Published in final edited form as:
Ann Stat. 2009 ; 37(4): 1733–1751. doi:10.1214/08-AOS625.

ON THE ADAPTIVE ELASTIC-NET WITH A DIVERGING NUMBER OF PARAMETERS

Hui Zou[*] and
School of Statistics, University of Minnesota, Minneapolis, Mn 55455, E-Mail: hzou@stat.umn.edu

Hao Helen Zhang[†]
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, E-Mail: hzhang2@stat.ncsu.edu

Abstract

We consider the problem of model selection and estimation in situations where the number of parameters diverges with the sample size. When the dimension is high, an ideal method should have the oracle property (Fan and Li, 2001; Fan and Peng, 2004) which ensures the optimal large sample performance. Furthermore, the high-dimensionality often induces the collinearity problem which should be properly handled by the ideal method. Many existing variable selection methods fail to achieve both goals simultaneously. In this paper, we propose the adaptive Elastic-Net that combines the strengths of the quadratic regularization and the adaptively weighted lasso shrinkage. Under weak regularity conditions, we establish the oracle property of the adaptive Elastic-Net. We show by simulations that the adaptive Elastic-Net deals with the collinearity problem better than the other oracle-like methods, thus enjoying much improved finite sample performance.

Keywords and phrases

Adaptive regularization; Elastic-Net; High dimensionality; Model selection; Oracle property; Shrinkage methods

1. Introduction

1.1. Background

Consider the problem of model selection and estimation in the classical linear regression model

$$\mathbf{y}=\mathbf{X}\beta^{*}+\epsilon, \tag{1.1}$$

where $\mathbf{y} = (y_1,\ldots,y_n)^T$ is the response vector and $\mathbf{x}_j = (x_{1j},\ldots,x_{nj})^T$, $j = 1,\ldots,p$, are the linearly independent predictors. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$ be the predictor matrix. Without loss of generality we assume the data are centered, so the intercept is not included in the regression function. Throughout this paper, we assume the errors are identically and independent distributed with zero mean and finite variance σ^2. We are interested in the sparse modeling problem where the true model has a sparse representation, i.e., some components of β^* are exactly zero. Let

[*] Zou's research is supported by National Science Foundation grant DMS 0706733.
[†] Zhang's research is supported by National Science Foundation grant DMS 0645293 and National Institutes of Health grant NIH/NCI R01 CA-085848.

$\mathcal{A}=\{j:\beta_j^* \neq 0, j=1,2,\dots,p\}$. In this work we call the size of $\mathcal{A}$ the intrinsic dimension of the underlying model. We wish to discover the set $\mathcal{A}$ and estimate the corresponding coefficients.

Variable selection is fundamentally important for knowledge discovery with high-dimensional data (Fan & Li 2006) and it could greatly enhance the prediction performance of the fitted model. Traditional model selection procedures follow best-subset selection and its step-wise variants. However, best-subset selection is computationally prohibitive when the number of predictors is large. Furthermore, as analyzed by Breiman (1996), subset selection is unstable, thus the resulting model has poor prediction accuracy. To overcome the fundamental drawbacks of subset selection, statisticians have recently proposed various penalization methods to perform simultaneous model selection and estimation. In particular, the lasso (Tibshirani 1996) and the SCAD (Fan & Li 2001) are two very popular methods due to their good computational and statistical properties. Efron, Hastie, Johnstone & Tibshirani (2004) proposed the LARS algorithm for computing the entire lasso solution path. Knight & Fu (2000) studied the asymptotic properties of the lasso. Fan & Li (2001) showed that the SCAD enjoys the oracle property, that is, the SCAD estimator can perform as well as the oracle if the penalization parameter is appropriately chosen.

1.2. Two fundamental issues with the ℓ_1 penalty

The lasso estimator (Tibshirani 1996) is obtained by solving the ℓ_1 penalized least squares problem

$$\widehat{\beta}(\text{lasso})=\arg\min_{\beta} \|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1, \tag{1.2}$$

where $\|\beta\|_1=\sum_{j=1}^{p}|\beta_j|$ is the ℓ_1-norm of β. The ℓ_1 penalty enables the lasso to simultaneously regularize the least squares fit and shrink some components of $\hat{\beta}$(lasso) to zero for some appropriately chosen λ. The entire lasso solution paths can be computed by the LARS algorithm (Efron et al. 2004). These nice properties make the lasso a very popular variable selection method.

Despite its popularity the lasso does have two serious drawbacks: namely the lack of oracle property and instability with high-dimensional data. First of all, the lasso does not have the oracle property. Fan & Li (2001) first pointed out that asymptotically the lasso has non-ignorable bias for estimating the nonzero coefficients. They further conjectured that the lasso may not have the oracle property because of the bias problem. This conjecture was recently proven in Zou (2006). Zou (2006) further showed that the lasso could be inconsistent for model selection unless the predictor matrix (or the design matrix) satisfies a rather strong condition. Zou (2006) proposed the following adaptive lasso estimator

$$\widehat{\beta}(\text{AdaLasso})=\arg\min_{\beta} \|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\sum_{j=1}^{p}\widehat{w}_j|\beta_j|, \tag{1.3}$$

where $\{\widehat{w}_j\}_{j=1}^{p}$ are the adaptive data-driven weights and can be computed by $\widehat{w}_j=(|\widehat{\beta}_j^{ini}|)^{-\gamma}$, where γ is a positive constant and $\hat{\beta}^{ini}$ is an initial root-*n* consistent estimate of β. Zou (2006) showed that with an appropriately chosen λ, the adaptive lasso performs as well as the oracle. Candes, Wakin & Boyd (2007) used the adaptive lasso idea to enhance sparsity in sparse signal recovery via the reweighted ℓ_1 minimization.

Secondly the ℓ_1 penalization methods can have very poor performance when there are highly correlated variables in the predictor set. The collinearity problem is often encountered in high-dimensional data analysis. Even when the predictors are independent, as long as the dimension is high, the maximum sample correlation can be large, as shown in Fan & Lv (2007). Collinearity can severely degrade the performance of the lasso. As shown in Zou & Hastie (2005), the lasso solution paths are unstable when predictors are highly correlated. Zou & Hastie (2005) proposed the Elastic-Net as an improved version of the lasso for analyzing high-dimensional data. The Elastic-Net estimator is defined as follows:

$$\widehat{\beta}(\text{enet})=(1+\frac{\lambda_2}{n})\left\{\arg\min_{\beta} \| \mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda_2 \|\beta\|_2^2+\lambda_1 \|\beta\|_1\right\}. \tag{1.4}$$

If the predictors are standardized (each variable has mean zero and L_2-norm one), then we should change $(1+\frac{\lambda_2}{n})$ to $(1 + \lambda_2)$ as in Zou & Hastie (2005). The ℓ_1 part of the Elastic-Net performs automatic variable selection while the ℓ_2 part stabilizes the solution paths and hence improves the prediction. In an orthogonal design where the lasso is shown to be optimal (Donoho, Johnstone, Kerkyacharian & Picard 1995), the Elastic-Net automatically reduces to the lasso. However, when the correlations among the predictors become high, the Elastic-Net can significantly improve the prediction accuracy of the lasso.

1.3. The adaptive Elastic-Net

The adaptively weighted ℓ_1 penalty and the Elastic-Net penalty improve the lasso in two different directions. The adaptive lasso achieves the oracle property of the SCAD and the Elastic-Net handles the collinearity. However, following the arguments in Zou & Hastie (2005) and Zou (2006), we can easily see that the adaptive lasso inherits the instability of the lasso for high-dimensional data, while the Elastic-Net is lack of the oracle property. Thus, it is natural to consider combining the ideas of the adaptively weighted ℓ_1 penalty and the Elastic-Net regularization to obtain a better method which can improve the lasso in both directions. To this end, we propose the adaptive Elastic-Net that penalizes the squared error loss using a combination of the ℓ_2 penalty and the adaptive ℓ_1 penalty. Since the adaptive Elastic-Net is designed for high-dimensional data analysis, we study its asymptotic properties under the assumption that the dimension diverges with the sample size.

