TR 682

Model Selection via Standard Error Adjusted Adaptive Lasso

Wei Qian and Yuhong Yang
School of Statistics
University of Minnesota

Abstract. The adaptive lasso is a model selection method shown to be both consistent in variable selection and asymptotically normal in coefficient estimation. The actual variable selection performance of the adaptive lasso depends on the weight used. It turns out that the weight assignment using the OLS estimate (OLS-adaptive lasso) can result in very poor performance when collinearity of the model matrix is a concern. To achieve better variable selection results, we take into account the standard errors of the OLS estimate for weight calculation, and propose two different versions of the adaptive lasso denoted by SEA-lasso and NSEA-lasso. We show through numerical studies that when the predictors are highly correlated, SEA-lasso and NSEA-lasso can outperform OLS-adaptive lasso under a variety of linear regression settings while maintaining the same theoretical properties of the adaptive lasso.

Key Words: BIC; model selection consistency; solution path; variable selection.

1 Introduction

Reliable variable selection is an important problem in statistical learning. In solving problems with a number of possible predictors, it is desirable for a variable selection method to produce a parsimonious model that describes the pattern of the data well. Exhaustive subset selection with traditional information criteria with possible modifications can generate sparse results, but it is not computationally feasible when the number of predictors is large.

The lasso (Tibshirani, 1996) is a popular approach to achieving sparse model selection. Suppose we have data from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the response vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is the $n \times p$ model matrix, $\mathbf{x}_i$'s $(i = 1, 2, \cdots, p)$ are predictor vectors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the coefficient vector and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$ is a vector of i.i.d random variables with mean 0 and variance σ^2. Also assume that the predictors are scaled, meaning that for each column vector, the mean is 0 and the l_2-norm is $\sqrt{n}$. The lasso estimate is defined by

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

where $\|\cdot\|_2$ is the l_2 norm and $\|\cdot\|_1$ is the l_1 norm. The regularization parameter λ continuously shrinks the coefficient estimates towards zeros as λ increases from zero to a large value, resulting in a solution path from the full model to the null model. Both variable selection and coefficient estimation can be achieved efficiently by algorithms such as LARS (Efron, Hastie, Johnstone and Tibshirani, 2004). However, it is now well-known that variable selection by the lasso can be inconsistent (Zou, 2006; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006).

The adaptive lasso proposed by Zou (2006) assigns data-dependent weights to the l_1 penalty components. Suppose $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$ and define a weight vector $\boldsymbol{w} = (w_1, w_2, \cdots, w_p)^T = 1/|\hat{\boldsymbol{\beta}}|^\gamma$, where γ is some positive constant. The adaptive lasso estimate is defined by

$$\hat{\boldsymbol{\beta}}(\text{adaptive lasso}) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{p} w_i|\beta_i|.$$

It has been shown that the adaptive lasso can achieve both consistency in variable selection and asymptotic normality in coefficient estimation.

An important question on the adaptive lasso is how to choose the weight. In practice, the adaptive lasso usually uses $\hat{\boldsymbol{\beta}}(\text{ols})$ as a convenient $\sqrt{n}$-consistent estimate to calculate the weight vector. We denote this weight selection method by OLS-adaptive lasso. OLS-adaptive lasso works well when the initial estimator $\hat{\boldsymbol{\beta}}(\text{ols})$ is reasonably reliable for small or zero coefficients. However, as will be seen later, when the predictors are correlated and the sample size is not large relative to p, OLS can give poor coefficient estimation, making the performance of the adaptive lasso unreliable. Thus, it is desirable to have a weight in the adaptive lasso that is less vulnerable to strong correlation.

In this article, we introduce the standard error adjusted adaptive lasso (SEA-lasso), a new version of the adaptive lasso, which incorporates standard errors of OLS estimate to the weight. To further improve the performance of SEA-lasso, we propose a two-stage model selection method denoted by NSEA-lasso. Numerical results show that SEA-lasso

and NSEA-lasso can perform better than OLS-adaptive lasso under a variety of settings. We also propose an empirical index to help decide whether SEA-lasso and NSEA-lasso should be used in practice. In addition, the consistency and asymptotic normality of SEA-lasso are established.

The rest of the paper is organized as follows. In section 2, we introduce the notation and motivation for SEA-lasso. In section 3, we provide numerical results on model selection performance of SEA-lasso and NSEA-lasso, and explain why they can outperform OLS-adaptive lasso when high correlations exist among the predictors. We give concluding remarks in section 4, and leave all theorems and technical proofs to the Appendix.

2 SEA-lasso

2.1 Definition

Throughout the paper, we assume $p < n$. Unless stated otherwise, p is fixed. Without loss of generality, assume that $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q, \beta_{q+1}, \cdots, \beta_p)^T$ for some $q \leq p$, $\beta_j \neq 0$ for $j = 1, \cdots, q$ and $\beta_j = 0$ for $j = q+1, \cdots, p$. Let $\mathbf{C}_n = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and further assume that $\mathbf{C}_n \to \mathbf{C}$, where both $\mathbf{C}_n$ and $\mathbf{C}$ are non-singular. The matrix $\mathbf{C}$ can be partitioned as

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix},$$

where $\mathbf{C}_{11}$ is a $q \times q$ matrix. Define an empirical value $\kappa = \log \frac{\lambda_{\max}(\mathbf{C}_n)}{\lambda_{\min}(\mathbf{C}_n)}$, where $\lambda_{\max}(\mathbf{C}_n)$ and $\lambda_{\min}(\mathbf{C}_n)$ are the maximum and minimum eigenvalues of $\mathbf{C}_n$, respectively. We call κ the condition index of the model matrix. Note that κ can be large even if the pairwise correlations of the predictors are low.

Let $\hat{\boldsymbol{\beta}}(\text{ols}) = (\hat{\beta}(\text{ols})_1, \hat{\beta}(\text{ols})_2, \cdots, \hat{\beta}(\text{ols})_p)^T$ be the vector of OLS estimate, $\mathbf{s} = (s_1, s_2, \cdots, s_p)$ be the standard error vector of OLS estimate, and $\boldsymbol{w} = (w_1, w_2, \cdots, w_p)^T$ be the standard error adjusted weight vector where $w_i = s_i^\gamma / \hat{\beta}(\text{ols})_i^\gamma$ $(i = 1, 2, \cdots, p)$. For simplicity, we choose $\gamma = 1$ from now on unless stated otherwise. Then define the SEA-lasso estimate by

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_n \sum_{i=1}^{p} w_i |\beta_i|. \tag{1}$$

The motivation for choosing the standard error adjusted weight is from a straightforword intuition. If $\beta_j = 0$, we want to assign a big value to the corresponding weight w_j, which can be achieved if $\hat{\beta}(\text{ols})_j$ is close to zero. But in reality, when p is not small

relative to n and some predictors are highly correlated, a large OLS standard error can result in a poor OLS estimate, and $\hat{\beta}(\text{ols})_j$ can be far away from zero, leading to an under-penalized l_1 term. One possible improvement would be to multiply this term by the OLS standard error. Our simulation shows that when the true model is highly sparse ($q \ll p$), this new weight assignment strategy works well on average compared with OLS-adaptive lasso. When q is close to p, however, SEA-lasso does not necessarily have better model selection performance, possibly due to over-penalization of the nonzero coefficient terms. To achieve better model selection results, we design a two-stage procedure NSEA-lasso. The numerical results show that when the condition index κ is greater than 10, NSEA-lasso generally performs better than OLS-adaptive lasso regardless of the sparsity of the true model. We will also give a heuristic explaination on the improvement of NSEA-lasso over SEA-lasso with a simple numerical example in section 3.2.

