v) = \lambda_n [I(v \le n^{-1}\lambda_n) + \frac{(an^{-1}\lambda_n - v)_+}{(a-1)n^{-1}\lambda_n} I(v > n^{-1}\lambda_n)],$$

where I is the indicator function and $a > 2$ is a constant. The MC penalty function is

$$p_{\lambda_n}(v) = \lambda_n \int_0^v \left(1 - \frac{nx}{\gamma\lambda_n}\right)_+ dx$$

for some $\gamma > 0$. The bridge penalty function satisfies (A6)(ii)(b). The bridge penalty function is

$$p_{\lambda_n}(v) = \lambda_n |v|^\gamma,$$

where γ is a constant satisfying $0 < \gamma < 1$. The ordinary LASSO penalty function, $p_\lambda(v) = |v|$, does not satisfy (A6).

Now define $\Sigma_n = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}/n$. Assume that

(A8) $\lim_{n\to\infty} \Sigma_n = \Sigma$ for some nonsingular matrix Σ.

We now have the following result.

Theorem 2: *Let (A1)-(A3) and (A5)-(A7) hold. Let* $n^{-1/2}\lambda_n \to \infty$ *and* $n^{-\theta}\lambda_n \to 0$ *as* $n \to \infty$ *if (A6)(ii)(a) holds. Let* $n^{-\gamma}\lambda_n \to \infty$ *and* $n^{-1/2}\lambda_n \to 0$ *as* $n \to \infty$ *if (A6)(ii)(b) holds. Then*

$$P(\hat{\beta}_{nj} = 0 \ \forall \ j \in A_s) \to 1 \quad \textit{and} \quad P(\text{sgn}(\hat{\beta}_{nA_0}) = \text{sgn}(\beta_{0A_0})) \to 1.$$

Thus, under the conditions of Theorem 2, the second-stage estimator distinguishes correctly between large and small coefficients with probability approaching 1 as $n \to \infty$.

4. The Nonparametric Additive Model

This section presents a method for selecting and estimating the large additive components f_j in model (2). Horowitz and Mammen (2004) describe a method for estimating the f_j's that is oracle efficient when the dimension of model (2) remains fixed as $n \to \infty$. The estimator of each f_j has the same asymptotic distribution that it would have if the other f_j's were known. There is no need to distinguish between large and small f_j's. Here, we consider the case in which the dimension of the model increases and may exceed n as $n \to \infty$. We present a two-

step procedure for selecting and estimating the large f_j's. The first step of the procedure consists of penalized least-squares estimation of series approximations to the f_j's using a group LASSO penalty function. Huang, Horowitz, and Wei (2010) showed that this procedure reduces the number of f_j's to a fixed value when Ef_j^2 ($j = 1,..., p$) is either zero or bounded away from zero. We show that asymptotically, the same procedure reduces the number of f_j's to a fixed value and retains all f_j's for which Ef_j^2 is large in the sense defined in Section 1. The second step consists of using the estimator of Horowitz and Mammen (2004) to re-estimate the f_j's that are retained in the first step. Horowitz and Mammen (2004) present the properties of the second-step estimator. Therefore, we treat only the first step here.

Assume that each $X_{\cdot j}$ takes values in $[a,b]$, where $a < b$. Let $\{\phi_k : k = 1,..., m_n\}$ denote a normalized B-spline basis for polynomial splines of degree $l \geq 1$ on $[a,b]$, where $m_n = K_n + l$ and K_n is the number of spline knots in (a,b). Define the centered B-splines

$$\psi_k(X_{ij}) = \phi_k(X_{ij}) - n^{-1}\sum_{\ell=1}^{n}\phi_k(X_{\ell j}); \quad k = 1,..., m_n; \; j = 1,..., p.$$

Define

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

Let $Z_j = (Z_{1j},...,Z_{nj})'$, $Z = (Z_1,...,Z_p)$ and $Y = (Y_1 - \bar{Y},...,Y_n - \bar{Y})'$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. The first-step estimator of our procedure consists of solving the problem

$$\tilde{\beta}_{nj} = \arg\min_b \left\| Y - Zb_n \right\|_2^2 + \lambda_n \sum_{j=1}^{p} \left\| b_j \right\|_2,$$

where b_j is the $m_n \times 1$ vector $(b_{j1},...,b_{jm_n})'$ and λ_n is the penalty parameter. This is also the problem solved in the first estimation step of Huang, Horowitz, and Wei (2010).

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

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

Order the f_j's so that the first q are large and the rest are small or zero.

Make the following assumptions.

(A9) The number of large additive components, q, is fixed. Moreover, there is a constant $C_f > 0$ such that $\min_{1 \le j \le q} \left\| f_j \right\|_2 \ge C_f$.

(A10) The random variables $\varepsilon_1, ..., \varepsilon_n$ are independently and identically distributed with $E(\varepsilon_i) = 0$. Moreover, $P(|\varepsilon_i| > x) \le K \exp(-Cx^2)$ $(i = 1, ..., n)$ for all $x \ge 0$, where C and K are finite constants.

(A11) $Ef_j(X_{\cdot j}) = 0$ and $f_j \in \mathcal{F}$ for all $j = 1, ..., p$.

(A12) The covariate vector $(X_{\cdot 1}, ..., X_{\cdot j})$ has a continuous probability density function with respect to Lebesgue measure. Moreover, there exist constants C_1 and C_2 such that the probability density function g_j of $X_{\cdot j}$ satisfies $0 < C_1 \le g_j(x) \le C_2 < \infty$ for every $x \in [a, b]$ and every $j = 1, ..., p$.

(A12) Every additive component is either large or small. The small components satisfy equation (5).