Pioneering papers on asymptotic theories with diverging number of parameters include Huber (1988) and Portnoy (1984) which studied the M-estimators. Recently, Fan, Peng & Huang (2005) studied a semi-parametric model with a growing number of nuisance parameters, whereas Lam & Fan (2007) investigated the profile likelihood ratio inference for the growing number of parameters. In particular, our work is influenced by Fan & Peng (2004) who studied the oracle property of nonconcave penalized likelihood estimators. Fan & Peng (2004) provocatively argued why it is important to study the validity of the oracle property when the dimension diverges. We would like to know whether the adaptive Elastic-Net enjoys the oracle property with a diverging number of predictors. This question will be thoroughly investigated in this paper.

The rest of the paper is organized as follows. In Section 2 we introduce the adaptive Elastic-Net. Statistical theory, including the oracle property, of the adaptive Elastic-Net is established in Section 3. In Section 4 we use simulation to compare the finite sample performance of the adaptive Elastic-Net with the SCAD and other competitors. Section 5 discusses how to combine SIS of Fan & Lv (2007) and the adaptive Elastic-Net to deal with the ultra-high dimension cases. Technical proofs are presented in Section 6.

2. Method

The adaptive Elastic-Net can be viewed as a combination of the Elastic-Net and the adaptive lasso. Suppose we first compute the Elastic-Net estimator $\hat{\beta}(\text{enet})$ as defined in (1.4), and then we construct the adaptive weights by

$$\widehat{w}_j=(|\widehat{\beta}_j(\text{enet})|)^{-\gamma}, \ j=1,2,\ldots,p, \tag{2.1}$$

where γ is a positive constant. Now we solve the following optimization problem to get the adaptive Elastic-Net estimates

$$\widehat{\beta}(\text{AdaEnet})=(1+\frac{\lambda_2}{n})\left\{\underset{\beta}{\arg\min} \parallel \mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda_2 \parallel \beta\|_2^2+\lambda_1^*\sum_{j=1}^{p}\widehat{w}_j|\beta_j|\right\}. \tag{2.2}$$

From now on, we write $\hat{\beta} = \hat{\beta}(\text{AdaEnet})$ for the sake of convenience.

If we force λ_2 to be zero in (2.2), then the adaptive Elastic-Net reduces to the adaptive lasso. Following the arguments in Zou & Hastie (2005), we can easily show that in an orthogonal design the adaptive Elastic-Net reduces to the adaptive lasso, regardless the value of λ_2. This is desirable because in that setting the adaptive lasso achieves the optimal minimax risk bound (Zou 2006). The role of the ℓ_2 penalty in (2.2) is to further regularize the adaptive lasso fit whenever the collinearity may cause serious trouble.

We know the Elastic-Net naturally adopts a sparse representation. One can use $\hat{w}_j = (|\hat{\beta}_j(\text{enet})| + 1/n)^{-\gamma}$ to avoid dividing zeros. We can also define $\hat{w}_j = \infty$ when $\hat{\beta}_j(\text{enet}) = 0$. Let $\mathscr{A}_{\text{enet}} = \{j : \hat{\beta}_j\ (\text{enet}) \neq 0\}$ and $\widehat{\mathcal{A}}^c_{\text{enet}}$ denotes its complement set. Then we have $\widehat{\beta}_{\widehat{\mathcal{A}}^c_{\text{enet}}}=0$ and

$$\widehat{\beta}_{\widehat{\mathcal{A}}_{\text{enet}}}=(1+\frac{\lambda_2}{n})\left\{\underset{\beta}{\arg\min} \parallel \mathbf{y}-\mathbf{X}_{\widehat{\mathcal{A}}_{\text{enet}}}\beta\|_2^2+\lambda_2 \parallel \beta\|_2^2+\lambda_1^*\sum_{j\in\widehat{\text{A}}_{\text{enet}}}\widehat{w}_j|\beta_j|\right\}. \tag{2.3}$$

where β in (2.3) is a vector of length $|\mathscr{A}_{\text{enet}}|$, the size of $\mathscr{A}_{\text{enet}}$.

The ℓ_1 regularization parameters, λ_1^* and λ_1, are directly responsible for the sparsity of the estimates. Their values are allowed to be different. On the other hand, we use the same λ_2 for the ℓ_2 penalty component in the Elastic-Net and the adaptive Elastic-Net estimators, because the ℓ_2 penalty offers the same kind of contribution in both estimators.

3. Statistical Theory

In our theoretical analysis, we assume the following regularity conditions throughout.

(A1) We use λ_{min} ($\mathbf{M}$) and λ_{max} ($\mathbf{M}$) to denote the minimum and maximum eigenvalues of a positive definite matrix $\mathbf{M}$, respectively. Then we assume

$$b \le \lambda_{min}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \le \lambda_{max}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \le B$$

where b and B are two positive constants.

(A2)

$$\lim_{n\to\infty}\frac{\max_{i=1,2,\ldots,n}\sum_{j=1}^{p}x_{ij}^2}{n}=0.$$

(A3) $E[|\epsilon|^{2+\delta}] < \infty$ for some $\delta > 0$.

(A4)

$$\lim_{n\to\infty}\frac{\log(p)}{\log(n)}=\nu \text{ for some } 0 \leq \nu < 1.$$

To construct the adaptive weights ($\hat{w}$), we take a fixed γ such that $\gamma>\frac{2\nu}{1-\nu}\cdot$ In our numerical studies we let $\gamma=\lceil\frac{2\nu}{1-\nu}\rceil+1$ to avoid the tuning on γ. Once γ is chosen, we choose the regularization parameters according to the following conditions

(A5)

$$\lim_{n\to\infty}\frac{\lambda_2}{n}=0,\ \lim_{n\to\infty}\frac{\lambda_1}{\sqrt{n}}=0,$$

and

$$\lim_{n\to\infty}\frac{\lambda_1^*}{\sqrt{n}}=0,\ \lim_{n\to\infty}\frac{\lambda_1^*}{\sqrt{n}}n^{\frac{(1-\nu)(1+\gamma)-1}{2}}=\infty.$$

(A6)

$$\lim_{n\to\infty}\frac{\lambda_2}{\sqrt{n}}\sqrt{\sum_{j\in\mathcal{A}}\beta_j^{*2}}=0,\ \lim_{n\to\infty}\min(\frac{n}{\lambda_1\sqrt{p}},\left(\frac{\sqrt{n}}{\sqrt{p}\lambda_1^*}\right)^{\frac{1}{\gamma}})(\min_{j\in\mathcal{A}}|\beta_j^*|)\to\infty.$$

Conditions (A1) and (A2) assume the predictor matrix has a reasonably good behavior. Similar conditions were considered in Portnoy (1984). Note that in the linear regression setting, condition (A1) is exactly condition (F) in Fan & Peng (2004). Condition (A3) is used to establish the asymptotic normality of $\hat{\beta}$ (AdaEnet).

It is worth pointing out that condition (A4) is weaker than that used in Fan & Peng (2004) in which p is assumed to satisfy $p^4/n \to 0$ or at most $p^3/n \to 0$. It means their results require $\nu<\frac{1}{3}\cdot$ Our theory removes this limitation. For any $0 \leq \nu < 1$, we can choose an appropriate γ to construct the adaptive weights and the oracle property holds as long as $\gamma>\frac{2\nu}{1-\nu}$. Also note that in the finite dimension setting $\nu = 0$, thus any positive γ can be used, which agrees with the results in Zou (2006).

Condition (A6) is similar to condition (H) in Fan & Peng (2004). Basically, condition (A6) allows the nonzero coefficients to vanish but at a rate that can be distinguished by the penalized least squares. In the finite dimension setting the condition is implicitly assumed.