2.2 Consistency of SEA-lasso

Zou (2006) has shown that adaptive lasso has an oracle property (Fan and Li, 2001), which includes both consistency in variable selection and asymptotic normality in coefficient estimation. Define that $\mathcal{A} = \{1, \cdots, q\}$ and $\mathcal{A}_n^* = \{j : \hat{\beta}_j^* \neq 0, j = 1, \cdots, p\}$. Then we can show by adapting the proof of Zou (2006) that with a proper choice of λ_n and holding the number of predictors fixed, SEA-lasso maintains the oracle property (See the Appendix). In fact, we will also see in the Appendix that under some additional assumptions, consistency can be extended to situations where the number of predictors grows with the sample size.

2.3 Computation

We obtain the SEA-lasso estimate by Algorithm 1, which is proposed by Zou (2006) for the adaptive lasso. The LARS algorithm (Efron et al, 2004) is used to estimate the coefficients in step 2 of Algorithm 1.

Algorithm 1. (Obtain the solution path by LARS algorithm)

1. Define $\mathbf{x}_i^{**} = \mathbf{x}_i / w_i$, $i = 1, 2, \cdots, p$.

2. Solve the following lasso problem using LARS algorithm and obtain the entire

solution path:

$$\hat{\beta}^{**} = \arg\min_{\beta} \|\mathbf{y} - \sum_{i=1}^{p} \mathbf{x}_i^{**}\beta_i\|_2^2 + \lambda_n \sum_{i=1}^{p} |\beta_i|.$$

3. To compute $\hat{\beta}^*$, divide each element of $\hat{\beta}^{**}$ by the corresponding weight: $\hat{\beta}_i^* = \hat{\beta}^{**}/w_i$, $i = 1, 2, \cdots, p$.

In Algorithm 1, model selection via SEA-lasso is converted to a lasso model selection problem. The next goal is to obtain the optimal λ_n from the solution path. It is well known that if the true sparse model is included in the model candidates, for the purpose of model identification, delete-one cross-validation or fixed proportion k-fold cross validation cannot be consistent (Shao, 1993; Yang, 2007). Instead, BIC type of criteria can be used (Wang, Li and Tsai, 2007). Zou, Hastie and Tibshirani (2007) also suggest using BIC as model selection criterion when the sparsity of the model is the primary concern. They also point out that the optimal λ_n is achieved at one of the transition points in the solution path, which further simplifies the optimization procedure. Thus, in the following Model Selection Procedure 1, we employ BIC to select a model on the solution path.

Model Selection Procedure 1. (Identify the optimal model from the solution path by BIC)
Calculate the BIC values at transition points of the solution path, and identify the optimal model with the minimum BIC value:

$$\mathrm{BIC}(\hat{\mu}_\lambda) = \log\left(\frac{\|\mathbf{y} - \hat{\mu}_\lambda\|_2^2}{n}\right) + \frac{\log n}{n}\hat{df}(\hat{\mu}_\lambda),$$

where $\hat{\mu}_\lambda$ is the model fit at one of the transition points and $\hat{df}(\hat{\mu}_\lambda)$ is the degree of freedom. By Efron's conjecture (Efron et al, 2004; Zou et al, 2007), $\hat{df}(\hat{\mu}_\lambda)$ can be replaced by the number of nonzero estimates immediately before the transition point.

Next, we introduce NSEA-lasso, a two-stage model selection procedure.

Model Selection Procedure 2. (NSEA-lasso)

Stage 1 (weight computation):

1. Use the lasso to obtain a solution path, from which a preliminary model is selected by Model Selection Procedure 1. Let $\hat{\beta}_1 = (\hat{\beta}_{11}, \cdots, \hat{\beta}_{1p})$ be the lasso estimate of this model, and define $\mathcal{A}_1 = \{j : \hat{\beta}_{1j} \neq 0, j = 1, \cdots, p\}$.

2. Sort the elements in $\mathbf{s}$ with descending order, and denote such rearranged OLS-standard error vector by $\mathbf{s}^{(1)} = (s_{t_1}, s_{t_2}, \cdots, s_{t_p})$, where $t_1, t_2, \cdots, t_p$ are the corresponding subscripts in $\mathbf{s}$, and $s_{t_1} \geq s_{t_2} \geq \cdots \geq s_{t_p}$.
3. Based on the preliminary model and the rearranged OLS-standard error obtained in the previous steps, compute the weight vector $\mathbf{w} = (w_1, w_2, \cdots, w_p)$ for the adaptive lasso as follows. For $i = 1$ to p:
   (a) Let $s^{(i)}_{\max}$ and $s^{(i)}_{\min}$ be the maximum (first) and the minimum (last) elements in $\mathbf{s}^{(i)}$, respectively. Then,
   $$w_{t_i} = \begin{cases} s^{(i)}_{\max}/\hat{\beta}(\text{ols})_{t_i}, & \text{if } t_i \notin \mathcal{A}_1, \\ s^{(i)}_{\min}/\hat{\beta}(\text{ols})_{t_i}, & \text{if } t_i \in \mathcal{A}_1. \end{cases}$$
   (b) Delete the first element in $\mathbf{s}^{(i)}$ if $t_i \notin \mathcal{A}_1$, and delete the last element otherwise. Define the remaining elements in $\mathbf{s}^{(i)}$ to be $\mathbf{s}^{(i+1)}$, a $(p-i)$-dimensional vector. Like $\mathbf{s}^{(i)}$, elements in $\mathbf{s}^{(i+1)}$ are in the descending order.

Stage 2 (the adaptive lasso):

Based on the weight vector $\mathbf{w}$ obtained in Stage 1, compute the NSEA-lasso estimate by Algorithm 1.

Both NSEA-lasso and SEA-lasso use OLS-standard error $\mathbf{s}$ to adjust the weight. The difference is that, based on the preliminary model $\mathcal{A}_1$ by the lasso, NSEA-lasso chooses a permutation of elements in $\mathbf{s}$ for the weight computation. Intuitively, given a coefficient β_i and $i \notin \mathcal{A}$, we should choose an element in $\mathbf{s}$ with a relatively large value to compute w_i so as to put more weight on the l_1 penalty term. On the other hand, if $i \in \mathcal{A}$, we should choose an element in $\mathbf{s}$ with a relatively small value. We use this intuition to design the step 3 of Stage 1 in NSEA-lasso procedure. Starting with the coefficient β_{t_1}, which has the biggest OLS-standard error, we find the maximum element $s^{(1)}_{\max}$ in $\mathbf{s}^{(1)}$, and assign the weight w_{t_1} to be $s^{(1)}_{\max}/\hat{\beta}(\text{ols})_{t_1}$ if $t_1 \notin \mathcal{A}_1$. Otherwise, we find the minimum element $s^{(1)}_{\min}$ in $\mathbf{s}^{(1)}$, and assign w_{t_1} to be $s^{(1)}_{\min}/\hat{\beta}(\text{ols})_{t_1}$. Next, we consider β_{t_2}, the coefficient with the second largest OLS-standard error. Since one element in $\mathbf{s}^{(1)}$ has been selected for calculating w_{t_1}, we define a new standard error vector $\mathbf{s}^{(2)}$ by deleting this element from $\mathbf{s}^{(1)}$. Assign the weight w_{t_2} to be $s^{(2)}_{\max}/\hat{\beta}(\text{ols})_{t_2}$ if $t_i \notin \mathcal{A}_1$, and $s^{(2)}_{\min}/\hat{\beta}(\text{ols})_{t_2}$ otherwise. Continue on the process until all the elements in $\mathbf{w}$ are computed. Here, we use $\mathcal{A}_1$ as a practical guess on whether or not a coefficient is zero in the true model. Like OLS-adaptive lasso and SEA-lasso, we can apply Algorithm 1

to obtain the solution path for NSEA-lasso, and apply Model Selection Procedure 1 to find the optimal model. A simple R package for computing SEA-lasso and NSEA-lasso estimates is available upon request.

It is worth pointing out that the oracle property of SEA-lasso still holds when the OLS standard error vector $\mathbf{s}$ for weight computation is replaced by any permutation of its elements. Thus, under the same conditions used for SEA-lasso, NSEA-lasso is also consistent.

