ADAPTIVE LASSO FOR SPARSE HIGH-DIMENSIONAL REGRESSION MODELS

Jian Huang[1], Shuangge Ma[2], and Cun-Hui Zhang[3]

[1]University of Iowa, [2]Yale University, [3]Rutgers University

November 2006

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 374

Revision 1 (June 2007)

Summary. We study the asymptotic properties of the adaptive Lasso estimators in sparse, high-dimensional, linear regression models when the number of covariates may increase with the sample size. We consider variable selection using the adaptive Lasso, where the L_1 norms in the penalty are re-weighted by data-dependent weights. We show that, if a reasonable initial estimator is available, then under appropriate conditions, the adaptive Lasso correctly selects covariates with nonzero coefficients with probability converging to one and that the estimators of nonzero coefficients have the same asymptotic distribution that they would have if the zero coefficients were known in advance. Thus, the adaptive Lasso has an oracle property in the sense of Fan and Li (2001) and Fan and Peng (2004). In addition, under a partial orthogonality condition in which the covariates with zero coefficients are weakly correlated with the covariates with nonzero coefficients, marginal regression can be used to obtain the initial estimator. With this initial estimator, adaptive Lasso has the oracle property even when the number of covariates is much larger than the sample size.

Key Words and phrases. Penalized regression, high-dimensional data, variable selection, asymptotic normality, oracle property, zero-consistency.

Short title. Adaptive Lasso

AMS 2000 subject classification. Primary 62J05, 62J07; secondary 62E20, 60F05

1 Introduction

Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \in \mathrm{I\!R}^n, \tag{1}$$

where $\mathbf{X}$ is an $n \times p_n$ design matrix, $\boldsymbol{\beta}$ is a $p_n \times 1$ vector of unknown coefficients, and $\boldsymbol{\varepsilon}$ is a vector of i.i.d. random variables with mean zero and finite variance σ^2. We note that p_n, the length of $\boldsymbol{\beta}$, may depend on the sample size n. We assume that the response and covariates are centered, so the intercept term is zero. We are interested in estimating $\boldsymbol{\beta}$ when p_n is large or even larger than n and the regression parameter is sparse in the sense that many of its elements are zero. Our motivation comes from studies that try to correlate a certain phenotype with high-dimensional genomic data. With such data, the dimension of the covariate vector can be much larger than the sample size. The traditional least squares method is not applicable, and regularized or penalized methods are needed. The Lasso (Tibshirani, 1996) is a penalized method similar to the ridge regression (Hoerl and Kennard, 1970) but uses the L_1 penalty $\sum_{j=1}^{p_n} |\beta_j|$ instead of the L_2 penalty $\sum_{j=1}^{p_n} \beta_j^2$. So the Lasso estimator is the value that minimizes

$$\left\|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\right\|^2 + 2\lambda \sum_{j=1}^{p_n} |\beta_j|, \tag{2}$$

where λ is the penalty parameter. An important feature of the Lasso is that it can be used for variable selection. Compared to the classical variable selection methods such as subset selection, the Lasso has two advantages. First, the selection process in the Lasso is continuous and hence more stable than the subset selection. Second, the Lasso is computationally feasible for high-dimensional data. In contrast, computation in subset selection is combinatorial and not feasible when p_n is large.

Several authors have studied the properties of the Lasso. When p_n is fixed, Knight and Fu (2001) showed that, under appropriate conditions, the Lasso is consistent for estimating the regression parameter and its limiting distributions can have positive probability mass at 0 when the true value of the parameter is zero. Leng, Lin and Wahba (2005) showed that the Lasso is in general not path consistent in the sense that (a) with probability greater than zero, the whole Lasso path may not contain the true parameter value; (b) even if the true parameter value is contained in the Lasso path, it cannot be achieved by using prediction accuracy as the selection criterion. For fixed p_n, Zou (2006) further studied the variable selection and estimation properties of the Lasso. He showed that the positive probability mass at 0 of a Lasso estimator, when the true value of the parameter is 0, is in general less than 1, which implies that the Lasso is in general not variable selection consistent. He also provided a condition on the design matrix for the Lasso to be variable selection consistent. This condition was discovered by Meinshausen and Buhlmann (2006) and Zhao and Yu (2007). In particular, Zhao and Yu (2007) called this condition the irrepresentable condition on the design matrix. Meinshausen and Buhlmann (2006) and Zhao and Yu (2007) allowed the number of variables go to infinity faster than n. They showed that under the irrepresentable condition, the Lasso is consistent for variable selection, provided that p_n is not too large and the penalty parameter λ grows faster than $\sqrt{n \log p_n}$. Specifically, p_n is allowed to be as large as $\exp(n^a)$ for some $0 < a < 1$ when the errors have Gaussian tails. However, the value of λ required for variable selection consistency over shrinks the nonzero coefficients, which leads to asymptotically biased estimates. Thus the Lasso is variable-selection consistent under certain conditions, but not in general. Moreover, if the Lasso is variable-selection consistent, then it is not efficient for estimating the nonzero parameters. Therefore, these studies confirm the suggestion that the Lasso does not possess the oracle property (Fan and Li 2001, Fan and Peng 2004). Here the oracle property of a method means that it can correctly select the nonzero coefficients with probability converging to one and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariance that they would have if the zero coefficients were known in advance. On the other hand, Greenshtein and Ritov (2004) showed that the Lasso has certain persistence property for prediction, and under a sparse Riesz condition, Zhang and Huang (2006) proved that the Lasso possesses the right order of sparsity and selects all coefficients of greater order than $(\lambda/n)\sqrt{k_n}$,where k_n is the number of nonzero coefficients.

In addition to the Lasso, other penalized methods have been proposed for the purpose of simultaneous variable selection and shrinkage estimation. Examples include the bridge penalty (Frank and Friedman 1996) and the SCAD penalty (Fan 1997; Fan and Li, 2001). For the SCAD penalty, Fan and Li (2001) and Fan and Peng (2004) studied asymptotic properties of penalized likelihood methods. They showed that there exist local maximizers of the penalized likelihood that have the oracle property. Huang, Horowitz and Ma (2006) showed that the bridge estimator in a linear regression model has the oracle property under appropriate conditions, if the bridge index is strictly between 0 and 1. Their result also permits a divergent number of regression coefficients. While the SCAD and bridge estimators enjoy the oracle property, the objective functions with the SCAD and bridge penalties are not convex, so it is more difficult to compute these estimators. Another interesting estimator, the Dantzig selector, in high-dimensional settings was proposed and studied for the estimation of $\boldsymbol{\beta}$ by Candes and Tao (2005). This estimator achieves a loss within a logarithmic factor of the ideal mean squared error and can be solved by a convex minimization problem.

An approach to obtaining a convex objective function which yields oracle estimators is by using a weighted L_1 penalty with weights determined by an initial estimator (Zou (2006)). Suppose that an initial estimator $\widetilde{\boldsymbol{\beta}}_n$ is available. Let

$$w_{nj} = |\widetilde{\beta}_{nj}|^{-1}, \quad j = 1, \ldots, p_n. \tag{3}$$

Denote

$$L_n(\boldsymbol{\beta}) = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + 2\lambda_n \sum_{j=1}^{p_n} w_{nj} |\beta_j|. \tag{4}$$

The value $\widehat{\boldsymbol{\beta}}_n$ that minimizes L_n is called the adaptive Lasso estimator (Zou (2006)). By allowing relatively higher penalty for zero coefficients and lower penalty for nonzero coefficients, the adaptive Lasso hopes to reduce the estimation bias and improve variable selection accuracy, compared with the standard Lasso.

