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

Summary. We study the asymptotic properties of 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, adaptive LASSO correctly select 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, univariate 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 greater than the sample size.

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

Short title. Sparse high-dimensional regression

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

1 Introduction

Consider a linear regression model

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p_n \times 1$ vector, $\varepsilon_1, \ldots, \varepsilon_n$ are 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 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

$$\sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p_n} |\beta_j|, \tag{2}$$

where λ is the penalty parameter. An important feature of 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 LASSO is continuous and hence more stable than the subset selection, which is a discrete and non-continuous. 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 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 LASSO is in general not path consistent in the sense that (1) with probability greater than zero, the whole LASSO path may not contain the true parameter value; (2) 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 LASSO. He showed that the positive probability mass at 0 of the LASSO, when the true value of the parameter is 0, is in general less than 1, which implies that 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 also discovered by Meinshausen and Buhlmann (2006) and Zhao and Yu (2006). In particular, Zhao and Yu (2006) called this condition the irrepresentable condition on the design matrix. Meinshausen and Buhlmann (2006) and Zhao and Yu (2006) 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 $n^{1/2}$. Specifically, p_n is allowed to be as large as $\exp(n^a)$ for some $0 < a < 1$ when the errors have Gaussian tails. Thus their results are applicable to truly high-dimensional data. However, the value of λ required for variable selection consistency over shrinks the nonzero coefficients, which leads to asymptotically biased estimates of the nonzero coefficients. Therefore, LASSO is variable-selection consistent under certain conditions, but not in general. However, if LASSO is variable-selection consistent, then it is not consistent for estimating the nonzero parameters. Therefore, these studies confirm the suggestion that 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, LASSO has the persistence property in estimating $X\boldsymbol{\beta}$ when p is larger than n (Greenshtein and Ritov 2004).

In addition to 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) studied asymptotic properties of penalized likelihood methods when the number of parameters is finite. Fan and Peng (2004) considered the same problem when the number of parameters diverges. 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. However, there has been effort to devise efficient algorithms for non-convex penalized problems (Fan and Li 2001, Hunter and Li 2005). Another interesting estimator, the Dantzig selector, of $\boldsymbol{\beta}$ in high-dimensional settings was proposed and studied 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

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

Denote

$$L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} w_{nj}|\beta_j|. \tag{3}$$

The value $\widehat{\boldsymbol{\beta}}_n$ that minimizes L_n is called the adaptive LASSO estimator (Zou 2006). If the initial estimator $\widetilde{\boldsymbol{\beta}}_n$ is zero-consistent in the sense that estimators of zero coefficients converge to zero in probability and estimators of non-zero coefficients do not converge to zero, then the weights for the zero coefficients converge to infinity, while the weights for the nonzero coefficients are bounded. The precise definition of zero-consistency is given in the next section.

For fixed p_n, Zou (2005) 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, if an initial zero-consistent estimator is available and if $p_n = O(\exp(n^a))$ for some $0 < a < 1$, then the adaptive LASSO has the oracle property. Here a depends on the rate of the initial zero-consistent estimator and the tail behavior of the error term ε_i. Thus the number of covariates can be larger than the sample size if an initial zero-consistent estimator is available.

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 provide sufficient conditions for the marginal regression estimators to be zero-consistent. Thus under these conditions, marginal regression estimators can be used in the adaptive LASSO. In Section 4, we present results from simulation studies and a real data example. Some concluding remarks are given in Section 5.

2 Variable-selection consistency and asymptotic normality

Let the true parameter value be $\boldsymbol{\beta}_0^n$. For simplicity of notation, we will simply write $\boldsymbol{\beta}_0$. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'$, 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.

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})'$ be the $p_n \times 1$ vector of covariates of the ith observation, $i = 1, \ldots, n$. We assume that the Y_i's are centered and the covariates are standardized, i.e.,

$$\sum_{i=1}^{n} Y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, j = 1, \ldots, p_n. \tag{4}$$

We also write $\mathbf{x}_i = (\mathbf{x}_{i1}', \mathbf{x}_{i2}')'$ where $\mathbf{x}_{i1}$ consists of the first k_n covariates with nonzero coefficients, and $\mathbf{x}_{i2}$ consists of the remaining m_n covariates with zero coefficients. Let $\mathbf{X}_n$, $\mathbf{X}_{1n}$, and $\mathbf{X}_{2n}$ be the matrices whose transposes are $\mathbf{X}_n' = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{X}_{1n}' = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{n1})$, and $\mathbf{X}_{2n}' = (\mathbf{x}_{12}, \ldots, \mathbf{x}_{n2})$, respectively. Let

$$\Sigma_n = n^{-1}\mathbf{X}_n'\mathbf{X}_n, \ \Sigma_{n11} = n^{-1}\mathbf{X}_{1n}'\mathbf{X}_{1n}, \ \text{and} \ \ \Sigma_{n12} = \Sigma_{n21}' = n^{-1}\mathbf{X}_{1n}'\mathbf{X}_{2n}$$

Let $H_n = \mathbf{X}_{n1}(\mathbf{X}_{n1}'\mathbf{X}_{n1})^{-1}\mathbf{X}_{n1}'$. Let ρ_{1n} and ρ_{2n} be the smallest and largest eigenvalues of Σ_n, and let τ_{1n} and τ_{2n} be the smallest and largest eigenvalues of Σ_{1n}, respectively. For any vector $\mathbf{x} = (x_1, x_2, \ldots)'$, denote

$$|\mathbf{x}| = (|x_1|, |x_2|, \ldots)', \quad \text{and} \quad \text{sgn}(\mathbf{x}) = (\text{sgn}(x_1), \text{sgn}(x_2), \ldots)',$$

where $\text{sgn}(x_1) = 1$ if $x_1 > 0$; $= 0$ if $x_1 = 0$; and $= -1$ if $x_1 < 0$. Following Zhao and Yu (2005), we say that $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}$ if and only if $\text{sgn}(\widehat{\boldsymbol{\beta}}_n) = \text{sgn}(\boldsymbol{\beta})$.