3 Numerical Studies

This section gives numerical examples to illustrate the performance of SEA-lasso and NSEA-lasso on model seletion. In the first subsection, we use model matrices from randomly generated covariance structures to quantitatively evaluate the effects of weight selection on the performance of the adaptive lasso. In the second subsection, we use a simple example to explain why NSEA-lasso is proposed as an improvement over SEA-lasso for variable selection. In the third subsection, we propose the use of the condition index κ as a practical indicator on whether SEA-lasso and NSEA-lasso should be considered for model selection. The fourth subsection evaluates the model selection performance with model matrices generated from some special covariance structures. In the last two subsections, we apply SEA-lasso and NSEA-lasso to two real data sets, the diabetes data set and the Boston housing data set.

3.1 Effects of weighting on the performance of the adaptive lasso

We start exploring the effects of weighting on the adaptive lasso by looking at the performance of OLS-adaptive lasso and SEA-lasso. Take $p = 30$, $q = 4$, $n = 200$, $\boldsymbol{\beta} = (2, 1, 3, 2, 0, \cdots, 0)^T$, $\sigma^2 = 0.09$. Then, we generate 250 model matrices by repeating the following procedure: sample a $p \times p$ covariance matrix S from Wishart(p, I_p) where I_p is the p-dimensional identity matrix, generate an $n \times p$ model matrix $\mathbf{X}$ by sampling from $N_p(\mathbf{0}, S)$, and scale $\mathbf{X}$ so that for each column vector, the mean is 0 and l_2 norm is $\sqrt{n}$. Such generated model matrices can give us a variety of designs, and have been used by Zhao and Yu (2006) to illustrate the model selection performance of the lasso.

For each model matrix, generate 100 data sets by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$. Then, compute the solution paths using OLS-adaptive lasso and SEA-lasso. If a solution path contains a model with identical nonzero and zero terms to the true model, we say

the true model is found in the solution path. In addition, we use Model Selection Procedure 1 to find the optimal model from the solution path, and record the value C, the number of zero coefficients correctly estimated as zero, and the value IC, the number of nonezero coefficients incorrectly esimated as zero. After the simulation for the 100 data sets, compute $(\kappa, PCT, \bar{C}, \bar{IC})$, where κ is the condition index for model matrix $\mathbf{X}$, PCT is the percentage of runs that find the true model, $\bar{C}$ is the average of C's and $\bar{IC}$ is the average of IC's.

The simulation results of OLS-adaptive lasso are given in Figure 1. Not surprisingly, since the condition index κ is an indicator of the collinearity problem (Belsley, Kuh and Welsch, 1980), both PCT and $\bar{C}$ drop dramatically as κ increases over 10. Figure 2 shows the comparison of performance between SEA-lasso and OLS-adaptive lasso. Clearly, SEA-lasso improves over OLS-adaptive lasso in terms of both PCT and $\bar{C}$ when κ have large values.

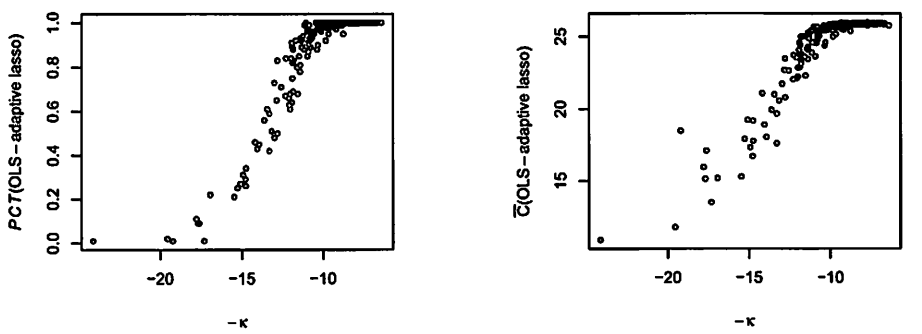

Figure 1: Performance of OLS-adaptive lasso. y-axis: (left panel) PCT(OLS-adaptive lasso); (right panel) $\bar{C}$(OLS-adaptive lasso). x-axis: $-\kappa$.

3.2 Rationalization of SEA-lasso and NSEA-lasso

To understand when and why SEA-lasso is favored over OLS-adaptive lasso, consider the following simple example. Let $p = 30$, $q = 3$, $n = 50$, $\boldsymbol{\beta} = (2, 1, 3, 0, \cdots, 0)^T$ and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. Generate an $n \times p$ model matrix $\mathbf{X}$ by taking a sample of size n from $N_p(\mathbf{0}, \Sigma)$, where Σ is a $p \times p$ covariance matrix. Then scale the model matrix as before. We consider two cases for Σ and σ^2:

Case (a): All the diagonal entries of Σ have value 1, $\Sigma_{4,5} = \Sigma_{5,4} = 0.99$ and all the remaining entries take value 0. Assume $\sigma^2 = 2.25$.

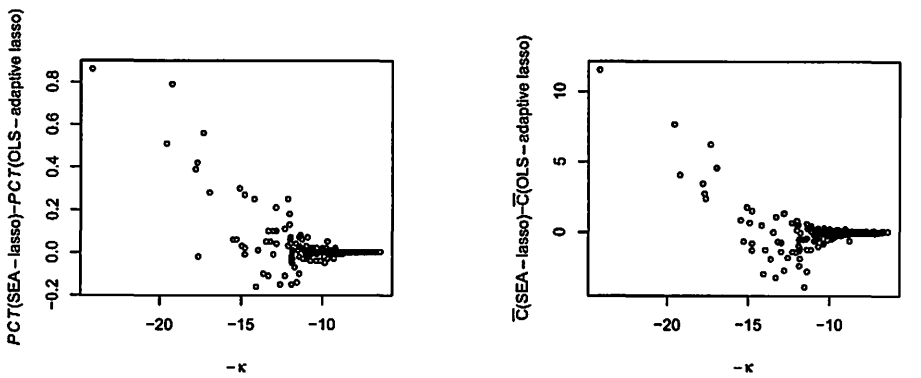

Figure 2: Comparison of PCT (left panel) and $\bar{C}$ (right panel) beween SEA-lasso and OLS-adaptive lasso. x-axis: $-\kappa$.

Case (b): All the diagonal entries of Σ have value 1, $\Sigma_{1,2} = \Sigma_{2,1} = 0.99$ and all the remaining entries take value 0. Assume $\sigma^2 = 0.25$.

For each case, generate 100 simulated response vectors by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, and compute ($\kappa$, PCT, $\bar{C}$, $\bar{IC}$) by OLS-adaptive lasso, SEA-lasso and NSEA-lasso with procedures described in section 3.1. The results are summarized in Table 1 (numbers in the parentheses are the standard errors).

In both of the two cases, only two predictors are highly correlated while the other pairs are independent. For case (a), the large elements in the OLS standard error vector are corresponding to the zero coefficient terms. As a result, SEA-lasso correctly penalizes more on these terms. As expected, SEA-lasso achieves higher PCT than adaptive lasso in case (a). On the other hand, OLS-adaptive lasso performs better than SEA-lasso in case (b). This is because the high correlation occurs between two predictors that correspond to nonzero coefficient terms, and SEA-lasso incorrectly puts more weight on these two terms. NSEA-lasso can avoid the drawback of SEA-lasso by rearranging the OLS standard error vector for the weight computation. Indeed, as shown in Table 1, although SEA-lasso does not give satisfactory PCT in case (b), NSEA-lasso still performs well compared to OLS-adaptive lasso. In the second part of the Appendix, we will show in detail how NSEA-lasso rearranges the OLS standard error vector and improves the weight assignment.