For fixed p_n, Zou (2006) proved that the adaptive Lasso has the oracle property. We consider the case when $p_n \to \infty$ as $n \to \infty$. We show that the adaptive Lasso has the oracle property under an adaptive irrepresentable and other regularity conditions and in particular, this can be achieved with marginal regression as the initial estimates under a partial orthogonal condition on the covariates. This result allows $p_n = O(\exp(n^a))$ for some constant $0 < a < 1$, where a depends on the regularity conditions. Thus, the number of covariates can be larger than the sample size if a proper initial estimator is used in the adaptive Lasso.

When $p_n > n$, the regression parameter is in general not identifiable without further assumptions on the covariate matrix. However, if there is suitable structure in the covariate matrix, it is possible to achieve consistent variable selection and estimation. We consider a partial orthogonality condition in which the covariates with zero coefficients are only weakly correlated with the covariates with nonzero coefficients. We show that for $p_n \gg n$ and under the partial orthogonality and certain other conditions, the adaptive Lasso achieves selection consistency and estimation efficiency when the marginal regression estimators are used as the initial estimators, although they do not yield consistent estimation of the parameters. The partial orthogonality condition is reasonable in microarray data analysis, where the genes that are correlated with the phenotype of interest may be in different functional pathways from the genes that are not related to the phenotype (Bair et al. 2006). The partial orthogonality condition was also discussed in the context of bridge estimation by Huang et al. (2006). Fan and Lv (2006) studied univariate screening in high-dimensional regression problems and provided conditions under which it can be used to reduce the exponentially growing dimensionality of a model. A new contribution of the present article is that we also investigate the effect of the tail behavior of the error distribution on the property of the marginal regression estimators in high-dimensional settings.

The rest of the paper is organized as follows. In Section 2, we state the results on variable-selection consistency and asymptotic normality of the adaptive Lasso estimator. In Section 3, we show that under the partial orthogonality and certain other regularity conditions, marginal regression estimators can be used in the adaptive Lasso to yield the desirable selection and estimation properties. In Section 4, we present results from simulation studies and a real data example. Some concluding remarks are given in Section 5. The proofs of the results stated in Sections 2 and 3 are provided in the online supplement to this article.

2 Variable-selection consistency and asymptotic normality

Let the true parameter value be $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ with dimension $p = p_n$. For simplicity of notation, we write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'$, where $\boldsymbol{\beta}_{10}$ is a $k_n \times 1$ vector and $\boldsymbol{\beta}_{20}$ is a $m_n \times 1$ vector. Suppose that $\boldsymbol{\beta}_{10} \neq \mathbf{0}$ and $\boldsymbol{\beta}_{20} = \mathbf{0}$, where $\mathbf{0}$ is the vector (with appropriate dimension) with all components zero. So k_n is the number of non-zero coefficients and m_n is the number of zero coefficients in the regression model. We note that it is unknown to us which coefficients are non-zero and which are zero. Most quantities and data objects in our discussion are functions of n, but this dependence on n is often made implicit, especially for n-vectors and matrices with n rows.

We center the response $\mathbf{y} = (y_1, \ldots, y_n)'$ and standardize the covariates $\mathbf{X} = (x_{ij})_{n \times p_n}$ so that

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 = 1, \; j = 1, \ldots, p_n. \tag{5}$$

Let $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$ be the j-th column of the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{p_n})$, and $\mathbf{y} = (y_1, \ldots, y_n)'$. The regression model is written as

$$\mathbf{y} = \sum_{j=1}^{p_n} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

with the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Let $J_{n1} = \{j : \beta_{0j} \neq 0\}$ and define

$$\mathbf{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_{k_n}), \quad \Sigma_{n11} = n^{-1}\mathbf{X}_1'\mathbf{X}_1$$

Let τ_{n1} be the smallest eigenvalue of Σ_{n11}. For any vector $\mathbf{x} = (x_1, x_2, \ldots)'$, denote its sign vector by $\text{sgn}(\mathbf{x}) = (\text{sgn}(x_1), \text{sgn}(x_2), \ldots)'$, with the convention $\text{sgn}(0) = 0$. Following Zhao and Yu (2007), we say that $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}$ if and only if $\text{sgn}(\widehat{\boldsymbol{\beta}}_n) = \text{sgn}(\boldsymbol{\beta})$. Let

$$b_{n1} = \min\{|\beta_{0j}| : j \in J_{n1}\}. \tag{7}$$

We assume the following conditions.

(A1) The errors $\varepsilon_i, \varepsilon_2, \ldots$ are independent and identically distributed random variables with mean zero and that for certain constants $1 \le d \le 2$, $C > 0$ and K, the tail probabilities of ε_i satisfy $P(|\varepsilon_i| > x) \le K\exp(-Cx^d)$ for all $x \ge 0$ and $i = 1, 2, \ldots$

(A2) The initial estimators $\widetilde{\beta}_{nj}$ are r_n-consistent for the estimation of certain η_{nj}:

$$r_n \max_{j \le p_n} \left| \widetilde{\beta}_{nj} - \eta_{nj} \right| = O_P(1), \quad r_n \to \infty,$$

where η_{nj} are unknown constants depending on $\boldsymbol{\beta}$ and satisfy

$$\max_{j \notin J_{n1}} |\eta_{nj}| \le M_{n2}, \quad \Big\{ \sum_{j \in J_{n1}} \Big(\frac{1}{|\eta_{nj}|} + \frac{M_{n2}}{|\eta_{nj}|^2} \Big)^2 \Big\}^{1/2} \le M_{n1} = o(r_n).$$

(A3) Adaptive irrepresentable condition: For $\mathbf{s}_{n1} = \big(|\eta_{nj}|^{-1}\mathrm{sgn}(\beta_{0j}), j \in J_{n1}\big)'$ and some $\kappa < 1$

$$n^{-1}\Big|\mathbf{x}_j'\mathbf{X}_1\Sigma_{n11}^{-1}\mathbf{s}_{n1}\Big| \le \kappa/|\eta_{nj}|, \quad \forall\, j \notin J_{n1}.$$

(A4) The constants $\{k_n, m_n, \lambda_n, M_{n1}, M_{n2}, b_{n1}\}$ satisfy the following condition

$$(\log n)^{I\{d=1\}}\Big\{ \frac{(\log k_n)^{1/d}}{n^{1/2}b_{n1}} + (\log m_n)^{1/d}\frac{n^{1/2}}{\lambda_n}\Big(M_{n2} + \frac{1}{r_n}\Big)\Big\} + \frac{M_{n1}\lambda_n}{b_{n1}n} \to 0.$$

(A5) There exists a constant $\tau_1 > 0$ such that $\tau_{n1} \ge \tau_1$ for all n.

Condition (A1) is standard for variable selection in linear regression. Condition (A2) assumes that the initial $\widetilde{\beta}_{nj}$ actually estimates some proxy η_{nj} of β_{nj} so that the weight $w_{nj} \approx |\eta_{nj}|^{-1}$ is not too large for $\beta_{0j} \ne 0$ and not too small for $\beta_{0j} = 0$. The adaptive irrepresentable condition (A3) becomes the strong irrespresentable condition for the sign-consistency of the Lasso if $|\eta_{nj}|$ are identical for all $j \le p_n$. It weakens the strong irrepresentable condition by allowing larger $|\eta_{nj}|$ in J_{n1} (smaller $\mathbf{s}_{n1}$) and smaller $|\eta_{nj}|$ outside J_{n1}. If $\mathrm{sgn}(\eta_{nj}) = \mathrm{sgn}(\beta_{nj})$ in (A2), we say that the initial estimates are *zero-consistent with rate* r_n. In this case, (A3) holds automatically and $M_{n2} = 0$ in (A2).