Assumptions (A9)-A(11) are made by Huang, Horowitz, and Wei (2010) and are explained there. Assumption (A12) defines the small additive components.

Define $\tilde{A}_0 = \{ j : \left\| \tilde{\beta}_{nj} \right\|_2 \ne 0;\ j = 1, ..., p \}$ and

$$\beta_{nj} = \arg \min_{b_1, ..., b_{m_n}} \left\| \sum_{k=1}^{m_n} b_k \psi_k(X_{\cdot j}) - f_j(X_{\cdot j}) \right\|_2^2 .$$

Also define

$$A_2 = A_1 \cup \{ j : \left\| \tilde{\beta}_{nj} \right\|_2 \ne 0 \}$$

and

$$\bar{A}_2 = \bar{A}_1 \cap \{ j : \left\| \tilde{\beta}_{nj} \right\|_2 = 0 \},$$

where $\bar{A}$ is the complement of any set A. Let $\tilde{\beta}_{nA_2}$ and β_{nA_2}, respectively, be the vectors consisting of the $\tilde{\beta}_{nj}$'s and β_{nj}'s for which $j \in A_2$.

We now have the following result, which extends Theorem 1 of Huang, Horowitz, and Wei (2010) to the case in which some additive components may be small but are not necessarily zero.

Theorem 3: *In model (2), let (A9)-(A12) hold. In addition, let* $\lambda_n \geq C\sqrt{n\log(pm_n)}$ *for some sufficiently large but finite constant* C *and let* $m_n \asymp n^{1/(2d+1)}$. *Then*

(i) With probability approaching 1 as $n \to \infty$ $|\tilde{A}_0| \leq M_1 |A_0| = M_1 q$, *for some finite constant* $M_1 > 1$.

(ii) If $m_n^2 \log(pm_n)/n \to 0$ *and* $\lambda_n^2 m_n / n^2 \to 0$ *as* $n \to \infty$, *then* $\left\|\tilde{\beta}_{nj}\right\|_2 \neq 0$ *with probability approaching 1 as* $n \to \infty$ *for every* $j \in A_0$.

(iii)

$$\left\|\tilde{\beta}_{nA_2} - \beta_{nA_2}\right\|_2^2 = O_p\left(\frac{m_n^2 \log(pm_n)}{n}\right) + O_p\left(\frac{1}{n}\right) + O\left(\frac{1}{m_n^{2d-1}}\right) + O\left(\frac{4m_n^2\lambda_n^2}{n^2}\right).$$

Under the conditions of Theorem 3, the group LASSO selects all of the large additive components of model (2) with probability approaching 1 as $n \to \infty$. Moreover, the group LASSO selects only a fixed number of additive components. Part (iii) of the theorem states the rate of convergence of the estimated components.

5. Monte Carlo Experiments

This section reports the results of a Monte Carlo investigation of the finite-sample performance of the LASSO and adaptive LASSO for model (1) when the small coefficients are not necessarily zero. We write model (1) in the form

$$Y_i = \sum_{j=1}^{d} \beta_j X_{ij} + \sum_{j=d+1}^{p} \beta_j X_{ij} + \varepsilon_i; \; i = 1,...,n,$$

where $\beta_1,...,\beta_d$ are large coefficients and the coefficients $\beta_{d+1},...,\beta_p$ are small or zero. The random variables ε_i are independently distributed as $N(0,\sigma_\varepsilon^2)$. The covariates are fixed in repeated samples and are centered and scaled so that

$$n^{-1}\sum_{i=1}^{n} X_{ij} = 0; \quad n^{-1}\sum_{i=1}^{n} X_{ij} = 0; \; i = 1,...,n$$

The covariates are generated as follows. Define

$$\xi_{ij} = \zeta_{ij} + \left(\frac{\rho_1}{1-\rho_1}\right)^{1/2} v_i; \quad i = 1,...,n; \; j = 1,..., p/2$$

$$\xi_{ij} = \zeta_{ij} + \left(\frac{\rho_2}{1-\rho_2}\right)^{1/2} v_i; \quad i = 1,...,n; \; j = p/2+1,..., p,$$

where the ζ_{ij}'s and ν_i's are independently distributed as $N(0,1)$ and $0 \le \rho_1, \rho_2 < 1$. Define

$$\bar{\xi}_j = n^{-1}\sum_{i=1}^{n}\xi_{ij};\ s_j^2 = n^{-1}\sum_{i=1}^{n}(\xi_{ij} - \bar{\xi}_j)^2 .$$

Then

$$X_{ij} = \frac{\xi_{ij} - \bar{\xi}_j}{s_j}.$$

Moreover,

$$corr(X_{ij}, X_{ik}) = \begin{cases} \rho_1 & \text{if } 1 \le j,k \le p/2 \\ \rho_2 & \text{if } p/2 < j,k < p \\ (\rho_1\rho_2)^{1/2} & \text{if } j \le p/2 < k \le p \end{cases}$$

In the experiments reported here,

$$\beta_j = \begin{cases} 1 \text{ if } 1 \le j \le d \\ 0.05 \text{ if } d+1 \le j \le p/2 \\ 0 \quad p/2+1 \le j \le p \end{cases}$$

In addition, $n = 100$, $p = 50$, $\sigma_\varepsilon^2 = 10$, $\rho_1 = 0.5$, and $\rho_2 = 0.1$. The coefficient of interest is β_1. Experiments are reported with $d = 2$, 4, and 6 and with the LASSO and adaptive LASSO. The penalization parameter is obtained by minimizing the BIC.