Table 1: Model selection results for simplified models with highly correlated predictors

Case	(a)	(b)
κ	6.5	6.5
PCT		
OLS-adaptive lasso	0.84	0.46
SEA-lasso	0.96	0.16
NSEA-lasso	0.91	0.54
$\bar{C}$		
OLS-adaptive lasso	24.99 (0.32)	25.79 (0.25)
SEA-lasso	25.43 (0.26)	25.10 (0.28)
NSEA-lasso	24.97 (0.30)	25.81 (0.20)
$\bar{IC}$		
OLS-adaptive lasso	0.02 (0.01)	0.58 (0.05)
SEA-lasso	0.03 (0.02)	0.75 (0.07)
NSEA-lasso	0.02 (0.01)	0.53 (0.05)

3.3 NSEA-lasso and the condition index

In the previous subsection, we used a simple example to give a heuristic explanation on why NSEA-lasso can work well even when SEA-lasso cannot. In this subsection, we will see the general applicability of NSEA-lasso under a variety of true model senarios. Consider the following four cases for the true model:

Case 1: $p = 30$, $q = 4$, $n = 200$, $\boldsymbol{\beta} = (2, 1, 3, 2, 0, \cdots, 0)^T$, $\sigma^2 = 0.25$;

Case 2: $p = 30$, $q = 10$, $n = 200$, $\boldsymbol{\beta} = (2, 2, \cdots, 2, 0, \cdots, 0)^T$, $\sigma^2 = 0.25$;

Case 3: $p = 30$, $q = 16$, $n = 200$, $\boldsymbol{\beta} = (2, 2, \cdots, 2, 0, \cdots, 0)^T$, $\sigma^2 = 0.25$;

Case 4: $p = 12$, $q = 4$, $n = 200$, $\boldsymbol{\beta} = (2, 1, 3, 2, 0, \cdots, 0)^T$, $\sigma^2 = 0.09$.

These cases have different sparsity and different number of predictors in the true model, and we want to use them to explore the performance of NSEA-lasso. For each of the true model senarios, we compute $(\kappa, PCT, \bar{C}, \bar{IC})$ the same way as described in section 3.1 except that the solution path is computed by NSEA-lasso. Figure 3 and Figure 4 compare the $\bar{C}$ and $\bar{IC}$ values between NSEA-lasso and OLS-adaptive lasso for the four

cases. In general, when κ is close to 10 or larger, NSEA-lasso performs better in terms of $\bar{C}$. Interestingly, these figures also show that NSEA-lasso can enjoy both higher $\bar{C}$ and lower $\bar{IC}$ than OLS-adaptive lasso, meaning that NSEA-lasso is capable of excluding more zero terms from and including more nonzero terms to the set of selected variables at the same time. Our simulation under many other true model senarios also show the advantage of NSEA-lasso over OLS-adaptive lasso in model selection when κ is large.

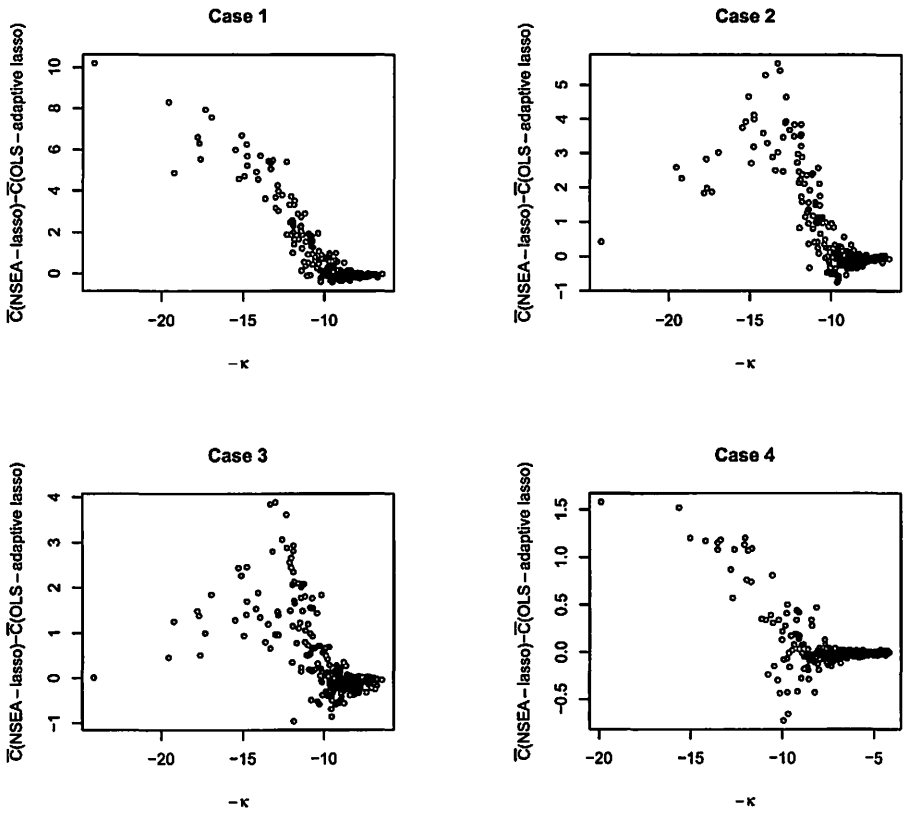

Figure 3: Comparison of $\bar{C}$ beween NSEA-lasso and OLS-adaptive lasso. y-axis: $\bar{C}$(NSEA-lasso) − $\bar{C}$(OLS-adaptive lasso). x-axis: $-\kappa$.

As described in NSEA-lasso procedure, the first stage of NSEA-lasso involves the selection of a preliminary model by the lasso. Due to the unreliability of the OLS weights when κ is large, the lasso turns out to be less vulnerable to collinearity problem than OLS-adaptive lasso. We can repeat the computation of (κ, PCT, $\bar{C}$, $\bar{IC}$) for case 1 using the lasso and OLS-adaptive lasso to obtain the respective solution paths. The difference of PCT values between the lasso and OLS-adaptive lasso is plotted in Figure 5.

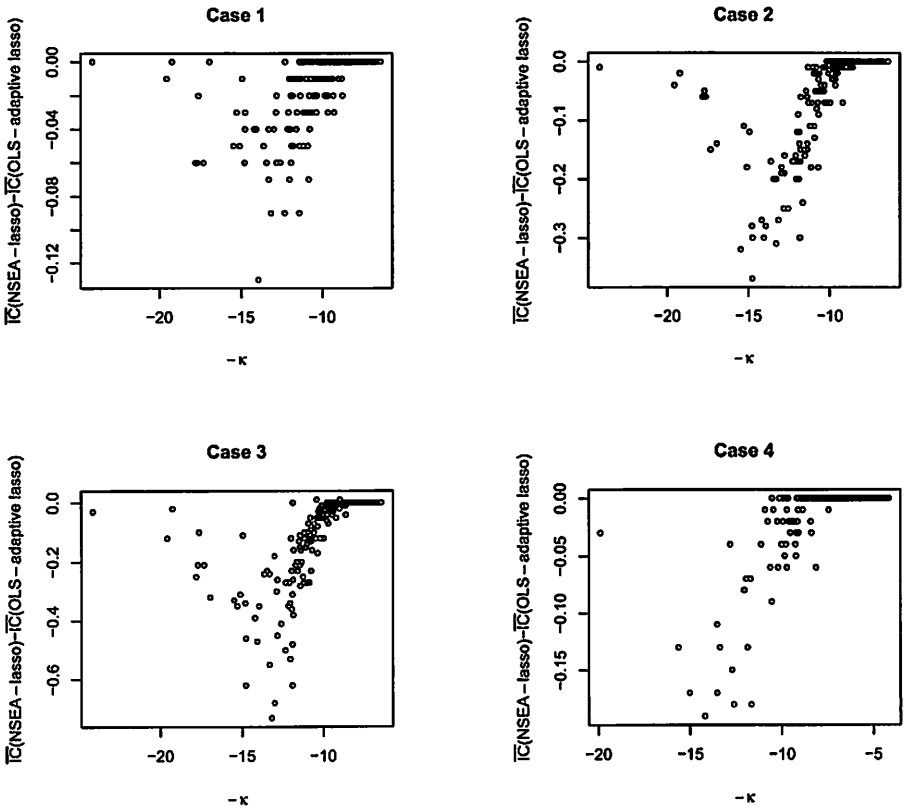

Figure 4: Comparison of $\bar{IC}$ beween NSEA-lasso and OLS-adaptive lasso. y-axis: $\bar{IC}$(NSEA-lasso) − $\bar{IC}$(OLS-adaptive lasso). x-axis: $-\kappa$.