Condition (A4) restricts the numbers of covariates with zero and nonzero coefficients, the penalty parameter, and the smallest non-zero coefficient. The number of covariates permitted depends on the tail behavior of the error terms. For sub-Gaussian tail, the model can include more covariates, while for exponential tail, the number of covariates allowed is fewer. We often have $n^{\delta-1/2}r_n \to \infty$ and $\lambda_n = n^a$ for some $0 < a < 1$ and small $\delta > 0$. In this case, the number m_n of zero coefficients can be as large as $\exp(n^{d(a-\delta)})$. But the number of nonzero coefficients allowed is of the order $\min\{n^{2(1-a)}, n^{1-2\delta}\}$, assuming $1/b_{n1} = O(1)$ and $M_{n1} = O(k_n^{1/2})$. Condition (A5) assumes that the eigenvalues of Σ_{n11} are bounded away from zero. This is reasonable since the number of nonzero covariates is small in a sparse model.

Among conditions (A1) to (A5), (A3) is the most critical one and is in general difficult to establish. It assumes that we can estimate certain η_{nj} satisfying the condition. On the other hand, this task can be reduced to establishing the simpler and stronger properties under a partial orthogonality condition described in Section 3.

Theorem 1 *Suppose that conditions (A1)-(A5) hold. Then*

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

The proof of this theorem can be found in the on-line supplement to this article.

Theorem 2 *Suppose that conditions (A1) to (A5) are satisfied. Let* $s_n^2 = \sigma^2 \boldsymbol{\alpha}_n' \Sigma_{n11}^{-1} \boldsymbol{\alpha}_n$ *for any* $k_n \times 1$ *vector* $\boldsymbol{\alpha}_n$ *satisfying* $\|\boldsymbol{\alpha}_n\|_2 \le 1$. *If* $M_{n1}\lambda_n/n^{1/2} \to 0$

$$n^{1/2} s_n^{-1} \boldsymbol{\alpha}_n'(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_0) = n^{-1/2} s_n^{-1} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{\alpha}_n' \Sigma_{n11}^{-1} \mathbf{x}_{1i} + o_p(1) \to_D N(0,1), \tag{8}$$

where $o_p(1)$ *is a term that converges to zero in probability uniformly with respect to* $\boldsymbol{\alpha}_n$.

This theorem can be proved by verifying the Lindeberg conditions the same way as in the proof of Theorem 2 of Huang et al. (2006). Thus we omit the proof here.

3 Zero-consistency, partial orthogonality and marginal regression

For the adaptive Lasso estimator to be variable selection consistent and have the oracle property, it is crucial to have an initial estimator that is zero-consistent or satisfies the weaker condition

(A3). When $p_n \leq n$, the least squares estimator is consistent and therefore zero-consistent under certain conditions on the design matrix and regression coefficients. In this case, we can use the least squares estimator as the initial estimators for the weights. However, when $p_n > n$, which is the case in many microarray gene expression studies, the least squares estimator is no longer feasible. In this section, we show that the marginal regression estimators are zero-consistent under a partial orthogonality condition.

With the centering and scaling given in (5), the estimated marginal regression coefficient is

$$\widetilde{\beta}_{nj} = \frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2} = \mathbf{x}_j' \mathbf{y}/n. \tag{9}$$

We take the η_{nj} in (A2) to be $\mathrm{E}\widetilde{\beta}_{nj}$. Since $\boldsymbol{\mu}_0 = \mathrm{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0$,

$$\eta_{nj} = \mathrm{E}\widetilde{\beta}_{nj} = \mathbf{x}_j' \boldsymbol{\mu}_0/n = \sum_{l=1}^{k_n} \beta_{0l} \mathbf{x}_j' \mathbf{x}_l/n. \tag{10}$$

It is also possible to consider $\widetilde{\beta}_{nj} = |\mathbf{x}_j' \mathbf{y}/n|^\gamma$ and $\eta_{nj} = |\mathbf{x}_j' \boldsymbol{\mu}_0/n|^\gamma$ with certain $\gamma > 0$, but we focus on the simpler (9) and (10) here.

We make the following assumptions:

(B1) The condition (A1) holds.

(B2) (Partial orthogonality) The covariates with zero coefficients and those with nonzero coefficients are only weakly correlated

$$\left| \frac{1}{n} \sum_{i=1}^n x_{ij} x_{ik} \right| = \left| \mathbf{x}_j' \mathbf{x}_k/n \right| \leq \rho_n, \quad j \notin J_{n1}, \ k \in J_{n1},$$

where for certain $0 < \kappa < 1$, ρ_n satisfies

$$c_n = \Big(\max_{j \notin J_{n1}} |\eta_{nj}| \Big) \Big(\sum_{j \in J_{n1}} \frac{|\eta_{nj}|^{-2}}{k_n} \Big)^{1/2} \leq \frac{\kappa \tau_{n1}}{k_n \rho_n}, \tag{11}$$

where κ is given in (A3).

(B3) The minimum $\widetilde{b}_{n1} = \min\{|\eta_{nj}|, j \in J_{n1}\}$ satisfies

$$\frac{k_n^{1/2}(1+c_n)}{\widetilde{b}_{n1} r_n} \to 0, \quad r_n = \frac{n^{1/2}}{(\log m_n)^{1/d}(\log n)^{I\{d=1\}}}.$$

Condition (B2) is the weak partial orthogonality assumption, which requires that the covariates with zero coefficients have weaker correlation to the mean $\boldsymbol{\mu}_0 = E\mathbf{y}$ than those with nonzero coefficients in an average sense. For $k_n \rho_n \leq \kappa \tau_{n1}$, (B2) holds for the Lasso with $\eta_{nj} = 1$. Thus, the adaptive Lasso has advantages only when $c_n < 1$. Condition (B3) requires that the non-zero coefficients are bounded away from zero at certain rates depending on the growth of k_n and m_n.

Theorem 3 *Suppose that conditions (B1) to (B3) hold. Then (A2) and (A3) hold for the η_{nj} in (10), i.e. the $\widetilde{\boldsymbol{\beta}}_n$ in (9) is r_n-consistent for η_{nj} and the adaptive irrepresentable condition holds.*

The proof of this theorem is given in the on-line supplement to this article.

Theorem 3 provides justification for using marginal regression estimator for adaptive Lasso as the initial estimator under the partial orthogonality condition. Under (B1)-(B3), (A4) follows from

(B4) Let $b_{n2} = O(1)$. Then $b_{n1} \leq |\beta_{0j}| \leq b_{n2}$ $\forall j \in J_{n1}$ and

$$\frac{(\log k_n / \log m_n)^{1/d}}{r_n b_{n1}} + \frac{n}{\lambda_n r_n}(k_n \rho_n + 1/r_n) + \frac{k_n^{1/2}\lambda_n}{n b_{n1} \widetilde{b}_{n1}} \to 0.$$

Thus, under (B1)-(B4) and (A5), we can first use the marginal regression to obtain the initial estimators, and use them as weights in the adaptive Lasso to achieve variable-selection consistency and oracle efficiency.

A special case of Theorem 3 is when $\rho_n = O(n^{-1/2})$, that is, the covariates with nonzero and zero coefficients are essentially uncorrelated. Then we can take $\eta_{nj} = 0, j \notin J_{n1}$ and (11) is satisfied. Consequently, the univariate regression estimator $\widetilde{\boldsymbol{\beta}}_n$ in (9) is zero-consistent with rate r_n. In this case, the adaptive irrepresentable condition (A3) is automatically satisfied.