Proposition 1 *Let* $W_{n1} = diag(w_{n1}, \ldots, w_{nk_n})$, $W_{n2} = diag(w_{n,k_n+1}, \ldots, w_{np_n})$, *and* $\mathbf{w}_{n2} = (w_{n,k_n+1}, \ldots, w_{np_n})'$. *Then*

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0) \geq \mathrm{P}(A_n \cap B_n), \tag{5}$$

where

$$A_n = \left\{ 2n^{-1/2} |\Sigma_{n11}^{-1} \mathbf{X}_1' \boldsymbol{\varepsilon}_n| < 2\sqrt{n}\ |\boldsymbol{\beta}_{n1}| - n^{-1/2} \lambda_n |\Sigma_{n11}^{-1} W_{n1} sgn(\boldsymbol{\beta}_{10})| \right\},$$

and

$$B_n = \left\{ 2n^{-1/2} |\mathbf{X}_{n2}'(I - H_n)\boldsymbol{\varepsilon}_n| \leq n^{-1/2}\lambda_n \mathbf{w}_{n2} - n^{-1/2}\lambda_n |\Sigma_{n21}\Sigma_{n11}^{-1} W_{n1} sgn(\boldsymbol{\beta}_{10})| \right\},$$

where the inequalities in A_n *and* B_n *are component-wise.*

Proposition 1 can be proved following the proof of Proposition 1 of Zhao and Yu (2005).

Let $J_{0n} = \{j : \beta_{0j} = 0\}$ and $J_{1n} = \{j : \beta_{0j} \neq 0\}$. Let

$$b_{n1} = \min\{|\beta_{0j}| : j \in J_{1n}\}, \quad \text{and} \quad b_{n2} = \max\{|\beta_{0j}| : j \in J_{1n}\}. \tag{6}$$

Definition 1 *We say that* $\widetilde{\boldsymbol{\beta}}_n$ *is zero-consistent if (a)* $\max_{j \in J_{0n}} |\widetilde{\beta}_{nj}| = o_p(1)$ *and, (b) There exists a constant* $\xi_b > 0$ *such that, for any* $\varepsilon > 0$,

$$\mathrm{P}\left(\min_{j \in J_{1n}} |\widetilde{\beta}_{nj}| \geq \xi_b\, b_{n1} \right) > 1 - \varepsilon$$

for all n *sufficiently large. In addition,* $\widetilde{\boldsymbol{\beta}}_n$ *is zero-consistent with rate* r_n *if (a) is strengthened to*

$$r_n \max_{j \in J_{0n}} |\widetilde{\beta}_{nj}| = O_p(1), \tag{7}$$

where $r_n \to \infty$.

We assume the following conditions.

(A1) (a) $\varepsilon_i, \varepsilon_2, \ldots$ are independent and identically distributed random variables with mean zero and variance σ^2, where $0 < \sigma^2 < \infty$; (b) For $1 \leq d \leq 2$, the tail probabilities of ε_i satisfy $P(|\varepsilon_i| > x) \leq K \exp(-Cx^d), i = 1, 2, \ldots$ for constants C and K.

(A2) The initial estimator $\widetilde{\boldsymbol{\beta}}_n$ is zero-consistent with rate $r_n \to \infty$.

(A3) $\lambda_n \to \infty$, $\lambda_n k_n / n^{1/2} \to 0$ and

(a) for $1 < d \leq 2$,

$$\frac{(\log k_n)^{1/d}}{\sqrt{n}\, b_{n1}} \to 0, \quad \text{and} \quad \frac{\lambda_n k_n}{n b_{n1}^2} \to 0, \tag{8}$$

$$\frac{\sqrt{n}(\log m_n)^{1/d}}{\lambda_n r_n} \to 0, \quad \text{and} \quad \frac{k_n^2}{r_n b_{n1}} \to 0. \tag{9}$$

(b) for $d = 1$,

$$\frac{(\log n)(\log k_n)}{\sqrt{n}\, b_{n1}} \to 0, \quad \text{and} \quad \frac{\lambda_n k_n}{n b_{n1}^2} \to 0, \tag{10}$$

$$\frac{\sqrt{n}(\log n)(\log m_n)}{\lambda_n r_n} \to 0, \quad \text{and} \quad \frac{k_n^2}{r_n b_{n1}} \to 0. \tag{11}$$

(A4) There exist constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 \leq \tau_{1n} \leq \tau_{2n} \leq \tau_2$ for all n;

Condition (A1a) is standard in linear regression. Condition (A1b) allows a range of tail behaviors of the error term ε_i, from sub-Gaussian to exponential. Condition (A2) assumes that an initial zero-consistent estimator exists. Condition (A3) puts restrictions on 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 term. For sub-Gaussian tail, the model can include more covariates, while for exponential tail, the number of covariates allowed is fewer. We often have $r_n = n^{1/2-\delta}$ and $\lambda_n = n^a$ for some $0 < a < 1/2$ and small $\delta > 0$. In this case, the number of zero coefficients can be as large as $\exp(n^{a(2-\delta)})$. But the number of nonzero coefficients allowed is on the order of $n^{a/2}$, assuming $b_{n1} > b_0 > 0$ for all n. (A4) assumes that the eigenvalues of Σ_{n11} are bounded away from zero and infinity. This is reasonable since the number of nonzero covariates is small in a sparse model.

Condition (A2) is the most critical one and it is in general difficult to establish. It assumes that we can consistently differentiate between zero and nonzero coefficients. On the other hand, this condition essentially reduces the task of establishing oracle property to the simpler property of zero-consistency.

Theorem 1 *Suppose that conditions (A1)-(A4) hold. Then*

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

Theorem 2 *Suppose that conditions (A1) to (A4) 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* $\|\boldsymbol{\alpha}_n\|_2 \leq 1$. *Then*

$$n^{1/2} s_n^{-1} \boldsymbol{\alpha}_n' (\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_0) = n^{-1/2} s_n^{-1} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{\alpha}_n' \Sigma_{n11}^{-1} \mathbf{x}_{1i} + o_p(1) \to_D N(0,1), \tag{12}$$

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 Initial zero-consistent estimator

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. When $p_n = o(n^{1/2})$, the least squares estimator is consistent and therefore zero-consistent. In this case, we can use the least squares estimator as the initial estimators for the weights. However, when $p_n > O(n^{1/2})$ or $p_n > n$, which is the case in many microarray gene expression studies, the least squares estimator is no longer feasible. When $p_n > n$, the regression parameter is in general not identified 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. For example, when the columns of the covariate matrix $\mathbf{X}$ are mutually orthogonal, each regression coefficient can be estimated by univariate regression. But in practice, mutual orthogonality is often too strong an assumption. Furthermore, when $p_n > n$, mutual orthogonality of all covariates is not possible, since the rank of $\mathbf{X}$ is at most $n - 1$. 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 under the partial orthogonality condition and certain other conditions, univariate regression estimator is zero-consistent even when the number of covariates is greater than n, although it does 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. 2004).