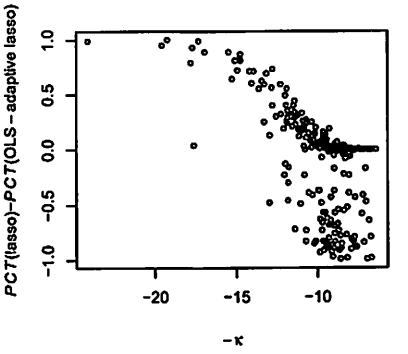

Figure 5: Comparison of PCT between the lasso and OLS-adaptive lasso for case 1. y-axis: PCT(lasso) − PCT(OLS-adaptive lasso). x-axis: $-\kappa$.

It is interesting to note from Figure 5 that OLS-adaptive lasso clearly wins when κ is small, but the lasso outperforms OLS-adaptive lasso when κ is well above 10. Therefore, in cases of high collinearity, the lasso can provide us with a reasonable initial guess on what variables may be included in the true model. To confirm our speculation that the preliminary model selection by the lasso is critical to the success of NSEA-lasso, we slightly modify the first step of the stage 1 in the NSEA-lasso procedure and select a preliminary model by OLS-adaptive lasso instead of the lasso. As expected, such modified NSEA-lasso procedure completely loses its adavantage over OLS-adaptive lasso (the numerical results are not included in this article).

Based on the observations described above, we propose the use of the condition index κ as an empirical guide on how to choose the weight vector for the adaptive lasso. If κ is less than 10, we can simply apply OLS-adaptive lasso for model selection; if κ is close to 10 or even larger than 10, we consider using SEA-lasso or NSEA-lasso to achieve more reliable selection results.

3.4 Models with special covariance structures

In this subsection, we evaluate the performance of SEA-lasso and NSEA-lasso on models with some special predictor covariance structures. Consider this model setting: $p = 100$, $q = 15$, $n = 200$, $\beta = (2, 2, \cdots, 2, 0, \cdots, 0)^T$, $\sigma^2 = 9$, $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. An $n \times p$ model matrix $\mathbf{X}$ is generated by sampling from $N_p(\mathbf{0}, \Sigma)$, where Σ is a $p \times p$ covariance matrix with one of the following structures:

Case 1 (compound symmetry): $\Sigma_{i,i} = 1$, $\Sigma_{i,j} = 0.5$ for $i \neq j$.

Case 2 (power decay): $\Sigma_{i,j} = 0.5^{|i-j|}$.

Case 3 (banded): $\Sigma_{i,i} = 1$, $\Sigma_{i,i+1} = \Sigma_{i+1,i} = 0.5$, all the other entries are 0.

Then we generate 100 simulated responses by $\mathbf{y} = \mathbf{X}\beta + \epsilon$. For each of the three cases, compute (κ, PCT, $\bar{C}$, $\bar{IC}$) by OLS-adapitve lasso, SEA-lasso and NSEA-lasso with procedures described in section 3.1 and 3.2. The results are summarized in Table 2. Interestingly, the banded correlation case has a condition index close to 10, and the performance of SEA-lasso and NSEA-lasso is indeed better than OLS-adaptive lasso. On the other hand, the compound symmetry and power decay correlation have smaller condition indices, and all the three adaptive lasso methods work well. As we will see from the following real data examples, unlike these special covariance structure cases, it is not unusual to encounter data sets with high condition indices, making SEA-lasso and NSEA-lasso attractive alternatives to OLS-adaptive lasso in practice.

Table 2: Model selection results for models with special predictor covariance structures

Case	compound symmetry	power decay	banded
κ	7.0	4.7	9.8
PCT			
OLS-adaptive lasso	0.91	0.99	0.01
SEA-lasso	0.91	0.99	0.01
NSEA-lasso	0.93	0.99	0.03
$\bar{C}$			
OLS-adaptive lasso	83.66 (0.16)	83.68 (0.19)	79.94 (0.44)
SEA-lasso	83.52 (0.17)	83.60 (0.19)	82.83 (0.24)
NSEA-lasso	83.68 (0.15)	83.55 (0.20)	81.51 (0.25)
$\bar{IC}$			
OLS-adaptive lasso	0.01 (0.01)	0.00 (0.00)	4.45 (0.20)
SEA-lasso	0.01 (0.01)	0.01 (0.01)	4.18 (0.18)
NSEA-lasso	0.00 (0.00)	0.00 (0.00)	3.63 (0.20)

3.5 Diabetes data example

In this example, we carry out a simulation study using the diabetes data set from the "Least Angle Regression" paper (Efron et al, 2003). This data set has one response and ten baseline predictors measured on 442 diabetes patients. These baseline predictors include age, sex, body mass index (bmi), average blood pressure (bp) and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu). We use $\mathbf{y}$ to represent the response vector and use $\mathbf{x_1}$, $\mathbf{x_2}$, $\cdots$, $\mathbf{x_{10}}$ to represent the baseline predictor vectors. The baseline model matrix $\mathbf{X_1}$ is comprised of these ten baseline predictor vectors. Besides these baseline predictors, we can also include the quadratic terms and interaction terms to generate an expanded model matrix $\mathbf{X_2}$. This expanded model matrix contains 64 predictors, including 10 baseline predictors, 45 interactions and 9 squares. The square term of $\mathbf{x_2}$ is not included because it is a dichotomous variable. All the column vectors in $\mathbf{X_1}$ and $\mathbf{X_2}$ are scaled as before, and the response $\mathbf{y}$ is centered. The t-tests based on simple linear regression of $\mathbf{y}$ on $\mathbf{X_1}$ show that $\mathbf{x_2}$, $\mathbf{x_3}$, $\mathbf{x_4}$ and $\mathbf{x_9}$ have significant estimated coefficients. Thus, in the simulation we generate the true mean by $\boldsymbol{\mu} = \beta_2\mathbf{x_2} + \beta_3\mathbf{x_3} + \beta_4\mathbf{x_4} + \beta_9\mathbf{x_9}$, where $(\beta_2, \beta_3, \beta_4, \beta_9) = (-6.011, 26.743, 19.468, 23.813)$, the OLS estimate of $\mathbf{y}$ on ($\mathbf{x_2}$, $\mathbf{x_3}$, $\mathbf{x_4}$, $\mathbf{x_9}$). Let $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\mu}$ and generate 100 simulated response vectors $\tilde{\mathbf{y}}$ from $\tilde{\mathbf{y}} = \boldsymbol{\mu} + \tilde{\boldsymbol{\epsilon}}$,

with components of $\tilde{\epsilon}$ a random sample, with replacement, from ϵ. Using either $(\mathbf{X}_1, \tilde{\mathbf{y}})$ or $(\mathbf{X}_2, \tilde{\mathbf{y}})$, compute ($\kappa$, PCT, $\bar{C}$, $\bar{IC}$) by OLS-adaptive lasso, SEA-lasso and NSEA-lasso with procedures described in section 3.1 and 3.2. The results are summarized in Table 3. As expected, for the expanded model matrix ($\kappa > 10$), we can see that SEA-lasso and NSEA-lasso performs better than OLS-adaptive lasso.