4 Numerical Studies

We conduct simulation studies to evaluate the finite sample performance of the adaptive Lasso estimate and use a real data example to illustrate the application of this method. Because our main interest is in when p_n is large and Zou (2006) has conducted simulation studies of adaptive Lasso in low dimensional settings, we focus on the case when $p_n > n$.

4.1 Simulation study

The adaptive Lasso estimate can be computed by a simple modification of the LARS algorithm (Efron et al. 2004). The computational algorithm is omitted here. In simulation study, we are interested in (1) accuracy of variable selection and (2) prediction performance measured by mse (mean squared error). For (1), we compute the frequency of correctly identifying zero and nonzero coefficients in repeated simulations. For (2), we compute the median prediction mse which is calculated based on the predicted and observed values of the response from independent data not used in model fitting. We also compare the results from the adaptive Lasso to those from the standard Lasso estimate.

We simulate data from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n).$$

Eight examples with $p_n > n$ are considered. In each example, the covariate vector is generated as normal distributed with mean zero and covariance matrix specified below. The value of $\mathbf{X}$ is generated once and then kept fixed. Replications are obtained by simulating the values of $\boldsymbol{\varepsilon}$ from $N(0, \sigma^2 \mathbf{I}_n)$ and then setting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for the fixed covariate value $\mathbf{X}$. The sample size used in estimation is $n = 100$. Summary statistics are computed based on 500 replications.

The eight examples we consider are

1. $p = 200$ and $\sigma = 1.5$. The n-rows of $\mathbf{X}$ are independent. For the i-th row, the first 15 covariates $(x_{i,1}, \ldots, x_{i,15})$ and the remaining 185 covarites $(x_{i,16}, \ldots, x_{i,200})$ are independent. The pairwise correlation between the k^{th} and the j^{th} components of $(x_{i,1}, \ldots, x_{i,15})$ is $r^{|k-j|}$

with $r = 0.5$, $k, j = 1, \ldots, 15$. The pairwise correlation between the k^{th} and the j^{th} components of $(x_{i,16}, \ldots, x_{i,200})$ is $r^{|k-j|}$ with $r = 0.5$, $k, j = 16, \ldots, 200$. The first 5 components of $\boldsymbol{\beta}$ are 2.5, components 6–10 are 1.5, components 11–15 are 0.5, and the rest are zero. The covariate matrix has the partial orthogonal structure.

2. The same as Example 1, except that $r = 0.95$.

3. The same as Example 1, except that $p = 400$.

4. The same as Example 2, except that $p = 400$.

5. $p = 200$ and $\sigma = 1.5$. The predictors are generated as follows: $x_{ij} = Z_{1j} + e_{ij}, i = 1, \ldots, 5$, $x_{ij} = Z_{2j} + e_{ij}, i = 6, \ldots, 10$, $x_{ij} = Z_{3j} + e_{ij}, i = 11, \ldots, 15$, and $x_{ij} = Z_{ij}$, where Z_{ij} are iid $N(0, 1)$ and e_{ij} are i.i.d $N(0, 1/100)$. The first 15 components of $\boldsymbol{\beta}$ are 1.5, the remaining ones are zero.

6. The same as Example 5, except that $p = 400$.

7. $p = 200$ and $\sigma = 1.5$. The pairwise correlation between the k^{th} and the j^{th} components of $(x_{i,1}, \ldots, x_{i,200})$ is $r^{|k-j|}$ with $r = 0.5, k, j = 1, \ldots, 300$. Components 1–5 of $\boldsymbol{\beta}$ are 2.5, components 11–15 are 1.5, components 21–25 are 0.5, and the rest are zero.

8. The same as example 7, except that $r = 0.95$.

Partial orthogonal condition is satisfied in Examples 1–6. Especially, Examples 1 and 3 represent cases with moderately correlated covariates; Examples 2 and 4 have strongly correlated covariates; while Examples 5 and 6 have the grouping structure (Zou and Hastie, 2005) with three equally important groups, where covariates within the same group are highly correlated. Examples 7 and 8 represent the cases where the partial orthogonality assumption is violated. Covariates with nonzero coefficients are correlated with the rest.

In each example, the simulated data consist of a training set and a testing set, each of size 100. For both the Lasso and Adaptive Lasso, tuning parameters are selected based on V-fold cross validation with the training set only. We set $V = 5$. After tuning parameter selection, the Lasso and adaptive Lasso estimates are computed using the training set. We then compute the prediction MSE for the testing set, based on the training set estimate. Specifically, in each data set of the 500 replications, let $\hat{y}_i$ be the fitted value based on the training data, and let y_i be the response value in the testing data whose corresponding covariate value is the same as that associated with $\hat{y}_i$. Then the prediction MSE for this data set is $n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ where $n = 100$. The PMSE included in Table 1 is the median of the prediction MSE's from 500 replications.

Summary statistics of variable selection and PMSE results are shown in Table 1. It can be seen that for Examples 1-6, the adaptive Lasso yields smaller models with better prediction performance. However, due to the very large number of covariates, the number of covariates identified by the adaptive Lasso is still larger than the true value (15). When the partial orthogonality condition is not satisfied (Examples 7 and 8), the adaptive Lasso still yields smaller models with satisfactory prediction performance (comparable to the Lasso). Extensive simulation studies with other value of p and different marginal and joint distributions of x_{ij} yield similar, satisfactory results. We show in Figures 1 and 2 the frequencies of individual covariate effects being properly classified: zero versus nonzero. For a better view, we only show the first 100 coefficients which include all the nonzero coefficients. The patterns of the results from the remaining coefficients are similar.

4.2 Data example

We use the data set reported in Scheetz et al. (2006) to illustrate the application of the adaptive Lasso in high-dimensional settings. In this data set, F1 animals were intercrossed and 120 twelve-week-old male offspring were selected for tissue harvesting from the eyes and microarray analysis. The microarrays used to analyze the RNA from the eyes of these F2 animals contain over 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the RMA (robust multi-chip averaging, Bolstad 2003, Irizzary 2003) method to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale. For the 31,042 probe sets on the array, we first excluded probes that were not expressed in the eye or that lacked sufficient variation. The definition of expressed was based on the empirical distribution of RMA normalized values. For a probe to be considered expressed, the maximum expression value observed for that probe among the 120 F2 rats was required to be greater than the 25th percentile of the entire set of RMA expression values. For a probe to be considered "sufficiently variable", it had to exhibit at least 2-fold variation in expression level among the 120 F2 animals. A total of 18,976 probes met these two criteria.

We are interested in finding the genes whose expression are correlated with that of 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. The probe from TRIM32 is 1389163_at, which is one of the $18,976$ probes that are sufficiently expressed and variable. One approach to finding the probes among the remaining $18,975$ probes that are most related to TRIM32 is to use regression analysis. Here the sample size $n = 120$ (i.e., there are 120 arrays from 120 rats), and the number of probes is $18,975$. Also, it is expected that only a few genes are related to TRIM32. Thus this is a sparse, high-dimensional regression problem. We use the proposed approach in the analysis. We first standardize the probes so that they have mean zero and standard deviation 1. We then do the following steps:

1. Select 3000 probes with the largest variances;

2. Compute the marginal correlation coefficients of the 3000 probes with the probe corresponding to TRIM32;

3. Select the top 200 covariates with the largest correlation coefficients. This is equivalent to selecting the covariates based on marginal regression, since covariates are standardized.