THEOREM 3.1

Given the data ($\mathbf{y}, \mathbf{X}$), *let* $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_p)$ *be a vector whose components are all non-negative and can depend on* ($\mathbf{y}, \mathbf{X}$). *Define*

$$\widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1) = \arg\min_{\beta} \left\{ \| \mathbf{y} - \mathbf{X}\beta \|_2^2 + \lambda_2 \| \beta \|_2^2 + \lambda_1 \sum_{j=1}^{p} \widehat{w}_j |\beta_j| \right\},$$

for non-negative parameters λ_2 *and* λ_1. *If* $\hat{w}_j = 1$ *for all j, we denote* $\hat{\beta}_{\hat{w}}(\lambda_2, \lambda_1)$ *by* $\hat{\beta}(\lambda_2, \lambda_1)$ *for convenience*.

*If we assume the model (*1.1*) and condition (A1), then*

$$E\left(\| \widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1) - \beta^* \|_2^2 \right) \leq 4 \frac{\lambda_2^2 \| \beta^* \|_2^2 + Bpn\sigma^2 + \lambda_1^2 E(\sum_{j=1}^{p} \widehat{w}_j^2)}{(bn + \lambda_2)^2}.$$

In particular, when $\hat{w}_j = 1$ *for all j, we have*

$$E\left(\| \widehat{\beta}(\lambda_2, \lambda_1) - \beta^* \|_2^2 \right) \leq 4 \frac{\lambda_2^2 \| \beta^* \|_2^2 + Bpn\sigma^2 + \lambda_1^2 p}{(bn + \lambda_2)^2}.$$

It is worth mentioning that the derived risk bounds are non-asymptotic. Theorem 3.1 is very useful for the asymptotic analysis. A direct corollary of Theorem 3.1 is that, under conditions (A1)–(A6), $\hat{\beta}(\lambda_2, \lambda_1)$ is a root-(*n/p*)-consistent estimator. This consistent rate is the same as the result of SCAD (Fan & Peng 2004). The root-(*n/p*) consistency result suggests that it is appropriate to use the Elastic-Net to construct the adaptive weights.

THEOREM 3.2

Let us write $\beta^* = (\beta^*_{\mathcal{A}}, 0)$ *and define*

$$\tilde{\beta}^*_{\mathcal{A}} = \arg\min_{\beta} \left\{ \| \mathbf{y} - \mathbf{X}_{\mathcal{A}}\beta \|_2^2 + \lambda_2 \sum_{j \in \mathcal{A}} \beta_j^2 + \lambda_1^* \sum_{j \in \mathcal{A}} \widehat{w}_j |\beta_j| \right\}. \tag{3.1}$$

Then with probability tending to 1, $((1 + \frac{\lambda_2}{n}) \tilde{\beta}^*_{\mathcal{A}}, 0)$ is the solution to (2.2).

Theorem 3.2 provides an asymptotic characterization of the solution to the adaptive Elastic-Net criterion. The definition of $\tilde{\beta}^*_{\mathcal{A}}$ borrows the concept of "oracle" (Donoho & Johnstone 1994, Fan & Li 2001, Fan & Peng 2004, Zou 2006). If there was an oracle informing us the true subset model, then we would use this oracle information and the adaptive Elastic-Net criterion would become that in (2.3). Theorem 3.2 tells us that asymptotically speaking, the adaptive Elastic-Net works as if it had such oracle information. Theorem 3.2 also suggests that the adaptive Elastic-Net should enjoy the oracle property, which is confirmed in the next theorem.

THEOREM 3.3

Under conditions (A1)–(A6), the adaptive Elastic-Net has the oracle property, that is, the estimator $\hat{\beta}(\text{AdaEnet})$ *must satisfy:*

1. *Consistency in selection :* $\Pr(\{j : \hat{\beta}(\text{AdaEnet})_j \neq 0\} = \mathcal{A}) \to 1$,
2. *Asymptotic normality :* $\alpha^T \frac{\mathbf{I}+\lambda_2 \sum_{\mathcal{A}}^{-1}}{1+\frac{\lambda_2}{n}} \sum_{\mathcal{A}}^{\frac{1}{2}} \left(\widehat{\beta}(\text{AdaEnet})_{\mathcal{A}} - \beta^*_{\mathcal{A}}\right) \to_d N(0,\sigma^2)$, where $\sum_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$ *and* α *is a vector of norm 1.*

By Theorem 3.3 the selection consistency and the asymptotic normality of the adaptive Elastic-Net are still valid when the number of parameters diverges. Technically speaking, the selection consistency result is stronger than that Theorem 3.2 implies, although Theorem 3.2 plays an important role in the proof of Theorem 3.3. As a special case, when we let $\lambda_2 = 0$, which is a choice satisfying conditions (A5) and (A6), Theorem 3.3 tell us that the adaptive lasso enjoys the selection consistency and the asymptotical normality:

$$\alpha^T \sum_{\mathcal{A}}^{\frac{1}{2}} \left(\widehat{\beta}(\text{AdaLasso})_{\mathcal{A}} - \beta^*_{\mathcal{A}}\right) \to_d N(0,\sigma^2).$$

4. Numerical Studies

In this section we present simulations to study the finite sample performance of the adaptive Elastic-Net. We considered five methods in the simulation study: the lasso(Lasso), the Elastic-Net(Enet), the adaptive lasso(ALasso), the adaptive Elastic-Net(AEnet) and the SCAD. In our implementation, we let $\lambda_2 = 0$ in the adaptive Elastic-Net to get the adaptive lasso fit. There are several commonly used tuning parameter selection methods, such as cross-validation, generalized cross-validation(GCV), AIC and BIC. Zou, Hastie & Tibshirani (2007) suggested using BIC to select the lasso tuning parameter. Wang, Li & Tsai (2007) showed that for the SCAD, BIC is a better tuning parameter selector than GCV and AIC. In this work, we used BIC to select the tuning parameter for each method.

Fan & Peng (2004) considered simulation models in which $p_n = [4n^{\frac{1}{4}}] - 5$ and $|\mathcal{A}| = 5$. Our theory allows $p_n = O(n^{\nu})$ for any $\nu < 1$. Thus, we are interested in models in which $p_n = O(n^{\nu})$ with $\nu > \frac{1}{3}$. In addition, we allow the intrinsic dimension $(\mathcal{A})$ to diverge with the sample size as well, because such designs make the model selection and estimation more challenging than in the fixed $|\mathcal{A}|$ situations.

Example 1

We generated data from the linear regression model,

$$y = \mathrm{x}^T \beta^* + \epsilon,$$

where β^* is a p-dim vector and $\varepsilon \sim N(0, \sigma^2)$, $\sigma = 6$ and $\mathbf{x}$ follows a p-dim multivariate normal distribution with zero mean and covariance Σ whose (j, k) entry is $\Sigma_{j,k} = \rho^{|j-k|}$ $1 \leq k, j \leq p$. We considered $\rho = 0.5$ and $\rho = 0.75$. Let $p = p_n = [4n^{1/2}] - 5$ for $n = 100,200,400$. Let $\mathbf{1}_m/\mathbf{0}_m$ denote a *m*-vector of 1s/0s. The true coefficients are $\beta^* = (3 \cdot \mathbf{1}_q, 3 \cdot \mathbf{1}_q, 3 \cdot \mathbf{1}_q, \mathbf{0}_{p-3q})^T$ and $|\mathscr{A}| = 3q$ and $q = [p_n/9]$. In this example $\nu=\frac{1}{2}$, hence we used $\gamma = 3$ for computing the adaptive weights in the adaptive Elastic-Net.

For each estimator $\hat{\beta}$, its estimation accuracy is measured by the mean squared error (MSE) defined as $E[(\hat{\beta} - \beta^*)^T \Sigma(\hat{\beta} - \beta^*)]$. The variable selection performance is gauged by (C, IC), where C is the number of zero coefficients that are correctly estimated by zero and IC is the number of nonzero coefficients that are incorrectly estimated by zero.

Table 1 documents the simulation results. Several interesting observations can be made.