With the centering and scaling given in (4), the estimated univariate regression coefficient

$$\widetilde{\beta}_j = \frac{\sum_{i=1}^n x_{ij} Y_i}{\sum_{i=1}^n x_{ij}^2} = n^{-1} \sum_{i=1}^n x_{ij} Y_i.$$

Let $\xi_{nj} = \mathrm{E}\widetilde{\beta}_j$. Since $\mathrm{E}y_i = \mathbf{x}'_{(1)i}\boldsymbol{\beta}_{10}$, we have

$$\xi_{nj} = n^{-1} \sum_{i=1}^n x_{ij} \mathbf{x}'_{(1)i} \boldsymbol{\beta}_{10}. \tag{13}$$

We make the following assumptions:

(B1) (a) $\varepsilon_i, \varepsilon_2, \ldots$ are independent and identically distributed random variables with mean zero and variance σ^2, where $0 < \sigma^2 < \infty$; (b) For $1 \le d \le 2$, the tail probabilities of ε_i satisfy $P(|\varepsilon_i| > x) \le K \exp(-Cx^d), i = 1, 2, \ldots$ for constants C and K.

(B2) The covariates of the nonzero coefficients and those of the zero coefficients are only weakly correlated

$$\frac{1}{n} \sum_{i=1}^n x_{ij} x_{ik} = O(n^{-1/2}), \quad j = J_{n0},\ k \in J_{n1}.$$

(B3) (a) There exists a constant $\xi_r > 0$ such that $\min\{|\xi_{nj}|, j \in J_{n1}\} > 2\xi_r b_{n1}$; (b) There exists a constant $0 < b_2 < \infty$ such that $b_{n2} \le b_2$. Here b_{n1} and b_{n2} are defined in (6).

(B4). (a) $r_n \to \infty$ and $r_n k_n n^{-1/2} \to 0$; (b) For $1 < d \le 2$, $r_n n^{-1/2} (\log m_n)^{1/d} \to 0$, and for $d = 1$, $r_n n^{-1/2} (\log n)(\log m_n) \to 0$; (c) $k_n \exp(-C\xi_r b_{n1}^d n^{d/2}) \to 0$.

We note that (B1) is the same as (A1). (B2) is the weak partial orthogonality assumption. (B3a) requires that the "correlation" between Y and x_j converges to zero no slower than the smallest non-zero coefficient. (B3b) requires that the non-zero coefficients are bounded above. (B4) puts restrictions on the rate of growth of k_n and m_n.

Theorem 3 *Let c_0 be a positive constant. Suppose that conditions (B1) to (B4) hold. Then*

$$\mathrm{P}\left(r_n \max_{j \in J_{n0}} |\widetilde{\beta}_{nj}| > c_0 \right) \to 0, \quad \textit{and} \quad \mathrm{P}\left(\min_{j \in J_{n1}} |\widetilde{\beta}_{nj}| > \xi_r b_{n1} \right) \to 1.$$

That is, $\widetilde{\boldsymbol{\beta}}_n$ is zero-consistent with rate r_n.

This theorem shows that condition (A2) is satisfied under (B1) to (B4), which provides justification for using univariate regression estimator for adaptive LASSO as the initial estimator under the partial orthogonality condition. Therefore, under (A1), (A3), (A4) and (B2)-(B4), we can first use the univariate regression to obtain the initial zero-consistent estimators, and use them as weights in the adaptive LASSO to achieve variable-selection consistency and oracle efficiency. In the special case when the number of nonzero coefficients k_n is finite and smallest absolute nonzero coefficient $b_{n1} > b_1$ for some $b_1 > 0$, the conditions can be much simplified. Specifically, (A3) simplifies to
(A3*) $\lambda_n \to \infty, \lambda_n/n^{1/2} \to 0$ and (a) for $1 < d \leq 2$, $\sqrt{n}(\log m_n)^{1/d}/(\lambda_n r_n) \to 0$. (b) for $d = 1$, $\sqrt{n}(\log n)(\log m_n)/(\lambda_n r_n) \to 0$.
(B4) simplifies to
(B4*) (a) $r_n \to \infty, r_n n^{-1/2} \to 0$; (b) for $1 < d \leq 2$, $r_n n^{-1/2}(\log m_n)^{1/d} \to 0$, and for $d = 1$, $r_n n^{-1/2}(\log n)(\log m_n) \to 0$.

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 this method. Because our main interest is in when p_n is large and Zhou (2005) 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

$$y = x'\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

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 x is generated once and then kept fixed. Replications are obtained by simulating the values of ϵ from $N(0, \sigma^2)$ and then setting $y = x'\beta + \epsilon$ for the fixed covariate value x. Summary statistics are computed based on 500 replications.

The eight models we consider are

1. $p = 200$ and $\sigma = 1.5$. The first 15 covariates $(x_1, \ldots, x_{15})$ and the remaining 185 covarites $(x_{16}, \ldots, x_{200})$ are independent. The pairwise correlation between the i^{th} and the j^{th} components of $(x_1, \ldots, x_{15})$ is $r^{|i-j|}$ with $r = 0.5, i, j = 1, \ldots, 15$. The pairwise correlation between the i^{th} and the j^{th} components of $(x_{16}, \ldots, x_{200})$ is $r^{|i-j|}$ with $r = 0.5, i, j = 16, \ldots, 200$. Components 1–5 of β 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. $p = 200$ and $\sigma = 1.5$. The predictors in Example 3 are generated as follows:

$$\begin{aligned}
x_i &= Z_1 + e_i, \quad Z_1 \sim N(0,1), \ i = 1, \ldots, 5; \\
x_i &= Z_2 + e_i, \quad Z_2 \sim N(0,1), \ i = 6, \ldots, 10; \\
x_i &= Z_3 + e_i, \quad Z_3 \sim N(0,1), \ i = 11, \ldots, 15; \\
X_i &\sim N(0,1), \quad X_i \ \ i.i.d. \ \ i = 16, \ldots, 200,
\end{aligned}$$

   where e_i are i.i.d $N(0, 0.01), i = 1, \ldots, 15$. The first 15 components of β are 1.5, the remaining ones are zero.