4. The estimation and prediction results from ada-lasso and lasso are provided below.

Table 2 lists the probes selected by the adaptive Lasso. For comparison, we also used the Lasso. The Lasso selected 5 more probes than the adaptive Lasso. To evaluate the performance of the adaptive Lasso relative to the Lasso, we use cross validation and compare the predictive mean square errors (MSEs). Table 3 gives the results when the number of covariates $p = 100, 200, 300, 400$ and 500. We randomly partition the data into a training set and a test set, the training set consists of 2/3 observations and the test set consists of the remaining 1/3 observations. We then fit the model with the training set, then calculate the prediction MSE for the testing set. We repeat this process 300 times, each time a new random partition is made. The values in Table 3 are the medians of the results from 300 random partitions. In the table, # cov is the number of covariates being considered; Nonzero is the number of covariates in the final model; Corr is the correlation coefficient between the predicted value based on the model and the observed value; Coef is the slope of the regression of the fitted values of Y against the observed values of Y, which shows the shrinkage effects of the two methods are similar. Overall, we see that the performance of the adaptive Lasso and Lasso are similar. However, there are some improvement of the adaptive Lasso over the Lasso in terms of prediction MSEs. Notably, the number of covariates selected by the adaptive Lasso is fewer than that selected by the Lasso, yet the prediction MSE of the adaptive Lasso is smaller.

5 Concluding remarks

The adaptive Lasso is a two-step approach. In the first step, an initial estimator is obtained. Then a penalized optimization problem with a weighted L_1 penalty must be solved. The initial estimator does not need to be consistent, but it must put more weight on the zero coefficients and less on nonzero ones in an average sense to improve upon the Lasso. Under the partial orthogonality condition, a simple initial estimator can be obtained from marginal regression. Comparing to the Lasso, the theoretical advantage of the adaptive Lasso is that it has the oracle property. Comparing to the SCAD and bridge methods which also have the oracle property, the advantage of the adaptive Lasso is its computational efficiency. Given the initial estimator, the computation of adaptive Lasso estimate is a convex optimization problem and its computational cost is the same as the Lasso. Indeed, the entire regularization path of the adaptive Lasso can be computed with the same computational complexity as the least squares solution using the LARS algorithm (Efron et al. 2004). Therefore, the adaptive Lasso is a useful method for analyzing high-dimensional data.

We have focused on the adaptive Lasso in the context of linear regression models. This method can be applied in a similar way to other models such as the generalized linear and Cox models. It would be interesting to generalized the results of this paper to these more complicated models.

Acknowledgements. The authors wish to thank two anonymous referees and an Associate Editor for their helpful comments.

REFERENCES

1. Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101**, 119-137.

2. Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185-193

3. Candes, E. and Tao, T. (2005) The Dantzig selector: statistical estimation when p is much larger than n. *Preprint*, Department of Computational and Applied Mathematics, Caltech. Accepted for publication by the *Ann. Statist.*

4. 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* (USA),**103**, 6287-6292.

5. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407499.

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

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

8. Fan, J. and Lv, J. (2006). Sure independence screening for ultra-high dimensional feature space. Preprint, Department of Operational Research & Financial Engineering, Princeton University.

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

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

11. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

12. Huang, J., Horowitz, J. L., and Ma, S. G. (2006). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Technical report No. 360, Department of Statistics and Actuarial Science, University of Iowa. Accepted for publication by the *Ann. Statist.*

13. 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.

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

15. Leng, C., Lin, Y., and Wahba, G. (2004). A Note on the Lasso and Related Procedures in Model Selection. *Statistica Sinica* **16**, 1273-1284.

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

17. 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.

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

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

20. Zhang, C.-H. and Huang, J. (2006) 2006-003: Model-selection consistency of the LASSO in high-dimensional linear regression. Technical report 2006-003. Department of Statistics, Rutgers University.

21. Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *J. Machine Learning Res.* **7**, 2541-2567.

22. Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

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

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

Division of Biostatistics
Department of Epidemiology and Public Health
Yale University
New Haven, Connecticut 06520-8034
E-mail: shuangge.ma@yale.edu

Department of Statistics
504 Hill Center, Busch Campus
Rutgers University
Piscataway NJ 08854-8019
E-mail: cunhui@stat.rutgers.edu

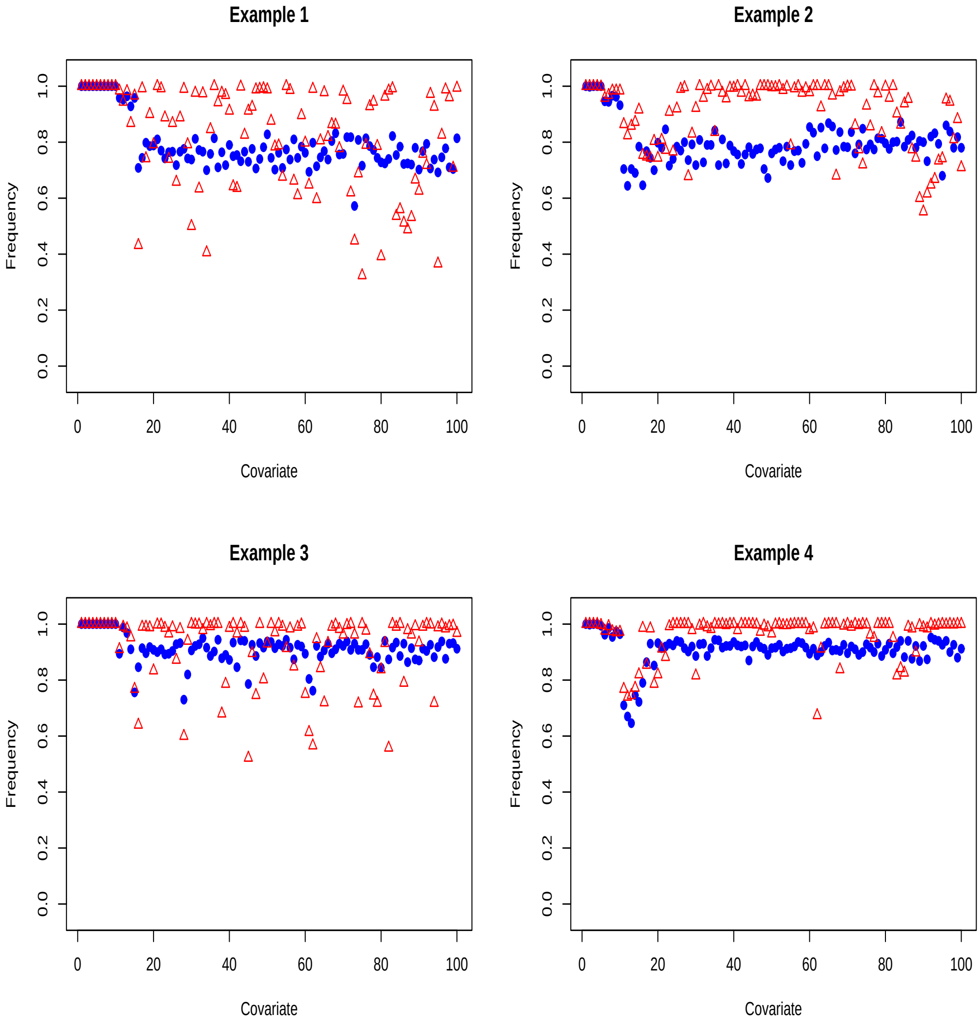

Figure 1: Simulation study (Examples 1–4): frequency of individual covariate effect being correctly identified. Circle: Lasso; Triangle: adaptive Lasso. Only the results of the first 100 coefficients are shown in the plots.