1. When the sample size is large ($n = 400$), the three oracle-like estimators outperform the lasso and the Elastic-Net which do not have the oracle property. That is expected according to the asymptotic theory.
2. The SCAD and the adaptive Elastic-Net are the best when the sample size is large and the correlation is moderate. However, the SCAD can perform much worse than the adaptive Elastic-Net when the correlation is high ($\rho = 0.75$) or the sample size is small.
3. Both the Elastic-Net and the adaptive lasso can do significantly better than the lasso. What is more interesting is that the adaptive Elastic-Net often outperforms the Elastic-Net and the adaptive lasso.

Example 2

We considered the same setup as in example 1, except that we let $p = p_n = [4n^{2/3}] - 5$ for $n =$ 100,200,800. Since $\nu=\frac{2}{3}$, we used $\gamma = 5$ for computing the adaptive weights in the adaptive Elastic-Net and the adaptive lasso. The estimation problem in this example is even more difficult than that in example 1. To see why, note that when $n = 200$ the dimension increases from 51 in example 1 to 131 in this example, and the intrinsic dimension ($|\mathscr{A}|$) is almost tripled.

The simulation results are presented in Table 2 from which we can see that the three observations made in example 1 are still valid in this example. Furthermore, we see that for every combination of $(n,p, |\mathscr{A}|,\rho)$, the adaptive Elastic-Net has the best performance.

5. Ultra-high dimensional data

In this section we discuss how the adaptive Elastic-Net can be applied to ultra-high dimensional data in which $p > n$. When p is much larger than n, Candes & Tao (2007) suggested using the Dantzig selector which can achieve the ideal estimation risk up to a $\log(p)$ factor under the uniform uncertainty condition. Fan & Lv (2007) showed that the uniform uncertainty condition may easily fail and the $\log(p)$ factor is too large when p is exponentially large. Moreover, the computational cost of the Dantzig selector would be very high when p is large. In order to overcome these difficulties, Fan & Lv (2007) introduced the Sure Independence Screening (SIS) idea which reduces the ultra-high dimensionality to a relatively large scale d_n but $d_n < n$. Then, the lower dimension methods such as the SCAD can be used to estimate the sparse model. This procedure is referred to as SIS+SCAD. Under regularity conditions, Fan & Lv (2007) proved that SIS misses true features with an exponentially small probability and SIS +SCAD holds the oracle property if $d_n=o(n^{\frac{1}{3}})$. Furthermore, with the help of SIS, the Dantzig selector can achieve the ideal risk up to a $\log(d_n)$ factor, rather than the original $\log(p)$.

Inspired by the results of Fan & Lv (2007), we consider combining the adaptive Elastic-Net and SIS when $p > n$. We first apply SIS to reduce the dimension to d_n and then fit the data by using the adaptive Elastic-Net. We call this procedure SIS+AEnet.

THEOREM 5.1

Suppose the conditions for Theorem 1 in Fan and Lv (2007) *hold. Let $d_n = O(n^{\nu})$, $\nu < 1$, then SIS+AEnet produces an estimator that holds the oracle property.*

We make a note here that Theorem 5.1 is a direct consequence of Theorem 1 in Fan & Lv (2007) and Theorem 3.3, thus its proof is omitted. Theorem 5.1 is similar to Theorem 5 in Fan & Lv (2007), but there is a difference. SIS+AEnent can hold the oracle property when d_n exceeds $O(n^{\frac{1}{3}})$, while Theorem 5 in Fan & Lv (2007) assumes $d_n=o(n^{\frac{1}{3}})$.

To demonstrate SIS+AEnet, we consider the simulation example used in Fan & Lv (2007) (Section 3.3.1). The model is $y = \mathbf{x}^T \beta^* + 1.5N(0,1)$, where $\beta^*=(\beta_1^T, 0_{p-|\mathscr{A}|})^T$ with $|\mathscr{A}| = \forall$. Here β_1 is a 8-dim vector and each component has the form $(-1)^u (a_n + |z|)$, where $a_n=4\log(n)/\sqrt{n}$, u is randomly drawn from $Ber(0.4)$ and z is randomly drawn from the standard normal distribution. We generated $n = 200$ data from the above model. Before applying the adaptive Elastic-Net, we used SIS to reduce the dimensionality from 1000 to $d_n=[5.5n^{\frac{2}{3}}]=188.$ The estimation problem is still rather challenging, as we need to estimate 188 parameters by using only 200 observations. From Table 3 we see that SIS+AEnet performs favorably compared to SIS+SCAD.

6. Proofs

PROOF OF THEOREM 3.1

We write

$$\widehat{\beta}(\lambda_2, 0)=\arg\min_{\beta} \| \mathbf{y} - \mathbf{X}\beta\|_2^2+\lambda_2 \| \beta\|_2^2$$

By the definition of $\hat{\beta}_{\hat{w}}(\lambda_2, \lambda_1)$ and $\hat{\beta}(\lambda_2, 0)$, we know

$$\begin{aligned} & \| \mathbf{y} - \mathbf{X}\widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1)\|_2^2+\lambda_2 \| \widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1)\|_2^2 \\ \geq \quad & \| \mathbf{y} - \mathbf{X}\widehat{\beta}(\lambda_2, 0)\|_2^2+\lambda_2 \| \widehat{\beta}(\lambda_2, 0)\|_2^2, \end{aligned}$$

and

$$\begin{aligned} & \| \mathbf{y} - \mathbf{X}\widehat{\beta}(\lambda_2, 0)\|_2^2+\lambda_2 \| \widehat{\beta}(\lambda_2, 0)\|_2^2+\lambda_1 \sum_{j=1}^{p} \widehat{w}_j|\widehat{\beta}(\lambda_2, 0)_j| \\ \geq \quad & \| \mathbf{y} - \mathbf{X}\widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1)\|_2^2+\lambda_2 \| \widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1)\|_2^2+\lambda_1 \sum_{j=1}^{p} \widehat{w}_j|\widehat{\beta}_{\widehat{w}}(\lambda_2, \lambda_1)_j|. \end{aligned}$$

From the above two inequalities, we have

$$\begin{aligned}\lambda_1 \sum_{j=1}^{p} \widehat{w}_j (|\widehat{\beta}(\lambda_2,0)_j| - |\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)_j|) \quad \geq \quad & (\| \mathbf{y} - \mathbf{X}\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2^2 + \lambda_2 \, \| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2^2) \\ & -(\| \mathbf{y} - \mathbf{X}\widehat{\beta}(\lambda_2,0)\|_2^2 + \lambda_2 \, \| \widehat{\beta}(\lambda_2,0)\|_2^2).\end{aligned} \tag{6.1}$$

On the other hand, we have

$$\begin{aligned} & (\| \mathbf{y} - \mathbf{X}\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2^2 + \lambda_2 \, \| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2^2) - (\| \mathbf{y} - \mathbf{X}\widehat{\beta}(\lambda_2,0)\|_2^2 + \lambda_2 \, \| \widehat{\beta}(\lambda_2,0)\|_2^2) \\ = \; & (\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0))^T (\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I})(\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0)),\end{aligned}$$

and

$$\begin{aligned}\sum_{j=1}^{p} \widehat{w}_j (|\widehat{\beta}(\lambda_2,0)_j| - |\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)_j|) \quad & \leq \sum_{j=1}^{p} \widehat{w}_j |\widehat{\beta}(\lambda_2,0)_j - \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)_j| \\ & \leq \sqrt{\sum_{j=1}^{p} \widehat{w}_j^2} \, \| \widehat{\beta}(\lambda_2,0) - \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2.\end{aligned}$$

Note that $\lambda_{min}(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}) = \lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2$. Therefore, we end up with

$$\begin{aligned} & (\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2) \, \| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0)\|_2^2 \\ \leq \; & (\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0))^T (\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I})(\widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0)) \\ \leq \; & \lambda_1 \sqrt{\sum_{j=1}^{p} \widehat{w}_j^2} \, \| \widehat{\beta}(\lambda_2,0) - \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1)\|_2,\end{aligned} \tag{6.2}$$

which results in the following inequality

$$\| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2,\lambda_1) - \widehat{\beta}(\lambda_2,0)\|_2 \leq \frac{\lambda_1 \sqrt{\sum_{j=1}^{p} \widehat{w}_j^2}}{\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2}. \tag{6.3}$$