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

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

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

7. $p = 200$ and $\sigma = 1.5$. The pairwise correlation between the i^{th} and the j^{th} components of $(x_1, \ldots, x_{200})$ is $r^{|i-j|}$ with $r = 0.5, i, j = 1, \ldots, 300$. Components 1–5 of β 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 4 represent cases with moderately correlated covariates; Examples 2 and 5 have strongly correlated covariates; while Examples 3 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 (1-5, 11-15, 21-25) 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 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, 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.

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 Lasso). Extensive simulation studies with other value of p and different marginal and joint distributions of x 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 covariates.

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 univariate 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 adaptive LASSO relative to 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 follow steps 3 and 4 above to 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 adaptive LASSO over LASSO in terms of prediction MSEs. Notably, the number of covariates selected by the adaptive LASSO is fewer than that selected by 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 should be zero-consistent. Under the partial orthogonality condition, a simple zero-consistent initial estimator can be obtained from univariate 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 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 LASSO. Indeed, the entire regularization path of 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.

6 Appendix

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), Section 2.2.

Lemma 1 *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are mutually uncorrelated 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 $1 \leq d \leq 2$. Let $c_1, \ldots, c_n$ be constants satisfying $\sum_{i=1}^n c_i^2 \leq \xi_2^2$, where $0 < \xi_2 < \infty$. Let $W = \sum_{i=1}^n a_i \varepsilon_i$.*

(i) (a) For $1 < d \leq 2$, $\|W\|_{\psi_d} \leq K_d \left\{\xi_2 \sigma + ((1+K)^{1/d} C^{-1/d} \xi_2\right\}$, where K_d is a constant.

(i)(b) For $d = 1$, $\|W\|_{\psi_1} \leq K_1 \left\{\xi_2\, \sigma + K'(1+K) C^{-1} \xi_2 \log(n)\right\}$.

(ii)(a) For $1 < d \leq 2$, let $W_1, \ldots, W_m$ be any random variables satisfying $\|W_j\|_{\psi_d} \leq c, 1 \leq j \leq m$. For any $w_n > 0$,

$$\mathrm{P}\left(\max_{1 \leq j \leq m} |W_j| \geq w_n\right) \leq \frac{c(\log m)^{1/d}}{w_n}$$

for a constant c not depending on n.

(ii)(b) For $d = 1$, let $W_1, \ldots, W_m$ be any random variables satisfying $\|W_j\|_{\psi_d} \leq c \log n, 1 \leq j \leq m$. For any $w_n > 0$,

$$\mathrm{P}\left(\max_{1 \leq j \leq m} |W_j| \geq w_n\right) \leq \frac{c\,(\log n)(\log m)}{w_n}$$

Proof of 1. (i) (a) 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

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

$$\leq \quad K_d \left\{ \xi_2 \, \sigma + (1+K)^{1/d} C^{-1/d} \left[\sum_{i=1}^{n} |a_i|^{d'} \right]^{1/d'} \right\}.$$

For $1 < d \leq 2$, $d' = d/(d-1) \geq 2$. Thus $\sum_{i=1}^{n} |a_i|^{d'} = \sum_{i=1}^{n} |a_i|^{d'-2} |a_i|^2 \leq \xi_2^{d'-2} \sum_{i=1}^{n} a_i^2 \leq \xi_2^{d'}$. It follows that

$$\left\| \sum_{i=1}^{n} a_i \varepsilon_i \right\|_{\psi_d} \leq K_d \left\{ \xi_2 \, \sigma + (1+K)^{1/d} C^{-1/d} \xi_2 \right\}.$$

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

$$\begin{aligned} \left\| \sum_{i=1}^{n} a_i \varepsilon_i \right\|_{\psi_1} &\leq K_1 \left\{ \mathrm{E} \left| \sum_{i=1}^{n} a_i \varepsilon_i \right| + \left\| \max_{1 \leq i \leq n} |c_i \varepsilon_i| \right\|_{\psi_1} \right\} \\ &\leq K_1 \left\{ \xi_2 \, \sigma + K' \log(n) \max_{1 \leq i \leq n} \|a_i \varepsilon_i\|_{\psi_1} \right\} \\ &\leq K_1 \left\{ \xi_2 \, \sigma + K'(1+K) C^{-1} \log(n) \max_{1 \leq i \leq n} |c_i| \right\} \\ &\leq K_1 \left\{ \xi_2 \, \sigma + K'(1+K) C^{-1} \xi_2 \log(n) \right\}, \end{aligned}$$

where the last inequality follows from the inequality $\max_{1 \leq i \leq n} a_i^2 \leq \sum_{i=1}^{n} a_i^2 \leq \xi_2^2$.

(ii) (a) By Lemma 2.2.2 of Van der Vaart and Wellner (1996),

$$\left\| \max_{1 \leq j \leq m} W_j \right\|_{\psi_d} \leq K \, (\log m)^{1/d}$$

for a constant K. Because $\mathrm{E}|W| \leq (\log 2)^{1/d} \|W\|_{\psi_d}$ for any random variable W, we have

$$\mathrm{E}\left(\max_{1 \leq j \leq m_n} |W_j| \right) \leq K' (\log m_n)^{1/d},$$

where $K' = K(\log 2)^{1/d}$. By the Markov inequality, we have

$$\mathrm{P}\left(\max_{1 \leq j \leq m_n} |W_j| \geq w_n \right) \leq \frac{K' (\log m_n)^{1/d}}{w_n}.$$

(ii) (b) can be proved similarly. This completes the proof.

Lemma 2 *Let $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)$ be given as in Lemma 1.*

(i) Let $\boldsymbol{\eta}_n = 2n^{-1/2} \Sigma_{n11}^{-1} \mathbf{X}_1' \boldsymbol{\varepsilon}_n \equiv (\eta_{n1}, \ldots, \eta_{nk_n})'$. Then η_j has the same Orlicz norm property

as ε_1. *That is, for* $1 < d \leq 2$, $\|\eta_j\|_{\psi_d} \leq c$ *and, for* $d = 1$, $\|\eta_j\|_{\psi_1} \leq c \log n$.

(ii) Let $\zeta_n = 2n^{-1/2}\mathbf{X}'_{n2}(\mathbf{I} - H_n)\boldsymbol{\varepsilon}_n = (\zeta_1, \ldots, \zeta_{m_n})'$. *Then* ζ_j *has the same Orlicz norm property as* ε_1. *That is, for* $1 < d \leq 2$, $\|\zeta_j\|_{\psi_d} \leq c$ *and, for* $d = 1$, $\|\zeta_j\|_{\psi_1} \leq c \log n$.

