Regularization after retention in ultrahigh dimensional linear regression models *

Haolei Weng, Yang Feng and Xingye Qiao

Abstract

Lasso has been both theoretically and empirically proved a successful variable selection approach. However, in the ultrahigh dimensional setting, the conditions of model selection consistency for lasso could easily fail. The independence screening framework tackles this problem by reducing the dimensionality based on marginal correlations before performing lasso. In this paper, we propose a two-step approach to relax the irrepresentable-type condition (Wainwright, 2009) of lasso by using marginal information in a different perspective from independence screening. In particular, we retain significant variables rather than screening out irrelevant ones. The new method is shown to be model selection consistent in the ultrahigh dimensional linear regression model. To improve the finite sample performance, we then introduce a three-step version and characterize its asymptotic behavior. Simulations and real data analysis show advantages of our method over lasso and independence screening in certain regimes.

Keywords: Independence screening; Lasso; Penalized least square; Retention; Selection consistency; Variable selection.

*Haolei Weng is Ph.D. Student, Department of Statistics, Columbia University, New York, NY 10027 (Email: hw2375@columbia.edu). Yang Feng is Assistant Professor, Department of Statistics, Columbia University, New York, NY 10027 (Email: yangfeng@stat.columbia.edu). Xingye Qiao is Assistant Professor, Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY 13902 (Email: qiao@math.binghamton.edu).
1 Introduction

High dimensional statistical learning has become increasingly important in many scientific areas. It mainly deals with statistical estimation and prediction in the setting where the dimensionality $p$ is substantially larger than the sample size $n$. An active philosophy of research imposes sparsity constraints on the model so as to achieve estimation consistency. Under this framework, variable selection plays a crucial role in three aspects: statistical accuracy, model interpretability and computational complexity.

Various penalized maximum likelihood methods have been proposed in recent years. Compared to traditional variable selection methods such as Akaike’s information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz, 1978), these regularization techniques essentially shrink the parameter space to reduce statistical variation and computational cost. Examples include bridge regression (Frank and Friedman, 1993), the least absolute shrinkage and selection operator (Tibshirani, 1996), the smoothly clipped absolute deviation (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the minimax concave penalty (Zhang, 2010), among others. Theoretical results on parameter estimation (Knight and Fu, 2000), model selection (Zhao and Yu, 2006; Wainwright, 2009), prediction (Greenshtein and Ritov, 2004) and oracle properties (Fan and Li, 2001) have been developed under different model contexts. However, in the ultrahigh dimensional feature space, where $\log p = O(n^\xi)$ ($\xi > 0$), the conditions for model selection/parameter estimation consistency associated with these techniques may easily fail due to high correlations between important and unimportant variables. Motivated by these concerns, Fan and Lv (2008) proposed a sure independence screening method in the linear regression setting. It has been further developed by Fan and Song (2010) and Fan et al. (2011) for more general models. The main idea
of independence screening methods is to utilize marginal information to screen out irrelevant variables. Fast computation and desirable statistical properties make them more attractive in large scale problems. After independence screening, other variable selection methods can be further applied to improve finite sample performance.

In this paper, we consider variable selection consistency in the ultrahigh dimensional linear regression model. A two-step approach is proposed, in a different direction from independence screening methods, in terms of how the marginal information is used. In the first step, we count on marginal regression coefficient estimates to retain a set of important predictors. In the second step, we impose $\ell_1$ penalty on the remaining variables using penalized least square. In theoretical development, we replace the assumption on the lower bound of marginal information for important variables (Fan and Lv, 2008) by an assumption on the upper bound of marginal information for irrelevant features. As a result, we can relax the irrepresentable-type condition of lasso (Wainwright, 2009) when $\log p = O(n^\xi)$ ($\xi > 0$). From the practical point of view, a permutation-based method is introduced to choose the threshold in the retention step. To enhance finite sample performance, we also introduce a three-step version to eliminate unimportant variables falsely selected during the retention step.
2 Model setup and relevant variable selection techniques