Note that

$$\widehat{\beta}(\lambda_2,0) - \beta^* = -\lambda_2 (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1} \beta^* + (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1} \mathbf{X}^T \epsilon,$$

which implies that

$$\begin{aligned} & E\left(\| \widehat{\beta}(\lambda_2,0) - \beta^* \|_2^2\right) \\ \leq \; & 2\lambda_2^2 \, \| (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1} \beta^* \|_2^2 + 2E\left(\| (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1} \mathbf{X}^T \epsilon \|_2^2\right) \\ \leq \; & 2\lambda_2^2 (\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2)^{-2} \, \| \beta^* \|_2^2 + 2(\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2)^{-2} E\left(\epsilon^T \mathbf{X}\mathbf{X}^T \epsilon\right) \\ = \; & 2(\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2)^{-2} (\lambda_2^2 \, \| \beta^* \|_2^2 + Tr(\mathbf{X}^T\mathbf{X})\sigma^2) \\ \leq \; & 2(\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2)^{-2} (\lambda_2^2 \, \| \beta^* \|_2^2 + p\lambda_{max}(\mathbf{X}^T\mathbf{X})\sigma^2).\end{aligned} \tag{6.4}$$

Combing (6.3) and (6.4), we have

$$\begin{aligned} & E\left(\| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2, \lambda_1) - \beta^* \|_2^2\right) \\ \leq \; & 2E\left(\| \widehat{\beta}(\lambda_2, 0) - \beta^* \|_2^2\right) + 2E\left(\| \widehat{\beta}_{\widehat{\boldsymbol{w}}}(\lambda_2, \lambda_1) - \widehat{\beta}(\lambda_2, 0) \|_2^2\right) \\ \leq \; & \frac{4\lambda_2^2 \|\beta^*\|_2^2 + 4p\lambda_{max}(\mathbf{X}^T\mathbf{X})\sigma^2 + 2\lambda_1^2 E[\sum_{j=1}^{p} \widehat{w}_j^2]}{(\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda_2)^2} \end{aligned} \qquad (6.5)$$

$$\leq \; 4\frac{\lambda_2^2 \; \| \beta^* \|_2^2 + Bpn\sigma^2 + \lambda_1^2 E[\sum_{j=1}^{p} \widehat{w}_j^2]}{(bn + \lambda_2)^2}. \qquad (6.6)$$

We have used condition (A1) in the last inequality. When $\hat{w}_j = 1$ for all j, we have

$$E\left(\| \widehat{\beta}(\lambda_2, \lambda_1) - \beta^* \|_2^2\right) \; \leq \; 4\frac{\lambda_2^2 \; \| \beta^* \|_2^2 + Bpn\sigma^2 + p\lambda_1^2}{(bn + \lambda_2)^2}.$$

PROOF OF THEOREM 3.2

We show that $((1+\frac{\lambda_2}{n})\tilde{\beta}^*_{\mathcal{A}}, 0)$ satisfies the Karush-Kuhn-Tucker (KKT) conditions of (2.2) with probability tending to 1. By the definition of $\tilde{\beta}^*_{\mathcal{A}}$, it suffices to show

$$\Pr(\forall j \in \mathcal{A}^c, | - 2X_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\beta}^*_{\mathcal{A}})| \leq \lambda_1^* \widehat{w}_j) \to 1,$$

or equivalently

$$\Pr(\exists j \in \mathcal{A}^c, | - 2X_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\beta}^*_{\mathcal{A}})| > \lambda_1^* \widehat{w}_j) \to 0.$$

Let $\eta = \min_{j \in \mathcal{A}}(|\beta_j^*|)$ and $\widehat{\eta} = \min_{j \in \mathcal{A}}(|\widehat{\beta}(\text{enet})_j^*|)$. We note that

$$\begin{aligned} & \Pr(\exists j \in \mathcal{A}^c, | - 2X_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\beta}^*_{\mathcal{A}})| > \lambda_1^* \widehat{w}_j) \\ \leq \; & \sum_{j \in \mathcal{A}^c} \Pr(| - 2X_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\beta}^*_{\mathcal{A}})| > \lambda_1^* \widehat{w}_j, \tilde{\eta} > \eta/2) + \Pr(\tilde{\eta} \leq \eta/2) \end{aligned}$$

$$\Pr(\tilde{\eta} \leq \eta/2) \leq \Pr(\| \widehat{\beta}(\text{enet}) - \beta^* \|_2 \geq \eta/2) \leq \frac{E(\| \widehat{\beta}(\text{enet}) - \beta^* \|_2^2)}{\eta^2/4}.$$

Then by Theorem 3.1 we obtain

$$\Pr(\widehat{\eta} \leq \eta/2) \leq 16\frac{\lambda_2^2 \; \| \beta^* \|_2^2 + Bpn\sigma^2 + \lambda_1^2 p}{(bn + \lambda_2)^2 \eta^2}. \qquad (6.7)$$

Moreover, let $M=\left(\frac{\lambda_1^*}{n}\right)^{\frac{1}{1+\gamma}}$ and we have

$$\begin{aligned}
& \sum_{j\in\mathcal{A}^c}\Pr(|-2X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|>\lambda_1^*\widehat{w}_j,\widehat{\eta}>\eta/2)\\
\leq\ & \sum_{j\in\mathcal{A}^c}\Pr(|-2X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|>\lambda_1^*\widehat{w}_j,\widehat{\eta}>\eta/2,|\widehat{\beta}(\text{enet})_j|\leq M)\\
& +\sum_{j\in\mathcal{A}^c}\Pr(|\widehat{\beta}(\text{enet})_j|>M)\\
\leq\ & \sum_{j\in\mathcal{A}^c}\Pr(|-2X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|>\lambda_1^*M^{-\gamma},\widehat{\eta}>\eta/2)+\sum_{j\in\mathcal{A}^c}\Pr(|\widehat{\beta}(\text{enet})_j|>M)\\
\leq\ & \frac{4M^{2\gamma}}{\lambda_1^{*2}}E\left(\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2 I(\widehat{\eta}>\eta/2)\right)+\frac{1}{M^2}E\left(\sum_{j\in\mathcal{A}^c}|\widehat{\beta}(\text{enet})_j|^2\right)\\
\leq\ & \frac{4M^{2\gamma}}{\lambda_1^{*2}}E\left(\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2 I(\widehat{\eta}>\eta/2)\right)+\frac{E\left(\|\widehat{\beta}(\text{enet})-\beta^*\|_2^2\right)}{M^2}\\
\leq\ & \frac{4M^{2\gamma}}{\lambda_1^{*2}}E\left(\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2 I(\widehat{\eta}>\eta/2)\right)+4\frac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^2 p}{(bn+\lambda_2)^2M^2},
\end{aligned} \tag{6.8}$$

where we have used Theorem 3.1 in the last step. By the model assumption, we have

$$\begin{aligned}
\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2 &= \sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{X}_{\mathcal{A}}\beta_{\mathcal{A}}^*-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)+X_j^T\epsilon|^2\\
&\leq 2\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{X}_{\mathcal{A}}\beta_{\mathcal{A}}^*-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2+2\sum_{j\in\mathcal{A}^c}|X_j^T\epsilon|^2\\
&\leq 2Bn\parallel\mathbf{X}_{\mathcal{A}}(\beta_{\mathcal{A}}^*-\tilde{\beta}_{\mathcal{A}}^*)\|_2^2+2Bn\parallel\epsilon\|_2^2\\
&\leq 2Bn\cdot Bn\parallel\beta_{\mathcal{A}}^*-\tilde{\beta}_{\mathcal{A}}^*\|_2^2+2Bn\parallel\epsilon\|_2^2,
\end{aligned}$$

which gives us the below inequality

$$\begin{aligned}
& E\left(\sum_{j\in\mathcal{A}^c}|X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}^*)|^2 I(\widehat{\eta}>\eta/2)\right)\\
\leq\ & 2B^2n^2E\left(\parallel\beta_{\mathcal{A}}^*-\tilde{\beta}_{\mathcal{A}}^*\|_2^2 I(\widehat{\eta}>\eta/2)\right)+2Bn\sigma^2.
\end{aligned} \tag{6.9}$$