Table 1 shows the mean-square errors of the estimates of β_1 obtained from applying OLS to the full model and to the model containing only the variables whose coefficients are large (the reduced model). These results are obtained analytically using the algebra of least squares. They show that the mean-square error is smaller when β_1 is estimated from the reduced model than when it is estimated from the full model.

Table 2 shows the results of estimation using the LASSO and adaptive LASSO. There are 1000 Monte Carlo replications in each experiment. Both versions of the LASSO reduce the mean-square estimation error by about a factor of two relative to OLS estimation with the full model. Not surprisingly, neither version achieves the mean-square error that is achievable when the variables with large coefficients are known. The model selected by the LASSO has a higher probability of containing all the important covariates than does the model selected by the adaptive LASSO.

If β_1 is the coefficient of interest, it is reasonable to consider versions of the LASSO and adaptive LASSO in which X_{i1} $(i = 1,.,.,n)$ is always in the chosen model. This can be achieved by leaving β_1 out of the penalty function. Table 3 shows the results of LASSO and adaptive

LASSO estimation with β_1 not in the penalty function. Forcing X_{i1} into the model greatly reduces the mean-square error of the adaptive LASSO estimator of β_1. It is essentially the same as the mean-square error that is obtained by applying OLS to the model with only the covariates with large coefficients.

6. An Empirical Example

This section presents an empirical example that illustrates the application of the LASSO and adaptive LASSO in a setting where many coefficients are plausibly small but non-zero. The application consists of estimating a wage equation for black males aged 40-49 years who reside in the northeastern U.S. The data are from the National Longitudinal Survey of Youth. There are 62 observations. The dependent variable is the logarithm of the hourly wage. There are 42 covariates, including scores on 10 sections of the armed forces qualification examination, indicators of education level, a variety of personal characteristics, a binary indicator of marital status (married or not), and a binary indicator of membership in a labor union. The variables of interest in this example are marital status and union membership. Their coefficients measure the fractional change in the wage associated with being married or belonging to a labor union. It is arguable that all of the covariates affect productivity and, therefore, the hourly wage but that the effects of many covariates may be small.

Application of the LASSO and adaptive LASSO using the BIC to select the penalty parameter resulted in selection of 7 and 4 covariates, respectively. An asymptotic chi-square test does not reject the hypotheses that the coefficients of the variables not selected by the LASSO or adaptive LASSO are zero ($p > 0.6$). This implies that the values of these coefficients are small enough to be within random sampling error of zero. They are not necessarily equal to zero. Table 4 shows the estimates and asymptotic standard errors of the two coefficients of interest that are obtained from applying ordinary least squares to the full model (all 42 covariates), the model selected by the LASSO, and the model selected by the adaptive LASSO. The three point estimates of the coefficient of labor union membership are similar, but the standard error of the estimate obtained from the full model is nearly twice as large as the standard errors obtained from the models selected by the LASSO and adaptive LASSO. The estimates of the coefficient of marital status obtained from the models selected by the LASSO and adaptive LASSO are nearly 4 times as large as the estimate obtained from the full model, and the standard errors of the estimates obtained from the selected models are about 55% of the standard error obtained with the full model.

7. Conclusions

In applications of mean regression analysis, it is often the case that there are many covariates whose effects on the conditional mean of the dependent variable are thought to be small but not necessarily zero and there relatively few covariates that have large effects on the conditional mean function. In such situations, the precision of estimating the large effects can be increased by leaving the covariates with small effects out of the model. However, it is rarely known a priori which covariates have large effects and which have small ones. This paper has given conditions under which the adaptive LASSO and several penalized least squares methods correctly distinguish between covariates with large and small effects in a linear model and a nonparametric additive model. Specifically, we have shown that with probability approaching one as the sample size increases, the adaptive LASSO and penalized least squares correctly distinguish between covariates with large and small effects under a generalized sparsity condition and other mild regularity conditions.

8. Proofs of Theorems

Proof of Theorem 1

Let $\psi(v) = \exp(v^2) - 1$. The ψ-Orlicz norm $\|x\|_\psi$ of any random variable x is defined as $\|x\|_\psi = \inf\{C > 0 : E\psi(|x|/C) \le 1\}$. The Orlicz norm is useful for obtaining maximal inequalities (Van der Vaart and Wellner 1996).

Lemma 2: *Suppose that* $\varepsilon_1, ..., \varepsilon$ *are iid random variables with* $E\varepsilon_i = 0$ *and* $Var(\varepsilon_i^2) = \sigma^2$. *Suppose that* $P(|\varepsilon_i| > z) \le K\exp(-Cz^2)$ *for* $i = 1, ..., n$ *and constants* C *and* K. *Then, for all constants* a_i *satisfying* $\sum_{i=1}^n a_i^2 = 1$,

$$(6) \qquad \left\| \sum_{i=1}^n a_i \varepsilon_i \right\|_\psi \le K[\sigma + (1+K)^{1/2} C^{-1/2}],$$

where K *is a constant. Consequently*

$$(7) \qquad g_n(t) \equiv \sup_{a_1^2 + ... + a_n^2 = 1} P\left(\sum_{i=1}^n a_i \varepsilon_i > t \right) \le \exp(-t^2/M)$$

for some constant M *that depends only on* K *and* C.