Proof This lemma follows from the scaling of the covariates given in (4), condition (A4), and Lemma 1.

Proof of Theorem 1. By Lemma 1,

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_{n0}) \geq 1 - \mathrm{P}(A_n^c \cup B_n^c) \geq 1 - \mathrm{P}(A_n^c) - \mathrm{P}(B_n^c).$$

To prove the theorem, it suffices to show that $\mathrm{P}(A_n^c) \to 0$ and $\mathrm{P}(B_n^c) \to 0$.

We first consider $\mathrm{P}(A_n^c)$. Let $\boldsymbol{\eta}_n = 2n^{-1/2}\Sigma_{n11}^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}_n \equiv (\eta_{n1}, \ldots, \eta_{nk_n})'$. By Lemma 1, $\eta_{n1}, \ldots, \eta_{nk_n}$ are sub-Gaussian. Let $\mathbf{u}_n \equiv \Sigma_{n11}^{-1}\mathbf{s}_n$, where $\mathbf{s}_n = W_{n1}\mathrm{sgn}(\boldsymbol{\beta}_{n1})$. We can write

$$A_n^c = \left\{ \boldsymbol{\eta}_n \geq 2\sqrt{n}\ |\boldsymbol{\beta}_{n1}| - n^{-1/2}\lambda_n |\mathbf{u}_n| \right\}.$$

Then

$$\begin{aligned} \mathrm{P}(A_n^c) &= \mathrm{P}(A_n^c \cap \{|\mathbf{w}_{n1}| \leq c_1 b_{n1}^{-1}\}) + \mathrm{P}(A_n^c \cap \{|\mathbf{w}_{n1} > c_1 b_{n1}^{-1}\}) \\ &\leq \mathrm{P}(A_n^c \cap \{|\mathbf{w}_{n1}| \leq c_1 b_{n1}^{-1}\}) + \mathrm{P}(|\mathbf{w}_{n1}| > c_1 b_{n1}^{-1}), \end{aligned} \tag{14}$$

where $\{|\mathbf{w}_{n1}| \leq c_1 b_{n1}^{-1}\} = \{|w_{n1j}| \leq c_1 b_{n1}^{-1}, 1 \leq j \leq k_n\}$. By condition (A2), $\mathrm{P}(|\mathbf{w}_{n1}| > c_1 b_{n1}^{-1}) \to 0$. So it suffices to show that the first term on the right-hand side of (14) converges to zero.

Let $\tau_{n1} \leq \cdots \leq \tau_{nk_n}$ be the eigenvalues of Σ_{n11} and $\gamma_1, \ldots, \gamma_{k_n}$ the associated eigenvectors. By spectrum decomposition,

$$\Sigma_{n11}^{-1} = \sum_{j=1}^{k_n} \tau_{nj}^{-1}\gamma_j\gamma_j'.$$

Then

$$\mathbf{u}_n = \Sigma_{n11}^{-1}W_{n1}\mathrm{sgn}(\boldsymbol{\beta}_{n1}) = \Sigma_{n11}^{-1}\mathbf{s}_n = \sum_{j=1}^{k_n} \tau_{nj}^{-1}\gamma_j\gamma_j'\mathbf{s}_n.$$

The lth element of $\mathbf{u}_n$ is

$$u_l = \sum_{j=1}^{k_n} \tau_{nj}^{-1}(\gamma_j'\mathbf{s}_n)\gamma_{jl},\ l = 1, \ldots, k_n.$$

By the Cauchy-Schwartz inequality,

$$|u_l|^2 \leq \tau_{kn}^{-2} \sum_{j=1}^{k_n} (\gamma_j' \mathbf{s}_n)^2 \sum_{j=1}^{k_n} (\gamma_{jl})^2 \leq \tau_{n1}^{-2} \sum_{j=1}^{k_n} \|\gamma_j\|^2 \|\mathbf{s}_n\|^2 = \tau_{n1}^{-2} k_n \|\mathbf{s}_n\|^2. \tag{15}$$

By the definition of $\mathbf{s}_n$, $\|\mathbf{s}_n\|^2 = \|\mathbf{w}_{n1}\|^2$. On $\{|\mathbf{w}_{n1}| \leq c_1 b_{n1}^{-1}\}$

$$\|\mathbf{s}_n\|^2 \leq c_1 k_n b_{n1}^{-2}. \tag{16}$$

From (15) and (16), we have

$$|u_l| \leq \tau_{n1}^{-1} k_n w_{n1} \leq c_1 \tau_1^{-1} k_n b_{n1}^{-1},\ l = 1, \ldots, k_n.$$

Let $c_1' = c_1 \tau_1^{-1}$, $\nu_n = 2\sqrt{n} b_{n1} - c_1' n^{-1/2} \lambda_n k_n b_{n1}^{-1}$ and,

$$C_{n1} = \{|\eta_j| \geq \nu_n, j = 1, \ldots, k_n\} = \{\max_{1 \leq j \leq k_n} |\eta_j| \geq \nu_n\}.$$

By the definition of A_n^c,

$$A_n^c \cap \{|\mathbf{w}_{n1} \leq c_1 b_{n1}^{-1}\} \subseteq C_{n1}. \tag{17}$$

By Lemma 1, for $1 < d \leq 2$,

$$\mathrm{P}(C_{n1}) = \mathrm{P}\left(\max_{1 \leq j \leq k_n} |\eta_j| \geq \nu_n\right) \leq \frac{K'(\log k_n)^{1/2}}{\nu_n}.$$

Write

$$\frac{(\log 2)^{1/2} K \log(k_n)}{\nu_n} = \frac{K'(\log k_n)^{1/d}}{\sqrt{n} b_{n1}[2 - (\lambda_n k_n / n b_{n1}^2)]}.$$

Under condition (A3a),

$$\frac{(\log k_n)^{1/d}}{\sqrt{n}\, b_{n1}} \to 0, \quad \text{and} \quad \frac{\lambda_n k_n}{n b_{n1}^2} \to 0,$$

we have

$$\frac{(\log k_n)^{1/d}}{\nu_n} \to 0.$$

For $d = 1$,

$$\mathrm{P}(C_{n1}) = \mathrm{P}\left(\max_{1 \leq j \leq k_n} |\eta_j| \geq \nu_n\right) \leq \frac{c(\log n)(\log k_n)}{\nu_n}.$$

Under condition (A3b), $(\log n)(\log k_n)/nu_n \to 0$. It follows that $\mathrm{P}(C_{n1}) \to 0$. By (17), it follows that $\mathrm{P}(A_n^c \cap \{|\mathbf{w}_{n1}| \le c_1 b_{n1}^{-1}\}) \to 0$.