We now bound $E\left(\parallel\beta_{\mathcal{A}}^*-\tilde{\beta}_{\mathcal{A}}^*\|_2^2 I(\widehat{\eta}>\eta/2)\right)$. Let

$$\tilde{\beta}_{\mathcal{A}}^*(\lambda_2,0)=\arg\min_{\beta}\left\{\parallel\mathbf{y}-\mathbf{X}_{\mathcal{A}}\beta\|_2^2+\lambda_2\sum_{j\in\mathcal{A}}\beta_j^2\right\}.$$

Then by using the same arguments for deriving (6.1), (6.2) and (6.3), we have

$$\parallel\tilde{\beta}_{\mathcal{A}}^*-\tilde{\beta}_{\mathcal{A}}^*(\lambda_2,0)\|_2\leq\frac{\lambda_1^*\cdot\max_{j\in\mathcal{A}}\widehat{w}_j\sqrt{|\mathcal{A}|}}{\lambda_{min}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})+\lambda_2}\leq\frac{\lambda_1^*\widehat{\eta}^{-\gamma}\sqrt{p}}{bn+\lambda_2} \tag{6.10}$$

Note that $\lambda_{min}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})\geq\lambda_{min}(\mathbf{X}^T\mathbf{X})\geq bn$ and $\lambda_{max}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})\leq\lambda_{max}(\mathbf{X}^T\mathbf{X})\leq Bn.$ Following the rest arguments in the proof of Theorem 3.1, we obtain

$$\begin{aligned} & E\left(\| \beta^*_{\mathcal{A}} - \tilde{\beta}^*_{\mathcal{A}}\|_2^2 I(\widehat{\eta}>\eta/2)\right) \\ \le\ & 4\frac{\lambda_2^2\|\beta^*_{\mathcal{A}}\|_2^2+\lambda_{max}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})|\mathcal{A}|\sigma^2+\lambda_1^{*2}(\eta/2)^{-2\gamma}|\mathcal{A}|}{(\lambda_{min}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})+\lambda_2)^2} \\ \le\ & 4\frac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^{*2}(\eta/2)^{-2\gamma}p}{(bn+\lambda_2)^2}. \end{aligned} \tag{6.11}$$

The combination of (6.7), (6.8), (6.9) and (6.11) yields

$$\begin{aligned} & \Pr(\exists j \in \mathcal{A}^c, |-2X_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\beta}^*_{\mathcal{A}})|>\lambda_1^*\widehat{w}_j) \\ \le\ & \frac{4M^{2\gamma}n}{\lambda_1^{*2}}\left(8B^2n\frac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^{*2}(\eta/2)^{-2\gamma}p}{(bn+\lambda_2)^2}+2B\sigma^2\right) \\ & +\frac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^2p}{(bn+\lambda_2)^2}\frac{4}{M^2} \\ & +\frac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^2p}{(bn+\lambda_2)^2}\frac{16}{\eta^2} \\ \widehat{\equiv}\ & K_1+K_2+K_3. \end{aligned}$$

We have chosen $\gamma>\frac{2\nu}{1-\nu}$, then under conditions (A1)–(A6) it follows that

$$\begin{aligned} K_1&=O(\left(\frac{\lambda_1^*}{\sqrt{n}}n^{\frac{(1+\gamma)(1-\nu)-1}{2}}\right)^{-\frac{2}{1+\gamma}}) \to 0, \\ K_2&=O(\frac{p}{n}\left(\frac{n}{\lambda_1^*}\right)^{\frac{2}{1+\gamma}}) \to 0, \\ K_3&=O(\frac{p}{n}\frac{1}{\eta^2})=O(\left(\lambda_1^*\sqrt{\frac{p}{n}}\eta^{-\gamma}\right)^{\frac{2}{\gamma}}\left(\frac{p}{n}\left(\frac{n}{\lambda_1^*}\right)^{\frac{2}{1+\gamma}}\right)^{\frac{1+\gamma}{\gamma}}p^{-\frac{2}{\gamma}}) \to 0. \end{aligned} \tag{6.12}$$

Thus the proof is completed.

PROOF OF THEOREM 3.3

From Theorem 3.2 we have shown that with probability tending to 1 the adaptive Elastic-Net estimator is equal to $((1+\frac{\lambda_2}{n})\tilde{\beta}^*_{\mathcal{A}}, 0).$ Therefore, in order to prove the model selection consistency result, we only need to show $\Pr\left(\min_{j\in\mathcal{A}}|\tilde{\beta}^*_j|>0\right) \to 1.$ By (6.10) we have

$$\min_{j\in\mathcal{A}}|\tilde{\beta}^*_j|>\min_{j\in\mathcal{A}}|\tilde{\beta}^*(\lambda_2,0)_j| - \frac{\lambda_1^*\sqrt{p}\widehat{\eta}^{-\gamma}}{bn+\lambda_2}.$$

Note that

$$\min_{j\in\mathcal{A}}|\tilde{\beta}^*(\lambda_2,0)_j|>\min_{j\in\mathcal{A}}|\beta^*_j|- \| \tilde{\beta}^*_{\mathcal{A}}(\lambda_2,0) - \beta^*_{\mathcal{A}}\|_2$$

Following (6.6) it is easy to see that

$$E\left(\| \tilde{\beta}^*_{\mathcal{A}}(\lambda_2,0) - \beta^*_{\mathcal{A}}\|_2^2\right) \le 4\frac{\lambda_2^2\ \|\beta^*\|_2^2+Bpn\sigma^2}{(bn+\lambda_2)^2}=O(\frac{p}{n})$$

Moreover, $\frac{\lambda_1^* \sqrt{p}\widehat{\eta}^{-\gamma}}{bn+\lambda_2} = O(\frac{1}{\sqrt{n}})\,(\frac{\lambda_1^* \sqrt{p}}{\sqrt{n}}\eta^{-\gamma})(\frac{\hat{\eta}}{\eta})^{-\gamma}$ and

$$\begin{aligned} E\left((\tfrac{\widehat{\eta}}{\eta})^2\right) &\le 2+\tfrac{2}{\eta^2}E\left((\widehat{\eta}-\eta)^2\right) \\ &\le 2+\tfrac{2}{\eta^2}E\left(\|\widehat{\beta}(\lambda_2,\lambda_1)-\beta^*\|_2^2\right) \\ &\le 2+\tfrac{8}{\eta^2}\,\tfrac{\lambda_2^2\|\beta^*\|_2^2+Bpn\sigma^2+\lambda_1^2 p}{(bn+\lambda_2)^2}. \end{aligned}$$

In (6.12) we have shown $\eta^2 \frac{n}{p} \to \infty$. Thus

$$\frac{\lambda_1^* \sqrt{p}\widehat{\eta}^{-\gamma}}{bn+\lambda_2} = o(\frac{1}{\sqrt{n}})O_P(1). \tag{6.13}$$

Hence, we have

$$\min_{j\in\mathcal{A}}|\tilde{\beta}_j^*| > \eta - \sqrt{\frac{p}{n}}O_P(1) - o(\frac{1}{\sqrt{n}})O_P(1),$$

and $\Pr\left(\min_{j\in\mathcal{A}}|\tilde{\beta}_j^*|>0\right) \to 1.$

We now prove the asymptotic normality. For convenience write

$$z_n = \alpha^T \frac{\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1}}{1+\frac{\lambda_2}{n}} \sum_{\mathcal{A}}^{\frac{1}{2}} \left(\widehat{\beta}(\text{AdaEnet})_{\mathcal{A}} - \beta_{\mathcal{A}}^*\right).$$