Table 3: Simulation results from diabetes data and Boston housing data

	diabetes		Boston housing
model matrix	baseline ($\mathbf{X}_1$)	expanded ($\mathbf{X}_2$)	expanded ($\mathbf{X}_T$)
p	10	64	103
q	4	4	13
κ	6.2	17.2	15.9
PCT			
OLS-adaptive lasso	0.43	0.00	0.00
SEA-lasso	0.74	0.23	0.00
NSEA-lasso	0.67	0.34	0.00
$\bar{C}$			
OLS-adaptive lasso	5.62 (0.06)	54.21 (0.23)	83.67 (0.21)
SEA-lasso	5.89 (0.03)	57.38 (0.30)	83.85 (0.28)
NSEA-lasso	5.78 (0.04)	59.39 (0.09)	87.78 (0.13)
$\bar{IC}$			
OLS-adaptive lasso	0.63 (0.06)	1.39 (0.06)	12.03 (0.06)
SEA-lasso	0.48 (0.06)	0.72 (0.07)	9.43 (0.16)
NSEA-lasso	0.58 (0.06)	0.75 (0.06)	9.26 (0.12)

3.6 Boston housing data example

The Boston housing data set (Harrison and Rubinfeld, 1978) has 506 observations for each census district of the Boston metropolitan area. The response y is median value of owner-occupied homes, and the data set includes 13 baseline predictors $x_1, x_2, \cdots, x_{13}$. Following the transformations proposed by Härdle and Simar (2007), we transform the original data by $y^* = \log(y)$, $x_1^* = \log(x_1)$, $x_2^* = x_2/10$, $x_3^* = \log(x_3)$, $x_4^* = x_4$, $x_5^* = \log(x_5)$, $x_6^* = \log(x_6)$, $x_7^* = x_7^{2.5}/10000$, $x_8^* = \log(x_8)$, $x_9^* = \log(x_9)$, $x_{10}^* = \log(x_{10})$, $x_{11}^* = \exp(0.4 \times x_{11})/1000$, $x_{12}^* = x_{12}/100$ and $x_{13}^* = \sqrt{x_{13}}$.

Like the diabetes example, we consider the quadratic terms and interaction terms, and expand the transformed data set to the model matrix $\mathbf{X}_T$. This expanded model matrix contains 103 predictors, including 13 baseline predictors, 78 interactions and 12 squares. The square term of x_4^* is not included because it is binary. Like before, column vectors of $\mathbf{X}_T$ are scaled, and the transformed response vector $\mathbf{y}$ is centered. In our simulation, the true mean vector $\boldsymbol{\mu}$ is $\boldsymbol{\mu} = \mathbf{X}_T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is obtained by running the lasso on $(\mathbf{X}_T, \mathbf{y})$ for 13 steps. The number of nonzero elements in $\boldsymbol{\beta}$ is also 13. Let $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\mu}$. Then 100 simulated response vectors $\mathbf{y}^*$ are generated from $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}^*$, with components in $\boldsymbol{\epsilon}^*$ a random sample, with replacement, from $\boldsymbol{\epsilon}$. Use $(\mathbf{X}_T, \mathbf{y}^*)$ to compute $(\kappa, PCT, \bar{C}, \bar{IC})$ by OLS-adaptive lasso, SEA-lasso and NSEA-lasso with procedures described in section 3.1 and 3.2. The results summarized in Table 3 shows that NSEA-lasso performs the best among the three adaptive lasso methods in terms of $\bar{C}$ and $\bar{IC}$.

4 Further Extensions

In the previous sections, we have discussed the performance of SEA-lasso and NSEA-lasso under the settings of fixed number of predictors. In many practical problems, we may encounter situations in which $p \to \infty$ as $n \to \infty$. Huang, Ma and Zhang (2008) has shown that under appropriate conditions, the oracle property holds for the adaptive lasso. By employing the technique used by Zou and Zhang (2009), we can specifically show that under some regularity conditions, the consistency of SEA-lasso still holds when $p \to \infty$ but $p < n$ (See the Appendix).

For problems that involves $p > n$, we cannot use OLS estimate for weight computation. One alternative is to use the l_2-penalized estimate and the corresponding standard error. It remains an interesting work to study its asymptotic property under these high-dimensional settings.

5 Appendix

The first part of the Appendix includes the theorems of SEA-lasso and their proofs. The second part is a continued discussion on the example in section 3.2.

5.1 Theorems and Proofs

This part of the Appendix contains three theorems. Theorem 1 shows that when p is fixed, the oracle property holds for SEA-lasso under some regularity conditions. With some additional conditions, Theorem 2 and Theorem 3 extend the oracle property of SEA-lasso to situations where $p \to \infty$ and $p < n$.

Theorem 1. Let $\hat{\boldsymbol{\beta}}^*_{\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{A}}$ be the vectors consisting of the corresponding components in $\hat{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}$, respectively. Suppose that $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$. Then the SEA-lasso estimate satisfies the following properties:

1. Consistency: $\lim_{n\to\infty} P(\mathcal{A}^*_n = \mathcal{A}) = 1$;
2. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{11}^{-1})$.

Since $\lim_{n\to\infty} \sqrt{n} s_j \to$ some constant, $j = 1, \cdots, p$, the proof of Theorem 1 can be immediately obtained by rescaling λ_n in Theorem 2 of Zou (2006). We omit its proof.

Next, we show the oracle property of SEA-lasso with a diverging number of predictors in Theorem 2.

Theorem 2. Assume the following conditions:

(a) Define $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ to be the minimum and maximum eigenvalues of matrix $\mathbf{M}$. Assume

$$0 < b \le \lambda_{\min}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \le \lambda_{\max}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \le B$$

and equivalently,

$$0 < d \le \lambda_{\min}(n(\mathbf{X}^T\mathbf{X})^{-1}) \le \lambda_{\max}(n(\mathbf{X}^T\mathbf{X})^{-1}) \le D;$$

(b) $\lim_{n\to\infty} \frac{\max_{i=1,2,\cdots,n} \sum_{j=1}^p x_{ij}^2}{n} = 0$;

(c) $E|\epsilon|^{2+\delta} < \infty$ for some $\delta > 0$;

(d) $\lim_{n\to\infty} \frac{\log(p)}{\log(n)} = \nu$ for some $0 \le \nu < 1$;

(e)

$$\lim_{n\to\infty} \frac{\lambda_n}{n^{\frac{1+\gamma}{2}}} = 0, \qquad \lim_{n\to\infty} \frac{\lambda_n}{n^{\frac{\nu(1+\gamma)+1}{2}}} = \infty;$$

(f)

$$\lim_{n\to\infty}\left(\frac{n^{\frac{1+\gamma}{2}}}{\sqrt{p}\lambda_n}\right)^{\frac{1}{\gamma}}\left(\min_{j\in\mathcal{A}}|\beta_j^*|\right)=\infty.$$

Then the SEA-lasso estimate $\hat{\boldsymbol{\beta}}^*$ satisfies:

1. Consistency: $\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}) = 1$;
2. Asymptotic normality: $\alpha^T\Sigma_{\mathcal{A}}^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^* - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} \mathrm{N}(0,\sigma^2)$,

where $\Sigma_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}$ and α is a vector of norm 1.

To prove Theorem 2, we need to first show Theorem 3.