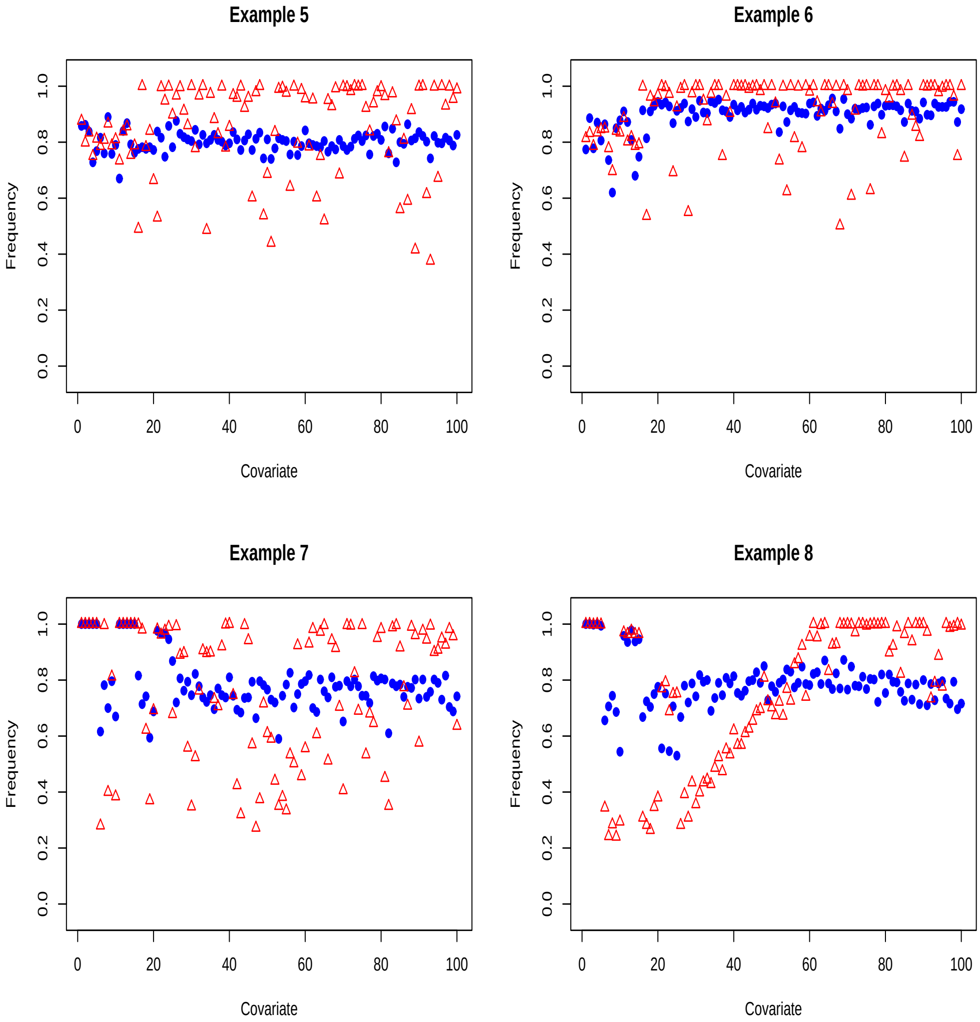

Figure 2: Simulation study (examples 5–8): frequency of individual covariate effect being correctly identified. Circle: Lasso; Triangle: adaptive Lasso. Only the results of the first 100 coefficients are are shown in the plots.

Table 1. Simulation study, comparison of adaptive Lasso with Lasso. PMSE: median of PMSE, inside "()" are the corresponding standard deviations. Covariate: median of number of covariates with nonzero coefficients.

	Lasso		Adaptive-Lasso	
Example	PMSE	Covariate	PMSE	Covariate
1	3.829 (0.769)	58	3.625 (0.695)	50
2	3.548 (0.636)	54	2.955 (0.551)	33
3	3.604 (0.681)	50	3.369 (0.631)	43
4	3.304 (0.572)	50	2.887 (0.499)	33
5	3.148 (0.557)	48	2.982 (0.540)	40
6	3.098 (0.551)	42	2.898 (0.502)	36
7	3.740 (0.753)	59	3.746 (0.723)	53
8	3.558 (0.647)	55	3.218 (0.578)	44

Table 2. The probe sets identified by Lasso and adaptive Lasso that correlated with TRIM32.

Probe ID	Lasso	Adaptive-Lasso
1369353_at	-0.021	-0.028
1370429_at	-0.012	
1371242_at	-0.025	-0.015
1374106_at	0.027	0.026
1374131_at	0.018	0.011
1389584_at	0.056	0.054
1393979_at	-0.004	-0.007
1398255_at	-0.022	-0.009
1378935_at	-0.009	
1379920_at	0.002	
1379971_at	0.038	0.041
1380033_at	0.030	0.023
1381787_at	-0.007	-0.007
1382835_at	0.045	0.038
1383110_at	0.023	0.034
1383522_at	0.016	0.01
1383673_at	0.010	0.02
1383749_at	-0.041	-0.045
1383996_at	0.082	0.081
1390788_a_at	0.013	0.001
1393382_at	0.006	0.004
1393684_at	0.008	0.003
1394107_at	-0.004	
1395415_at	0.004	

Table 3. Prediction results using cross validation. 300 random partitions of the data set are made, in each partition, the training set consists of 2/3 observations and the test set consists of the remaing 1/3 observations. The values in the table are medians of the results from 300 random partitions. In the table, # cov is the number of covariates being considered; nonzero is the number of covariates in the final model; corr is correlation coefficient between the fitted and observed values of Y; coef is the slope of the regression of the fitted values of Y against the observed values of Y, which shows the shrinkage effect of the methods.

	Lasso				Adaptive-Lasso			
# cov	nonzero	mse	corr	coef	nonzero	mse	corr	coef
100	20	0.005	0.654	0.486	18	0.006	0.659	0.469
200	19	0.005	0.676	0.468	17	0.005	0.678	0.476
300	18	0.005	0.669	0.443	17	0.005	0.671	0.462
400	22	0.005	0.676	0.442	19	0.005	0.686	0.476
500	25	0.005	0.665	0.449	22	0.005	0.670	0.463

ADAPTIVE LASSO FOR SPARSE HIGH-DIMENSIONAL REGRESSION MODELS: ON-LINE SUPPLEMENT

Jian Huang[1], Shuangge Ma[2], and Cun-Hui Zhang[3]

[1]University of Iowa, [2]Yale University, [3]Rutgers University

In this supplement, we prove Theorems 1 and 3.

Let $\psi_d(x) = \exp(x^d) - 1$ for $d \geq 1$. For any random variable X its ψ_d-Orlicz norm $\|X\|_{\psi_d}$ is defined as $\|X\|_{\psi_d} = \inf\{C > 0 : E\psi_d(|X|/C) \leq 1\}$. Orlicz norm is useful for obtaining maximal inequalities, see Van der Vaart and Wellner (1996) (hereafter referred to as VW (1996)).