Now consider $\mathrm{P}(B_n^c)$. Let $\boldsymbol{\zeta}^n = 2n^{-1/2}\mathbf{X}'_{n2}(I - H_n)\boldsymbol{\varepsilon}_n = (\zeta_1^n, \ldots, \zeta_{m_n}^n)'$. Then by Lemma 1, $\zeta_1^n, \ldots, \zeta_{m_n}^n$ are sub-Gaussian.

Let $\mathbf{v}_n = \Sigma_{n21}\Sigma_{n11}^{-1}W_{n1}\mathrm{sgn}(\boldsymbol{\beta}_{10}) = \Sigma_{n21}\mathbf{u}_n$. We can write

$$B_n^c = \left\{|\boldsymbol{\zeta}_n| > n^{-1/2}\lambda_n \mathbf{w}_{n2} - n^{-1/2}\lambda_n |\mathbf{v}_n|\right\}.$$

Let $D_n = \{\mathbf{w}_{n2} > r_n, \mathbf{w}_{n1} \le c_1 b_{n1}^{-1}\}$. We have

$$\begin{aligned} \mathrm{P}(B_n^c) &= \mathrm{P}(B_n^c \cap D_n) + \mathrm{P}(B_n^c \cap D_n^c) \\ &\le \mathrm{P}(B_n^c \cap D_n) + \mathrm{P}(D_n^c). \end{aligned} \tag{18}$$

By conditions (A2),

$$\mathrm{P}(D_n^c) \le \mathrm{P}(\mathbf{w}_{n2} < r_n) + \mathrm{P}(\mathbf{w}_{n1} > c_1 b_{n1}^{-1}) \to 0.$$

It suffices to show that $\mathrm{P}(B_n^c \cap D_n) \to 0$. By the scaling of the covariates, the (j,l)th element of Σ_{n21}

$$\left| n^{-1}\sum_{i=1}^n x_{i2j}x_{i1l} \right| \le \left(n^{-1}\sum_{i=1}^n x_{i2j}^2 \cdot n^{-1}\sum_{i=1}^n x_{i1l}^2 \right)^{1/2} = 1.$$

On D_n, by the definition of $\mathbf{v}_n$ and (15), the jth element of $\mathbf{v}_n$

$$|v_{nj}| = n^{-1}\left| \sum_{l=1}^{k_n}\sum_{i=1}^n x_{i2l}x_{i1j}u_{nl} \right| \le \sum_{l=1}^{k_n} |u_{nl}| \le c_1' k_n^2 b_{n1}^{-1},\ j = 1, \ldots, m_n.$$

Let $\xi_n = n^{-1/2}\lambda_n r_n - n^{-1/2}\lambda_n c_1' k_n^2 b_{n1}^{-1}$ and

$$C_{n2} = \{|\boldsymbol{\zeta}_n| \le \xi_n\} = \left\{ \max_{1\le j\le m_n} |\zeta_{nj}| \le \xi_n \right\}.$$

Then $B_n^c \cap D_n \subseteq C_{n2}$. For $1 < d \le 2$, we have

$$\mathrm{P}(C_{n2}) = \mathrm{P}\left(\max_{1\le j\le m_n} |\zeta_{nj}| > \xi_n \right) \le \frac{(K'(\log m_n)^{1/d}}{\xi_n}.$$

Now

$$\frac{K'(\log m_n)^{1/d}}{\xi_n} = \frac{K'(\log m_n)^{1/d}}{n^{-1/2}\lambda_n\psi_n[1-(c_1'k_n^2/r_nb_{n1})]}.$$

Under condition (A3a),

$$\frac{\sqrt{n}(\log m_n)^{1/d}}{\lambda_n r_n} \to 0, \quad \text{and} \quad \frac{k_n^2}{r_n b_{n1}} \to 0.$$

For $d = 1$, we have

$$\mathrm{P}(C_{n2}) = \mathrm{P}\left(\max_{1\le j\le m_n} |\zeta_{nj}| > \xi_n\right) \le \frac{c(\log n)(\log m_n)}{\xi_n}.$$

Under condition (A3b), $(\log n)(\log m_n)/\xi_n \to 0$. Thus in either case, $\mathrm{P}(C_{n2}) \to 0$. It follows that $\mathrm{P}(B_n^c) \to 0$.

Proof of Theorem 3. Because $\boldsymbol{\beta}_{20} = \mathbf{0}$, we have

$$\widetilde{\beta}_j = n^{-1}\sum_{i=1}^n x_{ij}(\mathbf{x}'_{(1)i}\boldsymbol{\beta}_{10} + \mathbf{x}'_{(2)i}\boldsymbol{\beta}_{20} + \varepsilon_i) = n^{-1}\sum_{i=1}^n x_{ij}(\mathbf{x}'_{(1)i}\boldsymbol{\beta}_{10} + \varepsilon_i).$$

For $j \in J_{n0}$, let $\mu_{nj}^0 = \mathrm{E}\widetilde{\beta}_{nj}$. Write

$$\mu_{nj}^0 = n^{-1}\sum_{i=1}^n x_{ij}\mathbf{x}'_{(1)i}\boldsymbol{\beta}_{10} = n^{-1}\sum_{k=1}^{k_n}\sum_{i=1}^n x_{ij}x_{ik}\beta_{10k}. \tag{19}$$

Let $\mu_n^0 = \max_{j\in J_{n0}} |\mu_{nj}^0|$. Under condition (B2), $\mu_n^0 = O(k_n n^{-1/2})$.

For $j \in J_{n1}$, let $\mu_{nj}^1 = \mathrm{E}\widetilde{\beta}_{nj}$. Then

$$\mu_{nj}^1 = \xi_{nj} = n^{-1}\sum_{i=1}^n x_{ij}\mathbf{x}'_{(1)i}\boldsymbol{\beta}_{10}. \tag{20}$$

Let $\mu_n^1 = \min_{j\in J_{n1}} |\xi_{nj}^1|$. By condition (B3), $\mu_n^1 > 2\xi_r b_{n1}$.