Note that

$$\begin{aligned} &\alpha^T(\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1})\sum_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}_{\mathcal{A}}^* - \frac{\beta_{\mathcal{A}}^*}{1+\frac{\lambda_2}{n}}\right) \\ = \;& \alpha^T(\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1})\sum_{\mathcal{A}}^{\frac{1}{2}}\frac{\lambda_2\beta_{\mathcal{A}}^*}{n+\lambda_2} + \alpha^T(\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1})\sum_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}_{\mathcal{A}}^* - \tilde{\beta}_{\mathcal{A}}^*(\lambda_2,0)\right) \\ &+ \alpha^T(\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1})\sum_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}_{\mathcal{A}}^*(\lambda_2,0) - \beta_{\mathcal{A}}^*\right). \end{aligned}$$

In addition, we have

$$(\mathbf{I}+\lambda_2\sum_{\mathcal{A}}^{-1})\sum_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}_{\mathcal{A}}^*(\lambda_2,0) - \beta_{\mathcal{A}}^*\right) = -\lambda_2\sum_{\mathcal{A}}^{-\frac{1}{2}}\beta_{\mathcal{A}}^* + \sum_{\mathcal{A}}^{-\frac{1}{2}}\mathbf{X}_{\mathcal{A}}^T\epsilon.$$

Therefore, by Theorem 3.2 it follows that with probability tending to 1, $z_n = T_1 + T_2 + T_3$, where

$$\begin{aligned} T_1 &= \alpha^T(\mathbf{I}+\lambda_2\sum\nolimits_{\mathcal{A}}^{-1})\sum\nolimits_{\mathcal{A}}^{\frac{1}{2}}\frac{\lambda_2\beta^*_{\mathcal{A}}}{n+\lambda_2} - \alpha^T\lambda_2\sum\nolimits_{\mathcal{A}}^{-\frac{1}{2}}\beta^*_{\mathcal{A}}, \\ T_2 &= \alpha^T(\mathbf{I}+\lambda_2\sum\nolimits_{\mathcal{A}}^{-1})\sum\nolimits_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}^*_{\mathcal{A}} - \tilde{\beta}^*_{\mathcal{A}}(\lambda_2, 0)\right), \\ T_3 &= \alpha^T\sum\nolimits_{\mathcal{A}}^{-\frac{1}{2}}\mathbf{X}^T_{\mathcal{A}} \in . \end{aligned}$$

We now show that $T_1 = o(1)$, $T_2 = o_P(1)$ and $T_3 \to N(0, \sigma^2)$ in distribution. Then by Slutsky's theorem we know $z_n \to_d N(0, \sigma^2)$. By (A1) and $\alpha^T\alpha = 1$, we have

$$\begin{aligned} T_1^2 &\le 2\,\|\,(\mathbf{I}+\lambda_2\sum\nolimits_{\mathcal{A}}^{-1})\sum\nolimits_{\mathcal{A}}^{\frac{1}{2}}\frac{\lambda_2\beta^*_{\mathcal{A}}}{n+\lambda_2}\|_2^2+2\,\|\,\lambda_2\sum\nolimits_{\mathcal{A}}^{-\frac{1}{2}}\beta^*_{\mathcal{A}}\|_2^2 \\ &\le 2\frac{\lambda_2^2}{(n+\lambda_2)^2}\,\|\,\sum\nolimits_{\mathcal{A}}^{\frac{1}{2}}\beta^*_{\mathcal{A}}\|_2^2(1+\frac{\lambda_2}{bn})^2+2\lambda^2\,\|\,\beta^*_{\mathcal{A}}\|_2^2\frac{1}{bn} \\ &\le \frac{2\lambda_2^2Bn}{(n+\lambda_2)^2}(1+\frac{\lambda_2}{bn})^2\,\|\,\beta^*_{\mathcal{A}}\|_2^2+2\lambda^2\,\|\,\beta^*_{\mathcal{A}}\|_2^2\frac{1}{bn}. \end{aligned}$$

Hence it follows by (A6) that $T_1 = o(1)$. Similarly, we can bound T_2 as follows

$$\begin{aligned} T_2^2 &\le (1+\frac{\lambda_2}{bn})^2\,\|\,\sum\nolimits_{\mathcal{A}}^{\frac{1}{2}}\left(\tilde{\beta}^*_{\mathcal{A}} - \tilde{\beta}^*_{\mathcal{A}}(\lambda_2, 0)\right)\|_2^2 \\ &\le (1+\frac{\lambda_2}{bn})^2Bn\,\|\,\tilde{\beta}^*_{\mathcal{A}} - \tilde{\beta}^*_{\mathcal{A}}(\lambda_2, 0)\|_2^2 \\ &\le (1+\frac{\lambda_2}{bn})^2Bn\left(\frac{\lambda_1^*\widehat{\eta}^{-\gamma}}{bn+\lambda_2}\right)^2 \end{aligned}$$

where we have used (6.10) in the last step. Then (6.13) tells us that $T_2^2=\frac{1}{n^2}O_P(1)$. Next we consider T_3. Let $\mathbf{X}_{\mathscr{A}}[i,]$ denote the ith row of the matrix $\mathbf{X}_{\mathscr{A}}$. With such notation we can write $T_3=\sum_{i=1}^n r_i\epsilon_i$, where $r_i=\alpha^T(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-\frac{1}{2}}(\mathbf{X}_{\mathcal{A}}[i,])^T$. Then it is easy to see that

$$\begin{aligned} \sum_{i=1}^n r_i^2 &= \sum_{i=1}^n \alpha^T(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-\frac{1}{2}}(\mathbf{X}_{\mathcal{A}}[i,])^T(\mathbf{X}_{\mathcal{A}}[i,])\,(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-\frac{1}{2}}\alpha \\ &= \alpha^T(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-\frac{1}{2}}(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-\frac{1}{2}}\alpha \\ &= \alpha^T\alpha=1. \end{aligned} \tag{6.14}$$

Furthermore, we have for $k = 2 + \delta$, $\delta > 0$

$$\sum_{i=1}^n E[\,|\epsilon_i|^{2+\delta}]|r_i^{2+\delta}| \;\le\; E[\,|\epsilon|^{2+\delta}](\sum_{i=1}^n|r_i^2|(\max_i|r_i|^\delta))=E[\,|\epsilon|^{2+\delta}](\max_i|r_i^2|)^{\frac{\delta}{2}}.$$

Note that $r_i^2 \;\le\|\,\sum\nolimits_{\mathcal{A}}^{-\frac{1}{2}}(\mathbf{X}_{\mathcal{A}}[i,])^T \;\le\; (\sum\nolimits_{j\in\mathcal{A}}x_{ij}^2)(\lambda_{max}(\sum\nolimits_{\mathcal{A}}^{-1})) \;\le\; \frac{\sum_{j=1}^p x_{ij}^2}{bn}$. Hence,

$$\sum_{i=1}^n E[\,|\epsilon_i|^{2+\delta}]|r_i^{2+\delta}| \;\le\; E[\,|\epsilon|^{2+\delta}]\left(\frac{\max_i(\sum_{j=1}^p x_{ij}^2)}{bn}\right)^{\frac{\delta}{2}} \to 0. \tag{6.15}$$

From (6.14) and (6.15) Lyapunov conditions for the central limit theorem are established. Thus, $T_3 \to_d N(0, \sigma^2)$. This completes the proof.

Acknowledgments

We sincerely thank an associate editor and referees for their help comments and suggestions.

REFERENCES

Breiman L. ‘Heuristics of instability and stabilization in model selection’. The Annals of Statistics 1996;24:2350–2383.

Candes E, Tao T. ‘The dantzig selector: statistical estimation when *p* is much larger than *n*’. The Annals of Statistics, to appear. 2007

Candes, E.; Wakin, M.; Boyd, S. California Institute of Technology; 2007. Enhancing sparsity by reweighted ℓ_1 minimization, Technical report.