Proof of Theorem 1: By the Karush-Kuhn-Tucker conditions, $\hat{\beta}_n = (\hat{\beta}_{n1}, ..., \hat{\beta}_{np})'$ is the unique adaptive LASSO estimator if

$$(8)\qquad \begin{cases} \boldsymbol{x}_j'(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_n)=\lambda_2 w_j \operatorname{sgn}(\hat{\beta}_{nj}) \ \text{ if } \ \hat{\beta}_{nj}\neq 0 \\ |\boldsymbol{x}_j'(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_n)|\le \lambda_2 w_{nj} \qquad \text{if } \hat{\beta}_{nj}=0 \end{cases}$$

and the vectors $\{\boldsymbol{x}_j, \hat{\beta}_{nj}\neq 0)$ are linearly independent. Let $\tilde{\boldsymbol{s}}_{A_0}=\{w_j \operatorname{sgn}(\beta_{0j}): j\equiv A_0\}$ and

$$\hat{\beta}_{A_0}=(X'_{A_0}X_{A_0})^{-1}(X'_{A_0}\boldsymbol{y}-\lambda_2\tilde{\boldsymbol{s}}_{A_0})$$

$$(9)\qquad =\beta_{0A_0}+n^{-1}C_{A_0}^{-1}(X'_{A_0}\varepsilon+X'_{A_0}X_{A_0}\beta_{0A_s}-\lambda_2\tilde{\boldsymbol{s}}_{A_0}),$$

where $C_{A_0}=X'_{A_0}X_{A_0}/n$. If $\operatorname{sgn}(\hat{\beta}_{A_0})=\operatorname{sgn}(\beta_{0A_0})$, then (8) holds for $\hat{\beta}_n=(\hat{\beta}_{A_0},\boldsymbol{0}_{A_0})'$, where $\boldsymbol{0}_{A_0}$ is a vector of zeros with length $|A_s|$. Let $\beta_0^*=(\beta'_{0A_0},\boldsymbol{0}'_{A_s})'$. To prove the theorem, it suffices to show that $P[\operatorname{sgn}(\hat{\beta}_{A_0})=\operatorname{sgn}(\beta_{0A_0})]\to 1$.

Since $X\hat{\beta}_n=X_{A_0}\hat{\beta}_{A_0}$ for this $\hat{\beta}_n$ and $\{\boldsymbol{x}_j: j\in A_0\}$ are linearly independent,

$$(10)\qquad \operatorname{sgn}(\hat{\beta}_n)=\operatorname{sgn}(\beta_0^*) \ \text{ if } \ \begin{cases} \operatorname{sgn}(\hat{\beta}_{A_0})=\operatorname{sgn}(\beta_{0A_0}) \\ |\boldsymbol{x}_j'(\boldsymbol{y}-X_{A_0}\hat{\beta}_{A_0})|\le \lambda_2 w_j \ \forall j\notin A_0. \end{cases}$$

Let $H_n=I_n-X_{A_0}C_{A_0}^{-1}X'_{A_0}/n$ be the projection onto the null of X'_{A_0}, where I_n is the $n\times n$ identity matrix. From (9), we have

$$(11)\qquad \boldsymbol{y}-X_{A_0}\hat{\beta}_{A_0}=H_n\varepsilon+n^{-1}\lambda_2 X_{A_0}C_{A_0}^{-1}\tilde{\boldsymbol{s}}_{A_0}+H_nX_{A_s}\beta_{0A_s}.$$

By (10) and (11), $\operatorname{sgn}(\hat{\beta}_n)=\operatorname{sgn}(\beta_0^*)$ if

$$(12)\qquad \begin{cases} \operatorname{sgn}(\beta_{0j})(\beta_{0j}-\hat{\beta}_{nj})<|\beta_{0j}| & \forall j\in A_0 \\ |\boldsymbol{x}_j'(H_n\varepsilon+n^{-1}\lambda_2 X_{A_0}C_{A_0}^{-1}\tilde{\boldsymbol{s}}_{A_0}+H_nX_{A_s}\beta_{0A_s})|\le \lambda_2 w_j \ \forall j\notin A_0. \end{cases}$$

Thus, by (9) and (12), for any $0<\kappa<\kappa+\nu<1$

$$P\{\mathrm{sgn}(\hat{\beta}_n) \neq \mathrm{sgn}(\beta_0^*)\} \leq$$

$$P\{n^{-1} |\, \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' X_{A_s} \beta_{0A_s} | \geq |\beta_{0j}|/3 \text{ for some } j \in A_0\}$$

$$+ P\{n^{-1} |\, \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' \varepsilon | \geq |\beta_{0j}|/3 \text{ for some } j \in A_0\}$$

$$+ P\{n^{-1}\lambda_2 |\, \boldsymbol{e}_j' C_{A_0}^{-1} \tilde{\boldsymbol{s}}_{A_0} | \geq |\beta_{0j}|/3 \text{ for some } j \in A_0\}$$

$$+ P\{|\, \boldsymbol{x}_j' H_n \varepsilon | \geq \lambda_2 w_j/3 \text{ for some } j \in A_0\}$$

$$+ P\{n^{-1} |\, \boldsymbol{x}_j' X_{A_0} C_{A_0}^{-1} \tilde{\boldsymbol{s}}_{A_0} | \geq w_j/3 \text{ for some } j \in A_0\}$$

$$+ P\{|\, \boldsymbol{x}_j' H_n X_{A_s} \beta_{0A_s} | \geq \lambda_2 w_j/3 \text{ for some } j \in A_0\}$$

$$\equiv P(B_{n1}) + P(B_{n2}) + P(B_{n3}) + P(B_{n4}) + P(B_{n5}) + P(B_{n6}),$$

where $\boldsymbol{e}_j$ is the unit vector in the direction of the j'th coordinate.

Consider B_{n1}. Because

$$n^{-1} |\, \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' X_{A_s} \beta_{0A_s} | \leq n^{-1} \left\| \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' \right\|_2 \cdot \left\| X_{A_s} \beta_{0A_s} \right\|_2$$

$$\leq n^{-1/2} \left\| C_{A_0}^{-1/2} \right\|_2 n^{1/2} \eta_n \leq \tau_{n1} \eta_n,$$

we have $P(B_{n1}) \to 0$ by (A5).