Theorem 3. Let $w_j = \frac{s_j^\gamma}{\hat{\beta}(\mathrm{ols})_j^\gamma}$ and write $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0})$. Define

$$\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^* = \arg\min_{\boldsymbol{\beta}}\|\mathbf{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}\|_2^2 + \lambda_n\sum_{j\in\mathcal{A}} w_j|\beta_j|.$$

Then, under conditions (a), (d), (e) and (f),

$$\lim_{n\to\infty} P((\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*, \mathbf{0}) = \hat{\boldsymbol{\beta}}^*) = 1.$$

Proof of Theorem 3. We need to show that $(\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*, \mathbf{0})$ satisfies the KKT conditions of (1) with probability tending to 1. It suffices to show

$$P(\exists j \in \mathcal{A}^c,\ |-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j) \to 0.$$

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

$$\begin{aligned}&P(\exists j \in \mathcal{A}^c,\ |-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j)\\ \le &\sum_{j\in\mathcal{A}^c} P(|-2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j,\ \hat{\eta} > \eta/2) + P(\hat{\eta} \le \eta/2)\end{aligned}$$

and

$$P(\hat{\eta} \le \eta/2) \le P(\|\hat{\boldsymbol{\beta}}(\mathrm{ols}) - \boldsymbol{\beta}^*\|_2 \ge \eta/2) \le \frac{E(\|\hat{\boldsymbol{\beta}}(\mathrm{ols}) - \boldsymbol{\beta}^*\|_2^2)}{\eta^2/4}.$$

Then, by Theorem 3.1 of Zou and Zhang (2009),

$$P(\hat{\eta} \le \eta/2) \le 16\frac{Bpn\sigma^2}{b^2n^2\eta^2}. \tag{2}$$

Moreover, let $M = (\frac{\lambda_n}{n^{\frac{\gamma+2}{2}}})^{\frac{1}{1+\gamma}}$, and we have

$$
\begin{aligned}
&\sum_{j\in\mathcal{A}^c} P(|-2\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j,\, \hat{\eta} > \eta/2) \\
\leq &\sum_{j\in\mathcal{A}^c} P(|-2\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j,\, \hat{\eta} > \eta/2,\, |\hat{\beta}(\text{ols})| \leq M) + \sum_{j\in\mathcal{A}^c} P(|\hat{\beta}(\text{ols})| > M) \\
\leq &\sum_{j\in\mathcal{A}^c} P(|-2\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \frac{\lambda_n d^{\frac{\gamma}{2}}}{n^{\frac{\gamma}{2}}M^{\gamma}},\, \hat{\eta} > \eta/2) + \sum_{j\in\mathcal{A}^c} P(|\hat{\beta}(\text{ols})| > M) \\
\leq &\frac{4M^{2\gamma}n^{\gamma}}{\lambda_n^2 d^{\gamma}} E(\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)|^2 I(\hat{\eta} > \eta/2)) + \frac{1}{M^2} E(\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}(\text{ols})\|^2) \\
\leq &\frac{4M^{2\gamma}n^{\gamma}}{\lambda_n^2 d^{\gamma}} E(\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)|^2 I(\hat{\eta} > \eta/2)) + \frac{4Bpn\sigma^2}{b^2n^2M^2}.
\end{aligned} \tag{3}
$$

By the model assumption, we have

$$
\begin{aligned}
&\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)|^2 \\
=&\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T(\mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}^* - \mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*) + \mathbf{x}_j^T\boldsymbol{\epsilon}|^2 \\
\leq &2BnBn\|\boldsymbol{\beta}_{\mathcal{A}}^* - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*\|_2^2 + 2\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T\boldsymbol{\epsilon}|^2,
\end{aligned}
$$

which gives us the inequality

$$
\begin{aligned}
&E(\sum_{j\in\mathcal{A}^c} |\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)|^2 I(\hat{\eta} > \eta/2)) \\
\leq &2B^2n^2E(\|\boldsymbol{\beta}_{\mathcal{A}}^* - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*\|_2^2 I(\hat{\eta} > \eta/2)) + 2Bnp\sigma^2.
\end{aligned} \tag{4}
$$

Define $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols}) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}\|_2^2$. Then by following the argument in the proof of Theorem 3.1 of Zou and Zhang (2009), we obtain that

$$
\begin{aligned}
&E(\|\boldsymbol{\beta}_{\mathcal{A}}^* - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*\|_2^2 I(\hat{\eta} > \eta/2)) \\
\leq &2E(\|\boldsymbol{\beta}_{\mathcal{A}}^* - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols})\|_2^2) + 2E(\|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols}) - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*\|_2^2 I(\hat{\eta} > \eta/2)) \\
\leq &\frac{4|\mathcal{A}|\lambda_{\max}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})\sigma^2 + 2\lambda_n^2 E(\sum_{j\in\mathcal{A}} w_j^2 I(\hat{\eta} > \eta/2))}{b^2n^2} \\
\leq &4\frac{nBp\sigma^2 + \frac{\lambda_n^2 D^{\gamma}p}{n^{\gamma}}(\frac{\eta}{2})^{-2\gamma}}{b^2n^2}.
\end{aligned} \tag{5}
$$

Combining (2), (3), (4) and (5), we obtain that

$$\begin{aligned}
&\sum_{j\in\mathcal{A}^c} P(|-2\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}_{\mathcal{A}}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*)| > \lambda_n w_j,\ \hat{\eta} > \eta/2) \\
&\le \frac{4M^{2\gamma}n^{\gamma}}{\lambda_n^2 d^{\gamma}}\Big(8B^2n^2\frac{nBp\sigma^2+\frac{\lambda_n^2 D^{\gamma}p(\eta/2)^{-2\gamma}}{n^{\gamma}}}{b^2n^2}+2Bnp\sigma^2\Big)+\frac{4Bpn\sigma^2}{b^2n^2M^2}+\frac{16Bpn\sigma^2}{b^2n^2\eta^2} \\
&\triangleq K_1+K_2+K_3.
\end{aligned}$$

It follows that

$$K_1 = O\Big(\frac{M^{2\gamma}n^{\gamma+1}p}{\lambda_n^2}\Big) = O\Big(\Big(\frac{\lambda_n}{n^{\frac{1+\gamma}{2}}}n^{((1-\nu)(1+\gamma)-1)/2}\Big)^{-\frac{2}{1+\gamma}}\Big) \to 0,$$

$$K_2 = O\Big(\frac{p}{n}\Big(\frac{\lambda_n}{n^{\frac{\gamma+2}{2}}}\Big)^{-\frac{2}{1+\gamma}}\Big) \to 0,$$

and

$$K_3 = O\Big(\frac{p}{n\eta^2}\Big) \to 0. \tag{6}$$

Here, $K_3 \to 0$ holds since

$$\frac{p}{n\eta^2} = \frac{1}{\eta^2\Big(\frac{n^{(1+\gamma)/2}}{\sqrt{p}\lambda_n}\Big)^{2/\gamma}}\Big(\frac{p}{n}\Big(\frac{n^{(\gamma+2)/2}}{\lambda_n}\Big)^{\frac{2}{1+\gamma}}\Big)^{\frac{1+\gamma}{\gamma}}\frac{1}{p^{2/\gamma}}.$$

Thus, the proof of theorem 3 is complete. □

Proof of Theorem 2. From theorem 3, we already show that SEA-lasso estimate is equal to $(\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^*, \mathbf{0})$ with probability tending to 1. To show the consistency, it suffices to prove

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

By the fact that $\sum_{j\in\mathcal{A}} w_j^2 \le \frac{D^{\gamma}p}{n^{\gamma}\hat{\eta}^{2\gamma}}$, we can follow the proof of Theorem 3.1 of Zou and Zhang (2009) to obtain that

$$\min_{j\in\mathcal{A}}|\tilde{\beta}_j^*| > \min_{j\in\mathcal{A}}|\tilde{\beta}(\text{ols})_j| - \frac{\lambda_n D^{\gamma/2}\sqrt{p}}{n^{\gamma/2}bn\hat{\eta}^{\gamma}}.$$

Note that

$$\min_{j\in\mathcal{A}}|\tilde{\beta}(\text{ols})_j| > \min_{j\in\mathcal{A}}|\beta_j^*| - \|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols}) - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2.$$

Following (5), we can show that

$$E\|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols}) - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2^2 \le \frac{4nBp\sigma^2}{b^2n^2} = O\Big(\frac{p}{n}\Big).$$