Lemma 1 *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are iid random variables with $\mathrm{E}\varepsilon_i = 0$ and $Var(\varepsilon_i) = \sigma^2$. Furthermore, suppose that their tail probabilities satisfy $P(|\varepsilon_i| > x) \leq K\exp(-Cx^d), i = 1, 2, \ldots$ for constants C and K, and for $1 \leq d \leq 2$. Then, for all constants a_i satisfying $\sum_{i=1}^n a_i^2 = 1$,*

$$\Big\|\sum_{i=1}^n a_i\varepsilon_i\Big\|_{\psi_d} \leq \begin{cases} K_d\left\{\sigma + (1+K)^{1/d}C^{-1/d}\right\}, & 1 < d \leq 2 \\ K_1\left\{\sigma + (1+K)C\log n\right\}, & d = 1. \end{cases}$$

where K_d is a constant depending on d only. Consequently

$$q_n^*(t) = \sup_{a_1^2+\cdots+a_n^2=1} P\left\{\sum_{i=1}^n a_i\varepsilon_i > t\right\} \leq \begin{cases} \exp(-t^d/M), & 1 < d \leq 2 \\ \exp(-t^d/\{M(1+\log n)\}), & d = 1, \end{cases}$$

for certain constant M depending on $\{d, K, C\}$ only.

Proof. Because ε_i satisfies $P(|\varepsilon_i| > x) \leq K\exp(-Cx^d)$, its Orlicz norm $\|\varepsilon_i\|_{\psi_2} \leq [(1+K)/C]^{1/d}$ (Lemma 2.2.1, VW 1996). Let d' be given by $1/d + 1/d' = 1$. By Proposition A.1.6 of VW (1996), there exists a constant K_d such that

$$\begin{aligned} \left\|\sum_{i=1}^n a_i\varepsilon_i\right\|_{\psi_d} &\leq K_d\left\{\mathrm{E}\left|\sum_{i=1}^n a_i\varepsilon_i\right| + \left[\sum_{i=1}^n \|a_i\varepsilon_i\|_{\psi_d}^{d'}\right]^{1/d'}\right\} \\ &\leq K_d\left\{\left[\mathrm{E}\left(\sum_{i=1}^n a_i\varepsilon_i\right)^2\right]^{1/2} + (1+K)^{1/d}C^{-1/d}\left[\sum_{i=1}^n |a_i|^{d'}\right]^{1/d'}\right\} \\ &\leq K_d\left\{\sigma + (1+K)^{1/d}C^{-1/d}\left[\sum_{i=1}^n |a_i|^{d'}\right]^{1/d'}\right\}. \end{aligned}$$

For $1 < d \leq 2$, $d' = d/(d-1) \geq 2$. Thus $\sum_{i=1}^n |a_i|^{d'} \leq (\sum_{i=1}^n |a_i|^2)^{d'/2} = 1$. It follows that

$$\left\| \sum_{i=1}^n a_i \varepsilon_i \right\|_{\psi_d} \leq K_d \left\{ \sigma + (1+K)^{1/d} C^{-1/d} \right\}.$$

For $d = 1$, by Proposition A.1.6 of VW (1996), there exists a constant K_1 such that

$$\begin{aligned} \left\| \sum_{i=1}^n a_i \varepsilon_i \right\|_{\psi_1} &\leq K_1 \left\{ \mathrm{E} \Big| \sum_{i=1}^n a_i \varepsilon_i \Big| + \| \max_{1 \leq i \leq n} |a_i \varepsilon_i| \|_{\psi_1} \right\} \\ &\leq K_1 \left\{ \sigma + K' \log(n) \max_{1 \leq i \leq n} \|a_i \varepsilon_i\|_{\psi_1} \right\} \\ &\leq K_1 \left\{ \sigma + K'(1+K) C^{-1} \log(n) \max_{1 \leq i \leq n} |a_i| \right\} \\ &\leq K_1 \left\{ \sigma + K'(1+K) C^{-1} \log(n) \right\}. \end{aligned}$$

The last inequality follows from

$$P(X > t \|X\|_{\psi_d}) \leq \{\psi_d(t) + 1\}^{-1} \Big(1 + E\psi_d(|X|/\|X\|_{\psi_d}) \Big) \leq 2e^{-t^d}, \ \forall t > 0$$

in view of the definition of $\|X\|_{\psi_d}$. □

Lemma 2 *Let* $\widetilde{\mathbf{s}}_{n1} = (|\widetilde{\beta}_{nj}|^{-1} sgn(\beta_{0j}), j \in J_{n1})'$ *and* $\mathbf{s}_{n1} = (|\eta_{nj}|^{-1} sgn(\beta_{0j}), j \in J_{n1})'$. *Suppose (A2) holds. Then,*

$$\Big\| \widetilde{\mathbf{s}}_{n1} \Big\| = (1 + o_P(1)) M_{n1}, \quad \max_{j \notin J_{n1}} \Big\| |\widetilde{\beta}_{nj}| \widetilde{\mathbf{s}}_{n1} - |\eta_{nj}| \mathbf{s}_{n1} \Big\| = o_P(1). \tag{12}$$

Proof. Since $M_{n1} = o(r_n)$, $\max_{j \in J_{n1}} \big| |\widetilde{\beta}_{nj}|/|\eta_{nj}| - 1 \big| \leq M_{1n} O_P(1/r_n) = o_P(1)$ by the r_n-consistency of $\widetilde{\beta}_{nj}$. Thus, $\|\widetilde{\mathbf{s}}_{n1}\| = (1 + o_P(1)) M_{n1}$. For the second part of (12), we have

$$\max_{j \notin J_{n1}} \|(|\eta_{nj}| \widetilde{\mathbf{s}}_{n1} - |\eta_{nj}| \mathbf{s}_{n1})\|^2 \leq M_{n2}^2 \sum_{j \in J_{n1}} \left| \frac{|\widetilde{\beta}_{nj}| - |\eta_{nj}|}{|\widetilde{\beta}_{nj}| \cdot |\eta_{nj}|} \right|^2 = O_P(M_{n1}^2 / r_n^2) = o_P(1) \tag{13}$$

and $\max_{j \notin J_{n1}} \|(|\widetilde{\beta}_{nj}| - |\eta_{nj}|) \widetilde{\mathbf{s}}_{n1}\| = O_P(M_{n1}/r_n) = o_P(1)$. □

Proof of Theorem 1. Let $J_{n1} = \{j : \beta_{0j} \neq 0\}$. It follows from the Karush-Kunh-Tucker conditions that $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1}, \ldots, \widehat{\beta}_{np})'$ is the unique solution of the adaptive Lasso if

$$\begin{cases} \mathbf{x}_j'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n) = \lambda_n w_{nj} \mathrm{sgn}(\widehat{\beta}_{nj}), & \widehat{\beta}_{nj} \neq 0 \\ |\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n)| < \lambda_n w_{nj}, & \widehat{\beta}_{nj} = 0 \end{cases} \tag{14}$$

and the vectors $\{\mathbf{x}_j, \widehat{\beta}_{nj} \neq 0\}$ are linearly independent. Let $\widetilde{\mathbf{s}}_{n1} = \big(w_{nj}\mathrm{sgn}(\beta_{0j}), j \in J_{n1}\big)'$ and

$$\widehat{\boldsymbol{\beta}}_{n1} = \Big(\mathbf{X}_1'\mathbf{X}_1\Big)^{-1}(\mathbf{X}_1'\mathbf{y} - \lambda_n\widetilde{\mathbf{s}}_{n1}) = \boldsymbol{\beta}_{01} + \Sigma_{n11}^{-1}(\mathbf{X}_1'\boldsymbol{\varepsilon} - \lambda_n\widetilde{\mathbf{s}}_{n1})/n, \tag{15}$$

where $\Sigma_{n11} = \mathbf{X}_1'\mathbf{X}_1/n$. If $\widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}$, then the equation in (14) holds for $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}', \mathbf{0}')'$. Thus, since $\mathbf{X}\widehat{\boldsymbol{\beta}}_n = \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1}$ for this $\widehat{\boldsymbol{\beta}}_n$ and $\{\mathbf{x}_j, j \in J_{n1}\}$ are linearly independent,

$$\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} \widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01} \\ \big|\mathbf{x}_j'\big(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1}\big)\big| < \lambda_n w_{nj}, \ \forall j \notin J_{n1}. \end{cases} \tag{16}$$