Now consider B_{n2}. Because $n^{-1} \left\| \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' \right\|_2 \leq n^{-1/2} \left\| C_{A_0}^{-1/2} \right\|_2 \leq (n\tau_{n1})^{-1/2}$ and $|\beta_{0j}| \geq b_{n1}$ for $j \in A_0$,

$$P(B_{n2}) = P(n^{-1} |\, \boldsymbol{e}_j' C_{A_0}^{-1} X_{A_0}' \varepsilon | \geq |\beta_{0j}|/3 \; \forall j \in A_0)$$

$$\leq q g_n[b_{n1}(\tau_{n1} n)^{1/2}/3]$$

with the tail probability $g_n(t)$ in Lemma 2. Therefore, $P(B_{n2}) \to 0$ by (A1), Lemma 2, (A4) and (A5).

Now $\left\| \tilde{\boldsymbol{s}}_{A_0} \right\|_2 = O_p[q^{1/2}/(nb_{n1})]$. Therefore, by (A5),

$$n^{-1}\lambda_2 \,|\, \boldsymbol{e}_j' C_{A_0}^{-1} \tilde{\boldsymbol{s}}_{A_0} \,| \le \frac{\lambda_2 \left\| \tilde{\boldsymbol{s}}_{A_0} \right\|_2}{n\tau_{n1}} = O_p(1)\frac{\lambda_2 q}{n\tau_{n1} b_{n1}} = o_p(b_{n1}).$$

This gives $P(B_{n3}) \to 0$.

For B_{n4}, we have $w_j^{-1} = |\, \tilde{\beta}_{nj} \,| \le |\, O_p(h_n) + \eta_n \,|$. Since $\left\| \boldsymbol{x}_j H_n \right\| \le n^{1/2}$, for large C

$$P(B_{n4}) \le P\{|\, \boldsymbol{x}_j' H_n \varepsilon \,| \ge (1/3)\lambda_2 / [Cn^{1/2}(h_n + \eta_n)] \;\; \forall j \notin A_0\} + o(1)$$

$$\le q_n g_n \{(1/3)\lambda_2 / [Cn^{1/2}(h_n + \eta_n)]\}.$$

Therefore, by Lemma 2 and (A5), $P(B_{n4}) \to 0$.

For B_{n5} we have

$$\frac{|\, \boldsymbol{x}_j' X_{A_0} C_{A_0}^{-1} \tilde{s}_{A_0} \,|}{n w_j} \le \left\| n^{-1} \boldsymbol{x}_j' X_{A_0} C_{A_0}^{-1} \right\|_2 \cdot \left\| \tilde{s}_{A_0} \right\|_2 |\, \tilde{\beta}_{nj} \,|$$

$$\le \frac{\tau_2^{1/2} q^{1/2}}{\tau_1 b_{n1}} [O_p(h_n) + \eta_n].$$

Therefore, $P(B_{n5}) \to 0$ by (A5).

Finally, for B_{n6} we have

$$|\, \boldsymbol{x}_j' H_n X_{A_s} \beta_{0A_s} \,| \le \left\| \boldsymbol{x}_j \right\|_2 \cdot \left\| X_{A_s} \beta_{0A_s} \right\|_2 \le n\eta_n .$$

Therefore,

$$\frac{|\, \boldsymbol{x}_j' H_n X_{A_s} \beta_{0A_s} \,|}{w_j} \le n\eta_n \,|\, \tilde{\beta}_{nj} \,| \le n\eta_n [O_p(h_n) + \eta_n].$$

Therefore, $P(B_{n6}) \to 0$ by (A5). This completes the proof. ■

Proof of Theorem 2: The proof takes place in 3 steps.

Step 1 consists of proving that $\left\| \hat{\beta} - \beta_0 \right\|_2^2 = O(n^{-1} + \lambda_n n^{-1})$ with probability approaching 1 as $n \to \infty$. Let r denote the (asymptotically fixed) number of covariates at this estimation stage. Denote the covariates by $\{X_{ij} : i = 1,...,n;\ j = 1,...,r\}$. Set $X_i = (X_{i1},...,X_{ir})$. Define

$$S_n(b) = \sum_{i=1}^{n} (Y_i - X_i b)^2 + \sum_{j=1}^{r} p_{\lambda_n}(|\, \beta_j \,|).$$

Then $S_n(\hat{\beta}_n) \le S_n(\beta_0)$. Therefore,

$$\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 + \sum_{j=1}^{r} p_{\lambda_n}(|\hat{\beta}_j|) \le \sum_{i=1}^{n}(Y_i - X_i\beta_0)^2 + \sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|)$$

and

$$\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 - \sum_{i=1}^{n}(Y_i - X_i\beta_0)^2 \le \sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|) - \sum_{j=1}^{r} p_{\lambda_n}(|\hat{\beta}_j|).$$

Some algebra shows that this is equivalent to

$$\sum_{i=1}^{n}[X_i(\hat{\beta} - \beta_0)]^2 - 2\sum_{i=1}^{n}\varepsilon_i X_i(\hat{\beta} - \beta_0) \le \sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|) - \sum_{j=1}^{r} p_{\lambda_n}(|\hat{\beta}_j|).$$