We first show that

$$\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) \to 0. \tag{21}$$

We have

$$\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) \;=\; \mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj} - \mu_{nj} + \mu_{nj}| > C\right)$$

$$
\begin{aligned}
&\leq \mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj} - \mu^0_{nj}| > C - r_n\mu_n\right) \\
&= \mathrm{P}\left(\sqrt{n} \max_{j\in J_{n0}} |\widetilde{\beta}_{nj} - \mu_{nj}| > r_n^{-1}\sqrt{n}C - \sqrt{n}\mu_n\right). \qquad (22)
\end{aligned}
$$

Note that

$$
n^{1/2}(\widetilde{\beta}_j - \mu^0_{nj}) = n^{-1/2}\sum_{i=1}^{n} x_{ij}\varepsilon_i.
$$

By Lemma 1, for $1 < d \leq 2$,

$$
\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) \leq \frac{K' \log^{1/d}(m_n)}{r_n^{-1}n^{1/2} - n^{1/2}\mu_n} = O(n^{-1/2} r_n \log^{1/2}(m_n)),
$$

and for $d = 1$,

$$
\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) \leq \frac{K'(\log n)(\log m_n)}{r_n^{-1}n^{1/2} - n^{1/2}\mu_n} = O(n^{-1/2} r_n (\log n)(\log m_n)).
$$

Since $n^{1/2}|\mu_n| = O(k_n)$ and $r_n k_n / n^{1/2} \to 0$, for $1 < d \leq 2$,

$$
\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) = O(n^{-1/2} r_n \log^{1/d}(m_n)).
$$

For $d = 1$,

$$
\mathrm{P}\left(r_n \max_{j\in J_{n0}} |\widetilde{\beta}_{nj}| > C\right) = O(n^{-1/2} r_n (\log n)(\log m_n)).
$$

Therefore, under condition (B4), (21) follows.

Next, we show that $\mathrm{P}(\min_{j\in J_{n1}} |\widetilde{\beta}_{nj}| \geq \xi_r b_{n1}) \to 1$, or equivalently,

$$
\mathrm{P}(\min_{j\in J_{n1}} |\widetilde{\beta}_{nj}| < \xi_r b_{n1}) \to 0. \qquad (23)
$$

We have

$$
\begin{aligned}
\mathrm{P}\left(\min_{j\in J_{n1}} |\widetilde{\beta}_{nj}| < \xi_r b_{n1}\right) &= \mathrm{P}\left(\bigcup_{j\in J_{n1}} \left\{|\widetilde{\beta}_{nj}| < \xi_r b_{n1}\right\}\right) \\
&\leq \sum_{j\in J_{n1}} \mathrm{P}\left(|\widetilde{\beta}_{nj}| < \xi_r b_{n1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j \in J_{n1}} \mathrm{P}\left(|\widetilde{\beta}_{nj} - \mu^1_{nj} + \mu^1_{nj}| < \xi_r b_{n1}\right) \\
&\leq \sum_{j \in J_{n1}} \mathrm{P}\left(n^{1/2}|\mu^1_{nj}| - n^{1/2}|\widetilde{\beta}_{nj} - \mu^1_{nj}| < \xi_r n^{1/2} b_{n1}\right) \\
&= \sum_{j \in J_{n1}} \mathrm{P}\left(n^{1/2}|\widetilde{\beta}_{nj} - \mu^1_{nj}| > n^{1/2}\xi^1_n - \xi_r n^{1/2} b_{n1}\right) \\
&\leq k_n K \exp[-C(n^{1/2}\mu^1_n - \xi_r n^{1/2} b_{n1})^d] \\
&\leq k_n K \exp[-C\xi^d_r n^{d/2} b^d_{n1}].
\end{aligned}
$$

Thus under condition (B4), (23) follows. Now the theorem follows from (21) and (23).

REFERENCES

1. Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2004). Prediction by supervised principal components. Stanford University Department of Statistics Tech Report, Sept 2004.

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. Bülhman, P. (2004). Boosting for high-dimensional linear models. *Tech Report No. 120*, ETH Zürich, Switzerland.

4. 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.

5. Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M., and Sheffield, V. C. (2006). Homozygosity Mapping with SNP Arrays Identifies a Novel Gene for Bardet-Biedl Syndrome (BBS10). *Proceedings of the National Academy of Sciences* (USA), vol. 103, no. 16 6287-6292.

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

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

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

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.

13. Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

14. Hunter, D.R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. of Statist. To appear.*

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

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

17. Kosorok, M.R. and Ma, S. (2005). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. U.W. Madison Department of Biostatistics/Medical Informatics TR188. *Tentatively accepted, Ann. Statist.*

18. Leng, C., Lin, Y., and Wahba, G. (2004). A Note on the LASSO and Related Procedures in Model Selection. To appear in *Statistica Sinica*.

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

20. Portnoy, S. (1984). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: I. Consistency. *Ann. Statist.* **12** 1298-1309.

21. Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: II. Normal approximation. *Ann. Statist.* **13** 1403-1417.

22. 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*. September 26, 2006. Vol 103: 14429-14434.

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

24. van der Laan, M. J., and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2** 445–461.

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

26. Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. Technical report No. 702. Department of Statistics, University of California, Berkeley.

27. Zou, H. (2006). The Adaptive Lasso and its Oracle Properties.

28. 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
60 College Street P.O. Box 208034
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