Thus, $\|\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\text{ols}) - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2 = \sqrt{\frac{p}{n}}O_p(1)$.
Moreover,

$$\frac{\lambda_n D^{\gamma/2}\sqrt{p}}{n^{\gamma/2}bn\hat{\eta}^{\gamma}} = \frac{D^{\gamma/2}}{\sqrt{n}}\Big(\frac{n^{\frac{\gamma+1}{2}}\eta^{\gamma}}{\sqrt{p}\lambda_n}\Big)^{-1}\Big(\frac{\eta}{\hat{\eta}}\Big).$$

From (2), it is easy to see that

$$P\Big(\frac{\eta}{\hat{\eta}} \geq 2\Big) \leq O\Big(\frac{p}{n\eta^2}\Big).$$

Since we already show $\frac{p}{n\eta^2} \to 0$, $\left(\frac{\eta}{\hat{\eta}}\right)^{\gamma} = O_p(1)$.
Therefore,

$$\frac{\lambda_n D^{\gamma/2}\sqrt{p}}{n^{\gamma/2}bn\hat{\eta}^{\gamma}} = o\Big(\frac{1}{\sqrt{n}}\Big)O_p(1). \tag{7}$$

Hence, we have

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

and

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

To prove asymptotic normality, we only need to follow the steps in the second half of the proof for Theorem 3.3 of Zou and Zhang (2009). Thus, we omit its proof here. □

5.2 A continued discussion on the example in section 3.2

Table 4: Solution paths of one repetition for case (a), first 10 steps

step	lasso	OLS-adaptive lasso	SEA-lasso	NSEA-lasso
1	3	3	1	3
2	1	1	3	1
3	2	4	2	2
4	22	2	21	21
5	6	21	7	7
6	13	12	14	14
7	21	7	12	12
8	14	14	22	22
9	7	22	8	8
10	12	8	29	29

Table 5: OLS-estimate and OLS-standard error vectors of one repetition for case (a)

predictor	$\hat{\beta}$(ols)	OLS-std.err.	rearranged std.err. by NSEA-lasso
1	2.7937	0.3635	0.2960
2	0.6427	0.3068	0.3008
3	2.4531	0.4065	0.2848
4	**-2.6876**	**2.1762**	**2.1762**
5	**2.7336**	**2.1647**	**2.1647**
6	-0.2030	0.3569	0.3711
7	-0.5310	0.3527	0.3569
8	0.6802	0.3443	0.3527
9	0.1012	0.3138	0.3279
10	-0.4104	0.5396	0.5396
11	0.3970	0.3522	0.3553
12	-0.8288	0.4143	0.4143
13	-0.0089	0.3874	0.3899
14	0.3634	0.3255	0.3337
15	0.2614	0.2848	0.3040
16	0.2991	0.3060	0.3255
17	0.0927	0.3899	0.4050
18	0.1112	0.2960	0.3060
19	0.1818	0.3418	0.3443
20	0.1270	0.3337	0.3418
21	-0.5628	0.3040	0.3138
22	-0.4245	0.4174	0.4174
23	0.2021	0.3553	0.3635
24	0.2241	0.3360	0.3426
25	0.4055	0.3279	0.3360
26	0.3275	0.3711	0.3874
27	-0.2947	0.4050	0.4065
28	0.3367	0.3008	0.3068
29	0.5872	0.3426	0.3522
30	0.2160	0.4315	0.4315

This part of the Appendix is a continued discussion on the example shown in section 3.2. We have observed that compared with OLS-adaptive lasso, SEA-lasso performs well in case (a), but not in case (b). NSEA-lasso, on the other hand, delivers good results in both cases. In the effort to further understand these results, we examine one of the 100 repetitions in detail by looking at the solution path, OLS estimate and OLS standard

Table 6: Solution paths of one repetition for case (b), first 10 steps

step	lasso	OLS-adaptive lasso	SEA-lasso	NSEA-lasso
1	1	3	3	1
2	3	1	1	3
3	2	2	28	2
4	28	28	18	28
5	7	18	29	18
6	20	29	25	29
7	18	25	7	7
8	4	7	23	25
9	29	23	11	23
10	24	26	10	20

error.

For case (a), the solution paths by the lasso, OLS-adaptive lasso, SEA-lasso and NSEA-lasso from one repetition are listed in Table 4 (only show 10 steps). Both of the solution paths by SEA-lasso and NSEA-lasso include the true model while OLS-adaptive lasso does not. This result can be understood by looking at the weight vectors. Table 5 lists the OLS estimate $\hat{\beta}(\text{ols})$ as well as the OLS-standard error vector used by SEA-lasso and the rearranged standard error vector used by NSEA-lasso. Due to the high correlation between predictor 4 and predictor 5, $\hat{\beta}(\text{ols})_4$ and $\hat{\beta}(\text{ols})_5$ have much larger absolute values than that of other non-important terms, leading to under-penalization in OLS-adaptive lasso. Therefore, it is helpful to adjust the weight and use the standard error vectors to put more weight on l_1 penalty terms of predictor 4 and predictor 5.

Different from case (a), high correlation occurs between predictor 1 and predictor 2 in case (b). Both of the two predictors are corresponding to nonzero terms. The solution paths from one repetition are listed in Table 6 (only show 10 steps). The solution path of SEA-lasso misses the true model while all the other three paths contain the true model. Therefore, SEA-lasso can perform poorer than OLS-adaptive lasso in this case. This is not a surprising result if we look at the OLS-standard error vectors shown in Table 7. Due to high correlation, the first two elements in OLS-standard error vector have large values. Applying this standard error vector directly for weight calculation in SEA-lasso results in over-penalization of the nonzero terms. Interestingly, the lasso correctly includes the first two terms into the preliminary model, and advises NSEA-lasso to rearrange the OLS-standard error vector. As shown in Table 7, the first two elements

Table 7: OLS-estimate and OLS-standard error vectors of one repetition for case (b)

predictor	$\hat{\beta}$(ols)	OLS-std.err.	rearranged std.err. by NSEA-lasso
1	2.4292	**0.7216**	0.0987
2	0.6795	**0.7254**	0.0949
3	2.9439	0.1023	0.1085
4	0.0032	0.1355	0.1003
5	0.1135	0.1212	0.1300
6	-0.1129	0.1190	0.1291
7	0.0778	0.1176	0.1013
8	0.0145	0.1148	0.1190
9	0.0097	0.1046	0.1148
10	-0.1436	0.1799	**0.7254**
11	0.0887	0.1174	0.1212
12	-0.1083	0.1381	0.1438
13	-0.0473	0.1291	0.1355
14	0.0455	0.1085	0.1174
15	0.0508	0.0949	0.1120
16	-0.0050	0.1020	0.1142
17	0.0676	0.1300	0.1381
18	-0.1623	0.0987	0.1112
19	0.0741	0.1139	0.1184
20	-0.0646	0.1112	0.1046
21	-0.0082	0.1013	0.1139
22	-0.0258	0.1391	0.1799
23	0.1278	0.1184	0.1237
24	-0.0224	0.1120	0.1023
25	0.1348	0.1093	0.1176
26	0.1409	0.1237	0.1350
27	-0.0254	0.1350	0.1391
28	0.1402	0.1003	0.1093
29	0.1513	0.1142	0.1020
30	0.0548	0.1438	**0.7216**

in the rearranged standard error vector no longer have the largest values; instead, the largestest values are now in positions 10 and 30, both of which are corresponding to zero terms. This observation explains why NSEA-lasso performs well in this case even though SEA-lasso does not.

References

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*, John Wiley and Sons, New York.

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

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

Harrison, D. and Rubinfeld, D. L. (1978), Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81-102.

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

Härdle, W. and Simar, L. (2007). *Applied Multivariate Statistical Analysis, 2nd ed.*, Springer, New York.

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

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

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

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Annals of Statistics*, 35, 2450-2473.

Zhao, P. and Yu B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7, 2541-2567.

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

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35, 2173-2192.

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