Define $\delta_n = n^{1/2}\Sigma_n^{1/2}(\hat{\beta} - \beta_0)$, and $D_n = n^{-1/2}\Sigma_n^{-1/2}X$. Then

$$\sum_{i=1}^{n}[X_i(\hat{\beta} - \beta_0)]^2 - 2\sum_{i=1}^{n}\varepsilon_i X_i(\hat{\beta} - \beta_0) = \delta_n'\delta - 2(D_n'\varepsilon)'\delta_n$$

$$= \|\delta_n - D_n'\varepsilon\|_2^2 - \|D_n'\varepsilon\|_2^2.$$

Therefore,

$$\|\delta_n - D_n'\varepsilon\|_2^2 - \|D_n'\varepsilon\|_2^2 \le \sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|) - \sum_{j=1}^{r} p_{\lambda_n}(|\hat{\beta}_j|). \tag{13}$$

Now use the inequality $(b-a)^2 \ge 0.5b^2 - a^2$ to get

$$\|\delta_n - D_n'\varepsilon\|_2^2 \ge 0.5\|\delta_n\|_2^2 - \|D_n'\varepsilon\|_2^2.$$

Substituting this inequality into (13) and rearranging terms gives

$$0.5\|\delta_n\|_2^2 \le 2\|D_n'\varepsilon\|_2^2 + \sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|) - \sum_{j=1}^{r} p_{\lambda_n}(|\hat{\beta}_j|).$$

Now $E\|D_n'\varepsilon\|_2^2 = \sigma^2 r$, where $\sigma^2 = E(\varepsilon^2)$. Moreover, $E\|\delta_n\|_2^2 = nE(\hat{\beta} - \beta_0)'\Sigma_n(\hat{\beta} - \beta_0)$.

Therefore,

$$E(\hat{\beta} - \beta_0)'\Sigma_n(\hat{\beta} - \beta_0) \le 4n^{-1}\sigma^2 r + 2n^{-1}E\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|) - p_{\lambda_n}(|\hat{\beta}_j|)]. \tag{14}$$

In particular,

$$E(\hat{\beta} - \beta_0)'\Sigma_n(\hat{\beta} - \beta_0) \le 4n^{-1}\sigma^2 r + 2n^{-1}\sum_{j=1}^{r} p_{\lambda_n}(|\beta_{0j}|).$$

But $p_\lambda = \lambda f$ by (A6), so,

$$E(\hat{\beta}-\beta_0)'\Sigma_n(\hat{\beta}-\beta_0) \le 2n^{-1}\sigma^2 r + 2\lambda_n n^{-1}\sum_{j=1}^{r} f(|\beta_{0j}|)$$

$$= O(n^{-1} + \lambda_n n^{-1}).$$

It follows that

$$E\left\|\hat{\beta}-\beta_0\right\|_2^2 = O\left(\frac{1+\lambda_n}{n}\right).$$

In addition, it follows from Markov's inequality, that for each $\varepsilon > 0$ there is an $M_\varepsilon < \infty$ such that

$$P\left[\left(\frac{n}{1+\lambda_n}\right)\left\|\hat{\beta}-\beta_0\right\|_2^2 \le M_\varepsilon\right] \ge 1-\varepsilon\,.$$

Step 2 of the proof consists of refining the result of step 1 to show that $\hat{\beta}_j$ is $n^{-1/2}$-consistent for β_{0j}. Let $A_s^* = \{j : j \in A_s;\ p\lim_{n\to\infty} |\tilde{\beta}_j| \ne 0\}$.

Now

$$p_{\lambda_n}(|\beta_{0j}|) - p_{\lambda_n}(|\hat{\beta}_j|) = \lambda_n[f(|\beta_{0j}|) - f(|\hat{\beta}_j|)]\,.$$

If (A7)(ii)(a) holds and $j \in A_0$, then it follows from step 1 that with probability approaching 1 as $n \to \infty$

$$|p_{\lambda_n}(|\beta_{0j}|) - p_{\lambda_n}(|\hat{\beta}_j|)| = 0\,.$$

If (A7)(ii)(b) holds and $j \in A_0$, then

$$|p_{\lambda_n}(|\beta_{0j}|) - p_{\lambda_n}(|\hat{\beta}_j|)| \le C\lambda_n|\hat{\beta}_j - \beta_{0j}|\,.$$

Therefore, if (A7)(ii)(a) holds,

$$\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]=\sum_{j\in A_s^*}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]$$

$$\leq \sum_{j\in A_s^*} p_{\lambda_n}(|\beta_{0j}|)$$

$$=\lambda_n \sum_{j\in A_s^*} f(|\beta_{0j}|)$$

$$\leq C\lambda_n \sum_{j\in A_s^*} |\beta_{0j}|$$

$$=o(\lambda_n n^{-\theta}).$$

with probability approaching 1. If (A7)(ii)(b) holds, then

$$\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]\leq \sum_{j\in A_0}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]+\sum_{j\in A_s^*}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]$$

$$\leq \sum_{j\in A_0}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]+\sum_{j\in A_s^*} p_{\lambda_n}(|\beta_{0j}|).$$

Therefore,

$$\left|\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]\right|\leq C\lambda_n \sum_{j\in A_0}|\hat{\beta}_j-\beta_{j0}|+C\lambda_n\sum_{j\in A_s^*}|\beta_{0j}|^{\gamma}$$

$$=C\lambda_n \sum_{j\in A_0}|\hat{\beta}_j-\beta_{j0}|+o(\lambda_n n^{-1/2})$$

if (A7)(ii)(b) holds. The Cauchy-Schwarz inequality gives

$$\sum_{j\in A_0}|\hat{\beta}_j-\beta_{j0}|,\ \sum_{j\in A_s^*}|\hat{\beta}_j-\beta_{j0}|\leq r^{1/2}\left\|\hat{\beta}-\beta_0\right\|_2.$$