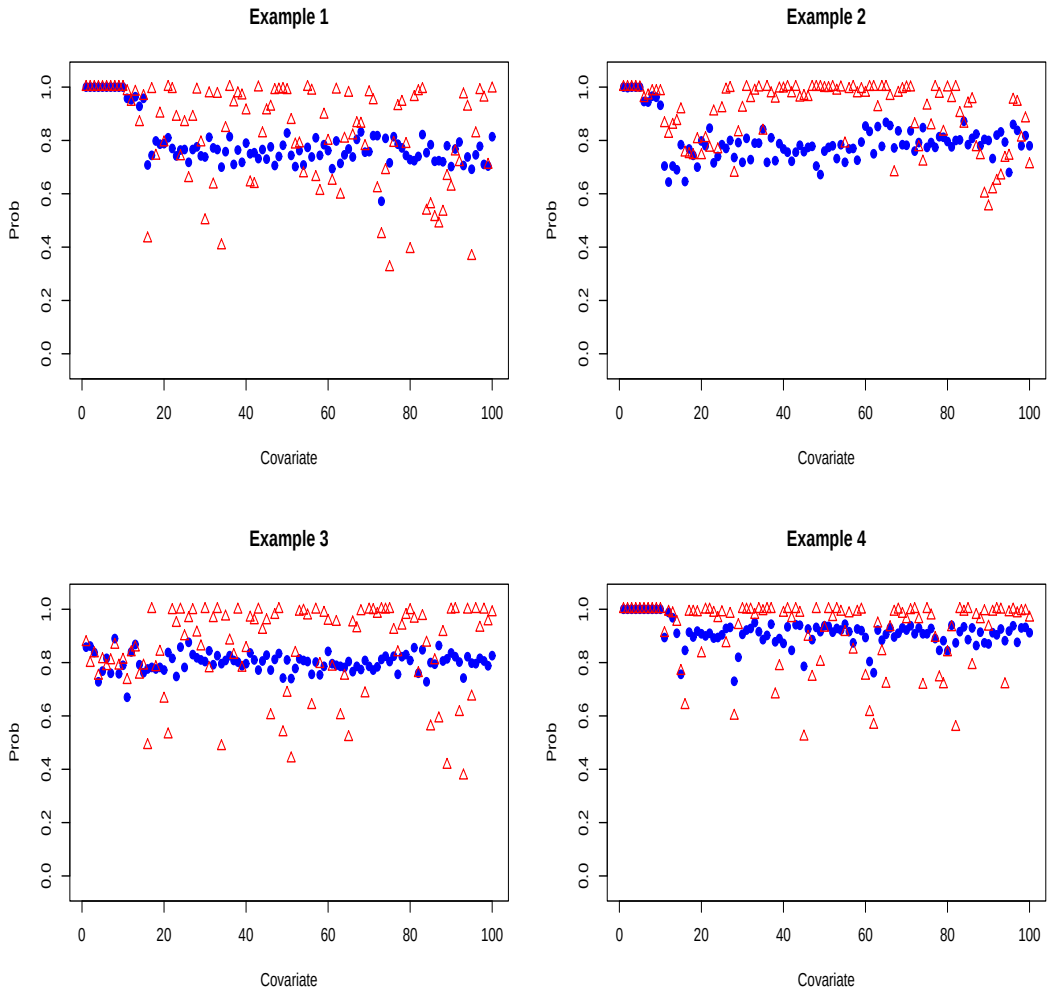

Figure 1: Simulation study (examples 1–4): probability of individual covariate effect being correctly identified. Circle (blue): LASSO; Triangle (red): adaptive Lasso.

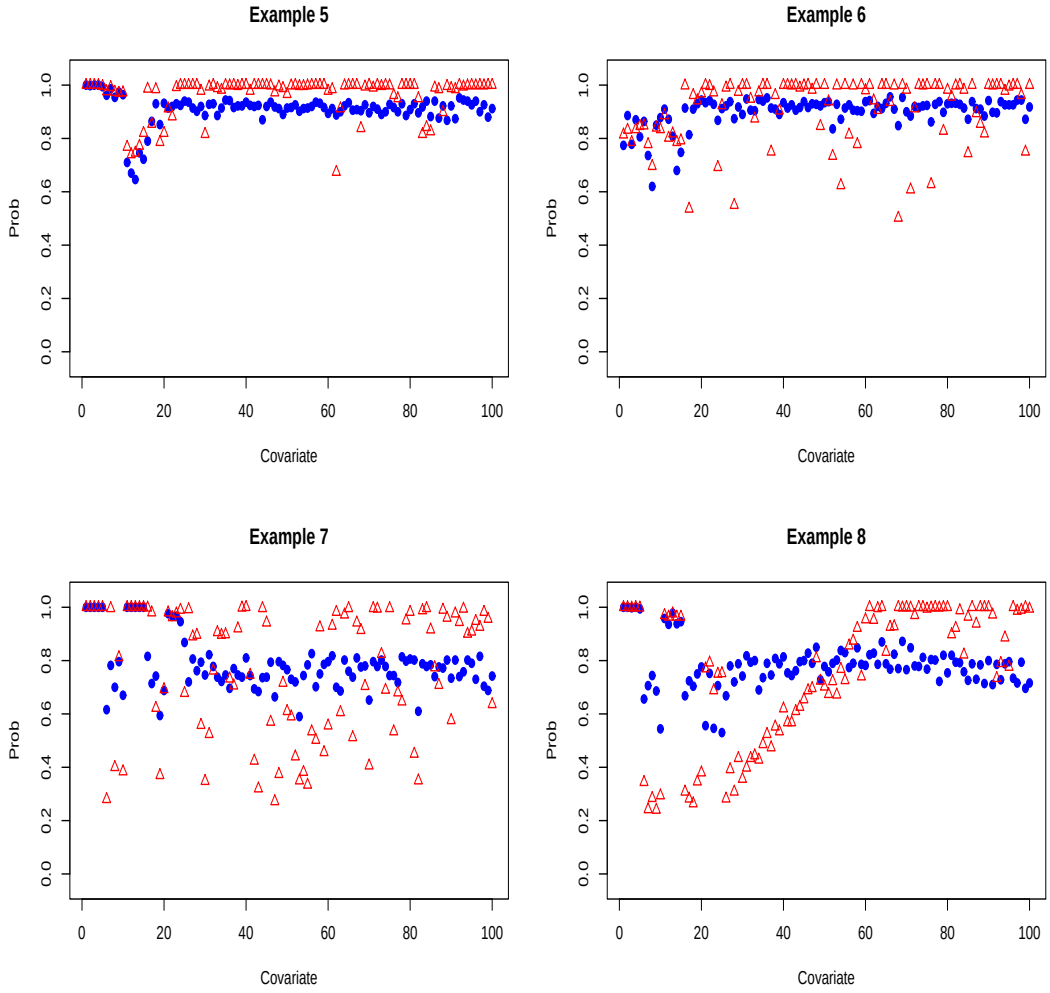

Figure 2: Simulation study (examples 5–8): probability of individual covariate effect being correctly identified. Circle (blue): LASSO; Triangle (red): adaptive Lasso.

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	3.148 (0.557)	48	2.982 (0.540)	40
4	3.604 (0.681)	50	3.369 (0.631)	43
5	3.304 (0.572)	50	2.887 (0.499)	33
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