This is a variation of Proposition 1 of Zhao and Yu (2007). Let $\mathbf{H}_n = \mathbf{I}_n - \mathbf{X}_1\Sigma_{n11}^{-1}\mathbf{X}_{n1}'/n$ be the projection to the null of $\mathbf{X}_1'$. It follows from (15) that $\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1} = \boldsymbol{\varepsilon} - \mathbf{X}_1(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) = \mathbf{H}_n\boldsymbol{\varepsilon} + \mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}\lambda_n/n$, so that by (16)

$$\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} \mathrm{sgn}(\beta_{0j})(\beta_{0j} - \widehat{\beta}_{nj}) < |\beta_{0j}|, & \forall j \in J_{n1} \\ \Big|\mathbf{x}_j'\Big(\mathbf{H}_n\boldsymbol{\varepsilon} + \mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}\lambda_n/n\Big)\Big| < \lambda_n w_{nj}, & \forall j \notin J_{n1}. \end{cases} \tag{17}$$

Thus, by (17) and (15), for any $0 < \kappa < \kappa + \epsilon < 1$

$$\begin{aligned} P\Big\{\widehat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}_0\Big\} \ &\leq \ P\Big\{|\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}|/n \geq |\beta_{0j}|/2 \text{ for some } j \in J_{n1}\Big\} \\ &\quad +P\Big\{|\mathbf{e}_j\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}|\lambda_n/n \geq |\beta_{0j}|/2 \text{ for some } j \in J_{n1}\Big\} \\ &\quad +P\Big\{|\mathbf{x}_j'\mathbf{H}_n\boldsymbol{\varepsilon}| \geq (1-\kappa-\epsilon)\lambda_n w_{nj} \text{ for some } j \notin J_{n1}\Big\} \\ &\quad +P\Big\{|\mathbf{x}_j'\mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}|/n \geq (\kappa+\epsilon)w_{nj} \text{ for some } j \notin J_{n1}\Big\} \\ &= \ P\{B_{n1}\} + P\{B_{n2}\} + P\{B_{n3}\} + P\{B_{n4}\}, \quad \text{say,} \end{aligned} \tag{18}$$

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

Since $\|(\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1')'\|/n \leq n^{-1/2}\|\Sigma^{-1/2}\| \leq (n\tau_{n1})^{-1/2}$ and $|\beta_{0j}| \geq b_{n1}$ for $j \in J_{n1}$,

$$P\{B_{n1}\} = P\Big\{|\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}|/n \geq |\beta_{0j}|/2, \exists j \in J_{n1}\Big\} \leq k_n q_n^*\big(\sqrt{\tau_{n1}n}b_{n1}/2\big)$$

with the tail probability $q_n^*(t)$ in Lemma 1. Thus, $P\{B_{n1}\} \to 0$ by (A1), Lemma 1, (A4) and (A5).

Since $w_{nj} = 1/|\widetilde{\beta}_{nj}|$, by Lemma 2 and conditions (A4) and (A5)

$$|\mathbf{e}_j \Sigma_{n11}^{-1} \widetilde{\mathbf{s}}_{n1}| \lambda_n / n \leq \frac{\|\widetilde{\mathbf{s}}_{n1}\| \lambda_n}{\tau_{n1} n} = O_P\Big(\frac{M_{n1}\lambda_n}{\tau_{n1} n}\Big) = o_P(b_{n1}),$$

where $b_{n1} = \min\{|\beta_{0j}|, j \in J_{n1}\}$. This gives $P\{B_{n2}\} = o(1)$.

For B_{n3}, we have $w_{nj}^{-1} = |\widetilde{\beta}_{nj}| \leq M_{n2} + O_P(1/r_n)$. Since $\|(\mathbf{x}_j \mathbf{H}_n)'\| \leq \sqrt{n}$, for large C

$$\begin{aligned} P\{B_{n3}\} &\leq P\Big\{|\mathbf{x}_j' \mathbf{H}_n \boldsymbol{\varepsilon}| \geq (1-\kappa-\epsilon)\lambda_n / \{C(M_{n2} + 1/r_n)\}, \exists j \notin J_{n1}\Big\} + o(1) \\ &\leq m_n q_n^*\big((1-\kappa-\epsilon)\lambda_n / \{C(M_{n2} + 1/r_n)\sqrt{n}\}\big). \end{aligned}$$

Thus, by Lemma 1 and (A4), $P\{B_{n3}\} \to 0$.

Finally for B_{n4}, Lemma 2 and condition (A5) imply

$$\begin{aligned} &\max_{j \notin J_{n1}} \Big(\frac{|\mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \widetilde{\mathbf{s}}_{n1}|}{n w_{nj}} - |\eta_{nj} \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1}|\Big) \\ \leq\ &\max_{j \notin J_{n1}} \Big(\|(\mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1})'\|/n\Big) \Big\| |\widetilde{\beta}_{nj}| \widetilde{\mathbf{s}}_{n1} - |\eta_{nj}| \mathbf{s}_{n1} \Big\| \leq \tau_{n1}^{-1/2} o_P(1) = o_P(1), \end{aligned}$$

due to $\|\mathbf{x}_j\|^2/n = 1$. Since $|\eta_{nj} \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1}| \leq \kappa$ by (A3), we have $P\{B_{n4}\} \to 0$. □

Proof of Theorem 3. Let $\boldsymbol{\mu}_0 = E\mathbf{y} = \sum_{j=1}^{p_n} \mathbf{x}_j \beta_{0j}$. Then,

$$\widetilde{\beta}_{nj} = \mathbf{x}_j' \mathbf{y}/n = \eta_{nj} + \mathbf{x}_j' \boldsymbol{\varepsilon}/n$$

with $\eta_{nj} = \mathbf{x}_j' \boldsymbol{\mu}_0 / n$. Since $\|\mathbf{x}_j\|^2/n = 1$, by Lemma 1, for all $\epsilon > 0$

$$P\Big\{r_n \max_{j \leq p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| > \epsilon\Big\} = P\Big\{r_n \max_{j \leq p_n} |\mathbf{x}_j' \boldsymbol{\varepsilon}|/n > \epsilon\Big\} \leq p_n q_n^*(\sqrt{n}\epsilon / r_n) = o(1)$$

due to $r_n (\log p)(\log n)^{I\{d=1\}}/\sqrt{n} = o(1)$. For the second part of (A2) with $M_{n2} = \max_{j \notin J_{n1}} |\eta_{nj}|$, we have by (B3)

$$\sum_{j \in J_{n1}} \Big(\frac{1}{\eta_{nj}^2} + \frac{M_{n2}^2}{\eta_{nj}^4}\Big) \leq \frac{k_n}{\widetilde{b}_{n1}^2}(1 + c_n^2) = o(r_n^2).$$

To verify (A3), we notice that

$$\|\mathbf{X}_1' \mathbf{x}_j\|^2 = \sum_{l \in J_{n1}} \Big(\mathbf{x}_l' \mathbf{x}_j\Big)^2 \leq k_n n^2 \rho_n^2$$

and $|\eta_{nj}| \times \|\mathbf{s}_{n1}\| \leq k_n^{1/2} c_n$ for all $j \notin J_{n1}$. Thus, for such j, (B2) implies

$$|\eta_{nj}| n^{-1} \left| \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1} \right| \leq c_n k_n \rho_n / \tau_{n1} \leq \kappa.$$

The proof is complete. □