Therefore,

$$\left|\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]\right|=o(\lambda_n n^{-\theta})$$

with probability approaching 1 if (A7)((ii)(a) holds, and

$$\left|\sum_{j=1}^{r}[p_{\lambda_n}(|\beta_{0j}|)-p_{\lambda_n}(|\hat{\beta}_j|)]\right|\leq C\lambda_n r^{1/2}\left\|\hat{\beta}_j-\beta_{0j}\right\|_2+o(\lambda_n n^{-1/2})$$

if (A7)(ii)(b) holds. Substituting these inequalities into (14) yields

$$E(\hat{\beta}-\beta_0)'\Sigma_n(\hat{\beta}-\beta_0)\le 4n^{-1}\sigma^2 r+o(\lambda_n n^{-1-\theta})$$

for all sufficiently large n if (A7)(ii)(a) holds and

$$E(\hat{\beta}-\beta_0)'\Sigma_n(\hat{\beta}-\beta_0)\le 2n^{-1}\sigma^2 r+Cn^{-1}\lambda_n r^{1/2}\left\|\hat{\beta}_j-\beta_{0j}\right\|_2+o(\lambda_n n^{-3/2})$$

if (A7)(ii)(b) holds. Now $E\left\|\hat{\beta}-\beta_0\right\|_2\le\left(E\left\|\hat{\beta}-\beta_0\right\|_2^2\right)^{1/2}$ by the Cauchy-Schwarz inequality. This combined with non-singularity of Σ implies that

$$E(\hat{\beta}-\beta_0)'\Sigma_n(\hat{\beta}-\beta_0)\ge cE\left\|\hat{\beta}-\beta_0\right\|_2^2$$

for some constant $c>0$. Therefore,

(15a) $$E\left\|\hat{\beta}-\beta_0\right\|_2^2\le\tilde{C}n^{-1}+o(\lambda_n n^{-1-\theta})$$

for all sufficiently large n and some $\tilde{C}<\infty$ if (A7)(ii)(a) holds and

(15b) $$E\left\|\hat{\beta}-\beta_0\right\|_2^2\le\tilde{C}n^{-1}+\tilde{C}n^{-1}\lambda_n\left\|\hat{\beta}_j-\beta_{0j}\right\|_2+o(\lambda_n n^{-3/2})$$

if (A7)(ii)(a) holds. Inequalities (15a) and (15b) imply that

$$E\left\|\hat{\beta}-\beta_0\right\|=O(n^{-1/2}).$$

The third step of the proof consists of showing that with probability approaching 1 as $n\to\infty$, all the large β_j's and none of the small ones are selected. Let $\hat{\beta}=(\hat{\beta}_1',\hat{\beta}_2')'$, where $\hat{\beta}_1$ is the second-stage estimator of the large coefficients and $\hat{\beta}_2$ is the second stage estimator of the small ones. We have $\left\|\hat{\beta}-\beta_0\right\|\le n^{-1/2}C_\varepsilon$ with probability at least $1-\varepsilon$ for any $\varepsilon>0$ and all sufficiently large C_ε. Let $\beta_{1n}=\beta_{01}+n^{-1/2}u_1$ and $\beta_{2n}=\beta_{02}+n^{-1/2}u_2$, where β_{01} and β_{02} are the true values of the large and small coefficients, respectively, and $\|u\|^2=\|u_1\|^2+\|u_2\|^2\le C_\varepsilon^2$. Define

$$V_n(u_1,u_2)=S_n(\beta_{1n},\beta_{2n})-S_n(\beta_{01},0).$$

Then $(\hat{\beta}_{1n}',\hat{\beta}_{2n}')$ minimizes $V_n(u_1,u_2)$ over $\|u\|\le C_\varepsilon$ with probability at least $1-\varepsilon$. Define $u_{20}=-n^{1/2}\beta_{02}$. It follows from $n^{-1/2}$ consistency of the $\hat{\beta}_j$'s that all the large β_j's are chosen with probability approaching 1 as $n\to\infty$. Therefore, it suffices to show that

$$V_n(u_1,u_2)-V_n(u_1,0)\ge 0$$

with probability at least $1-\varepsilon$ if $u_2 \neq 0$. Write $x=(w,z)$, where w corresponds to covariates with large coefficients and z corresponds to covariates with small ones. Then

$$V_n(u_1,u_2)-V_n(u_1,0)=n^{-1}\sum_{i=1}^{n}[(z_iu_2)^2-(z_iu_{20})^2]+2n^{-1}\sum_{i=1}^{n}(w_iu_1)z_i(u_2-u_{20})$$

$$-2n^{-1/2}\sum_{i=1}^{n}\varepsilon_i z_i(u_2-u_{20})+\Delta$$

$$\equiv R_{n1}+R_{n2}+R_{n3}+\Delta,$$

where

$$\Delta=\sum_{j=k+1}^{p}[p_{\lambda_n}(|\beta_{2j}|)-p_{\lambda_n}(|\beta_{02,j}|)]$$

$$=\sum_{j=k+1}^{p}[p_{\lambda_n}(|\beta_{02,j}+n^{-1/2}u_2|)-p_{\lambda_n}(|\beta_{02,j}|)].$$

Now $u_{20}=o(1)$, so

$$R_{n1}=n^{-1}\sum_{i=1}^{n}(z_iu_2)^2+o(1)$$

$$R_{n2}=2n^{-1}\sum_{i=1}^{n}(w_iu_1)z_iu_2+o(1),$$

and

$$R_{n3}=-2n^{-1/2}\sum_{i=1}^{n}\varepsilon_i z_iu_2+o(1).$$

As in Huang, Horowitz, and Ma (2008), $R_{n1}+R_{n2}\geq -C$ for some constant $C<\infty$, and $R_{n3}=O_p(1)$. Therefore,

$$V_n(u_1,u_2)-V_n(u_1,0)\geq -C+O(1)+\Delta$$

$$=-C+O(1)+\lambda_n\sum_{j\in A_s^*}[f(|\beta_{02,j}+n^{-1/2}u_{2j}|)-f(|\beta_{02,j}|)].$$