Donoho D, Johnstone I. ‘Ideal spatial adaptation via wavelet shrinkage’. Biometrika 1994;81:425–455.

Donoho D, Johnstone I, Kerkyacharian G, Picard D. ‘Wavelet shrinkage: asymptopia? (with discussion)’. Journal of Royal Statistical Society, Series B 1995;57:301–337.

Efron B, Hastie T, Johnstone I, Tibshirani R. ‘Least angle regression’. The Annals of Statistics 2004;32:407–499.

Fan J, Li R. ‘Variable selection via nonconcave penalized likelihood and its oracle properties’. Journal of the American Statistical Association 2001;96:1348–1360.

Fan J, Li R. ‘Statistical challenges with high dimensionality: Feature selection in knowledge discovery’. Proceedings of the Madrid International Congress of Mathematicians 2006 2006;Vol. III:595–622.

Fan, J.; Lv, J. Department of Operations Research and Financial Engineering, Princeton University; 2007. Sure independence screening for ultra-high dimensional feature space, Technical report.

Fan J, Peng H. ‘Nonconcave penalized likelihood with a diverging number of parameters’. The Annals of Statistics 2004;32:928–961.

Fan J, Peng H, Huang T. ‘Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency (with discussion)’. Journal of the American Statistical Association 2005;100:781–813.

Huber P. ‘Robust regression: Asymptotics, conjectures and monte carlo’. The Annals of Statistics 1988;1:799–821.

Knight K, Fu W. ‘Asymptotics for lasso-type estimators’. The Annals of Statistics 2000;28:1356–1378.

Lam C, Fan J. ‘Profile-kernel likelihood inference with diverging number of parameters’. The Annals of Statistics. 2007 to appear.

Portnoy S. ‘Asymptotic behavior of M-estimatiors of *p* regression parameters when p^2/n is large. I. consistency’. The Annals of Statistics 1984;12:1298–1309.

Tibshirani R. ‘Regression shrinkage and selection via the lasso’. Journal of the Royal Statistical Society, Series B 1996;58:267–288.

Wang H, Li R, Tsai C. ‘Tuning parameter selectors for the smoothly clipped absolute deviation method’. Biometrika 2007;94:553–568. [PubMed: 19343105]

Zou H. ‘The adaptive lasso and its oracle properties’. Journal of the American Statistical Association 2006;101:1418–1429.

Zou H, Hastie T. ‘Regularization and variable selection via the elastic net’. Journal of the Royal Statistical Society, Series B 2005;67:301–320.

Zou H, Hastie T, Tibshirani R. ‘On the degrees of freedom of the lasso’. The Annals of Statistics 2007;35:2173–2192.

TABLE 1

Simulation I: model selection and fitting results based on 100 replications.

$\rho = 0.5$	***n***	**p_n**	**$\lvert\mathscr{A}\rvert$**	**Model**	**MSE**	**C**	**IC**
	100	35	9	Truth		26	0
				Lasso	7.57 (0.31)	24.08	0.01
				ALasso	6.78 (0.42)	25.50	0.42
				Enet	5.91 (0.29)	24.06	0.00
				AEnet	5.07 (0.35)	25.47	0.15
				SCAD	10.55 (0.68)	22.54	0.35
	200	51	15	Truth		36	0
				Lasso	6.63 (0.24)	33.32	0.00
				ALasso	3.78 (0.18)	35.46	0.02
				Enet	4.86 (0.19)	33.36	0.00
				AEnet	3.46 (0.17)	35.47	0.01
				SCAD	4.76 (0.33)	34.63	0.10
	400	75	24	Truth		51	0
				Lasso	4.99 (0.15)	47.31	0.00
				ALasso	2.76 (0.09)	50.33	0.00
				Enet	3.37 (0.12)	48.00	0.00
				AEnet	2.47 (0.08)	50.45	0.00
				SCAD	2.42 (0.09)	50.88	0.00
$\rho = 0.75$	*n*	p_n	$\lvert\mathscr{A}\rvert$	*Model*	MSE	C	IC
	100	35	9	Truth		26	0
				Lasso	5.93 (0.26)	24.80	0.14
				ALasso	8.49 (0.39)	25.76	1.84
				Enet	4.18 (0.24)	24.77	0.05
				AEnet	5.24 (0.32)	25.70	0.74
				SCAD	11.59 (0.56)	22.46	1.34
	200	51	15	Truth		36	0

$\rho = 0.5$	n	p_n	$\lvert\mathscr{A}\rvert$	Model	MSE	C	IC
				Lasso	5.10 (0.18)	34.66	0.02
				ALasso	5.32 (0.31)	35.70	0.87
				Enet	3.79 (0.17)	34.79	0.00
				AEnet	3.32 (0.17)	35.80	0.19
				SCAD	5.99 (0.31)	33.10	0.35
	400	75	24	Truth		51	0
				Lasso	3.83 (0.12)	49.03	0.00
				ALasso	2.85 (0.12)	50.53	0.09
				Enet	3.24 (0.11)	49.07	0.00
				AEnet	2.71 (0.09)	50.54	0.03
				SCAD	3.64 (0.17)	48.43	0.09

TABLE 2

Example 2: model selection and fitting results based on 100 replications.

$\rho = 0.5$	n	p_n	$\|\mathscr{A}\|$	**Model**	**MSE**	**C**	**IC**
	100	81	27	Truth		54	0
				Lasso	31.73 (1.06)	47.06	0.19
				ALasso	28.78 (1.22)	53.01	2.12
				Enet	27.61 (1.04)	46.35	0.13
				AEnet	20.27 (0.94)	53.00	1.15
				SCAD	44.88 (2.65)	47.79	2.37
	200	131	42	Truth		89	0
				Lasso	23.41 (0.67)	80.51	0.00
				ALasso	12.70 (0.48)	87.99	0.14
				Enet	18.94 (0.61)	80.27	0.00
				AEnet	10.68 (0.37)	87.97	0.00
				SCAD	14.14 (0.64)	87.42	0.25
	800	339	111	Truth		228	0
				Lasso	13.72 (0.23)	212.10	0.00
				ALasso	6.44 (0.12)	226.61	0.00
				Enet	11.02 (0.18)	213.91	0.00
				AEnet	6.00 (0.10)	226.75	0.00
				SCAD	7.79 (0.30)	228.00	0.33
$\rho = 0.75$	n	p_n	$\|\mathscr{A}\|$	Model	MSE	C	IC
	100	81	27	Truth		54	0
				Lasso	22.04 (0.73)	50.74	0.71
				ALasso	33.98 (1.08)	53.73	7.19
				Enet	17.37 (0.62)	50.82	0.46
				AEnet	16.18 (0.80)	53.67	2.36
				SCAD	31.84 (1.77)	50.55	4.74
	200	131	42	Truth		89	0

$\rho = 0.5$	n	p_n	$\lvert\mathscr{A}\rvert$	Model	MSE	C	IC
				Lasso	16.71 (0.50)	85.17	0.06
				ALasso	20.98 (0.92)	88.64	3.98
				Enet	14.12 (0.48)	85.35	0.05
				AEnet	11.16 (0.46)	88.60	0.87
				SCAD	15.27 (0.61)	87.20	1.33
	800	339	111	Truth		228	0
				Lasso	10.01 (0.16)	221.74	0.00
				ALasso	6.39 (0.12)	226.89	0.00
				Enet	8.01 (0.13)	222.74	0.00
				AEnet	6.23 (0.11)	226.94	0.00
				SCAD	6.62 (0.17)	228.00	0.29

TABLE 3

A demonstration of SIS+AEnet: model selection and fitting results based on 100 replications.

$d_n = [5.5n^{2/3}]$	Model	MSE	C	IC
188	Truth		992	0
	SIS+AEnet	0.71 (0.18)	987.45	0.05
	SIS+SCAD	1.48 (0.90)	982.20	0.06