Therefore,

(15a) $$V_n(u_1,u_2)-V_n(u_1,0)\geq -C+O(1)+C_1\lambda_n n^{-1/2}\sum_{j\in A_s^*}|u_{2j}|$$

for all sufficiently large n under (A6)(ii)(a), and

(15b) $V_n(u_1,u_2)-V_n(u_1,0)\geq -C+O(1)+C_1\lambda_n n^{-\gamma/2}\sum_{j\in A_3^*}|u_{2j}|$

under (A6)(ii)(b), where C_1 is a constant. The right-hand sides if (15a) and (15b) increase without bound as $n\to\infty$. ■

Proof of Theorem 3:

The proofs of Theorem 3(i) and 3(ii) are identical to the proof of Theorem 1(i) and 1(ii) in Huang, Horowitz, and Wei (2010). To prove Theorem 3(iii), define η_n to be the $n\times 1$ vector whose i'th component is $\eta_i = Y_i - Z_i\beta_n$, where β_n is the $pm_n\times 1$ vector of stacked β_{nj}'s. Define $f_{0A_2}(X_i)=\sum_{j\in A_2} f_j(X_i)$, $f_{nA_2}(X_i)=\sum_{j\in A_2} Z_{ij}'\beta_{nj}$, and $f_{n\bar{A}_2}(X_i)=\sum_{j\in \bar{A}_2} Z_{ij}'\beta_{nj}$. Proceed as in the proof of Theorem 1(iii) of Huang, Horowitz, and Wei (2010) to obtain

$$\eta_i = Y_i-\mu-f_0(X_i)-\bar{Y}-\sum_{j\in A_2} Z_{ij}'\beta_{nj}$$

$$=\varepsilon_i+\mu+[f_{A_2}(X_i)-f_{nA_2}(X_i)]+f_{\bar{A}_2}(X_i).$$

Now proceed as in Huang, Horowitz, and Wei (2010) to obtain

$$\left\|\tilde{\beta}_{nA_2}-\beta_{nA_2}\right\|_2^2 = O_p\left[\frac{m_n^2\log(pm_n)}{n}\right]+O_p\left(\frac{m_n}{n}\right)+O_p\left(\frac{m_n^2\lambda_n^2}{n^2}\right)$$

$$+O\left(\frac{1}{m_n^{2d-1}}\right)+O_p\left(\frac{m_n}{n}\left\|f_{\bar{A}_2}\right\|_2^2\right).$$

The last term on the right-hand side is asymptotically negligible if $m_n \asymp n^{1/(2d+1)}$, which gives part (iii) of the theorem. ■

REFERENCES

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations (with discussion), *Journal of the American Statistical Association*, 96, 939-967.

Bickel, P.J., Y. Ritov, A.B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics*. 37, 1705-1732.

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

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

Fan, J., H. Peng, and T. Huang (2005). Semilinear high-dimensional model for normalization of microarray data, *Journal of the American Statistical Association*, 100, 781-796.

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparameterization, *Bernoulli*, 10, 971-988.

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

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

Knight, K. and W.J. Fu (2000). Asymptotics for lasso-type estimators, *Annals of Statistics*, 28, 1356-1378.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares, *Annals of Statistics*, 37, 3498-3528.

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

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

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

Zhang, C.-H. (2010) Nearly unbiased variable selection under minmax concave penalty, *Annals of Statistics*, 38, 894-932.

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

Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 penalization, *Annals of Statistics*, 37, 2109-2144.

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

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

Zou, H. and H.H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters, *Annals of Statistics*, 37, 1733–1751.

TABLE 1: Mean Square Errors of OLS Estimates of β_1 from Full and Reduced Models

d	Mean-Square Error of Estimate of β_1	
	Full Model	Reduced Model
2	0.67	0.22
4	0.67	0.19
6	0.67	0.16

TABLE 2: Results of LASSO and Adaptive LASSO Estimation

d	MSE of $\hat{\beta}_1$	Average Size of Selected Model	Prob. that Selected Model Contains Large Variables	Average λ
		LASSO		
2	0.31	9.7	0.84	4.57
4	0.34	12.3	0.75	4.52
6	0.29	15.0	0.67	4.48
		ADAPTIVE LASSO		
2	0.31	6.5	0.67	3.10
4	0.30	8.8	0.56	3.29
6	0.32	10.8	0.39	3.44

TABLE 3: Results of LASSO and Adaptive LASSO Estimation with β_1 Not in the Penalty Function

d	MSE of $\hat{\beta}_1$	Average Size of Selected Model	Prob. that Selected Model Contains Large Variables	Average λ
		LASSO		
2	0.27	7.9	0.88	4.66
4	0.29	10.6	0.81	4.64
6	0.40	13.3	0.67	4.58
		ADAPTIVE LASSO		
2	0.19	5.8	0.67	2.83
4	0.17	8.0	0.64	3.13
6	0.19	10.2	0.43	3.23

TABLE 4: Results of Estimating Effects of Union Membership and Marital Status on Wages

Variable	Coefficient (Standard Error) Obtained from OLS	LASSO	Adaptive LASSO
Union Member	0.21 (0.17)	0.21 (0.096)	0.22 (0.094)
Marital Status	0.051 (0.19)	0.19 (0.11)	0.20 (0.11)