2.1 Model setup and notations

Let $V_1, \ldots, V_n$ be independently and identically distributed random vectors, where $V_i = (X_i, Y_i)$, following the linear regression model,

$$Y_i = X_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i = (X_i^1, \ldots, X_i^p)^T$ is $p$-dimensional Gaussian distributed as $N(0, \Sigma)$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the true coefficient vector, $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed as $N(0, \sigma^2)$, and $\{X_i\}_{i=1}^n$ are independent of $\{\varepsilon_i\}_{i=1}^n$. Denote the support index set of $\beta$ by $S = \{j : \beta_j \neq 0\}$ and the cardinality of $S$ by $s$. For any set $A$, let $A^c$ be its complement set. For any $k$ dimensional vector $w$ and any subset $K \subseteq \{1, \ldots, k\}$, $w_K$ denotes the subvector of $w$ indexed by $K$, and let $\|w\|_1 = \sum_{i=1}^k |w_i|$, $\|w\|_2 = (\sum_{i=1}^k w_i^2)^{1/2}$, $\|w\|_\infty = \max_{i=1,\ldots,k} |w_i|$. For any $k_1 \times k_2$ matrix $M$, any subsets $K_1 \subseteq \{1, \ldots, k_1\}$, $K_2 \subseteq \{1, \ldots, k_2\}$, $M_{K_1,K_2}$ represents the submatrix of $M$ consisting of entries indexed by the Cartesian product $K_1 \times K_2$. Let $M_{K_2}$ be the columns of $M$ indexed by $K_2$ and $M^j$ be the $j$-th column of $M$. Denote $\|M\|_2 = \{\Lambda_{\max}(M^T M)\}^{1/2}$, $\|M\|_\infty = \max_{i=1,\ldots,k} \sum_{j=1}^k |M_{ij}|$. When $k_1 = k_2 = k$, let $\rho(M) = \max_{i=1,\ldots,k} M_{ii}$, $\Lambda_{\min}(M)$, $\Lambda_{\max}(M)$ be the minimum and maximum eigenvalues of $M$ respectively, and $\Sigma_{S \setminus S^c} = \Sigma_{S^c} \Sigma_{S^c}^{-1} \Sigma_{SS^c}$.

In the ultrahigh dimensional scenario, assuming $\beta$ is sparse, we are interested in recovering the sparsity pattern $S$ of $\beta$. For technical convenience, we consider a stronger result called sign consistency (Zhao and Yu, 2006), namely $\text{pr}(|\text{sign}(\hat{\beta}) = \text{sign}(\beta)| \rightarrow 1$, as $n \rightarrow \infty$, where $\text{sign}(\cdot)$ maps positive numbers to 1, negative numbers
to $-1$ and zero to zero. In asymptotic analysis, we denote the sparsity level by $s_n$ and dimension by $p_n$ to allow them to grow with the number of observations. For conciseness, we sometimes use signals and noises to represent relevant predictors $S$ and irrelevant predictors $S^c$ or their corresponding coefficients, respectively.

### 2.2 Lasso in random design

The least absolute shrinkage and selection operator (Tibshirani, 1996) solves

$$
\hat{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\}.
$$

For fixed design, model selection consistency has been well studied in Zhao and Yu (2006) and Wainwright (2009). They characterized the dependency between relevant and irrelevant predictors by an irrepresentable condition, which proved to be both sufficient and (almost) necessary for sign consistency. For random design, Wainwright (2009) established precise sufficient and necessary conditions on $(n, p_n, s_n)$ for sparse recovery. We state a corollary from his general results with a particular scaling of the triplet for further use in the sequel. Here are some key conditions.

**Condition 0.** $\log p_n = O(n^{a_1}), s_n = O(n^{a_2}), a_1 > 0, a_2 > 0, a_1 + 2a_2 < 1.$

**Condition 1.** $\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0.$

**Condition 2.** $\|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_{\infty} \leq 1 - \gamma, \ \gamma \in (0, 1].$

**Condition 3.** $\rho(\Sigma_{S^c|S}) = o(n^{\delta}), \ 0 < \delta < 1 - a_1 - 2a_2.$

**Condition 4.** $\min_{j \in S} |\beta_j| \geq Cn^{(\delta + a_1 + 2a_2 - 1)/2}$ for a sufficient large $C$, where $\delta$ is the same as in Condition 3.
Condition 2 is the population analog of the irrepresentable condition in Zhao and Yu (2006), in which \((\Sigma_{SS})^{-1}\Sigma_{SSc}\) is the regression coefficient matrix by regressing noises on signals. Hence, \(\|\Sigma_{SSc}(\Sigma_{SS})^{-1}\|_{\infty}\) can be viewed as a reasonable measurement of the dependency between signals and noises. In the ultrahigh dimensional scenario, noises are likely to be highly correlated with signals, which could make this condition fail. To relax this condition, the corresponding matrix in the regularization step for our method is a submatrix of \(\Sigma_{SSc}(\Sigma_{SS})^{-1}\) with smaller number of columns. As a result, the corresponding quantity in Condition 2 is reduced. In Condition 3, \(\Sigma_{S\cdot|S}\) is the conditional covariance matrix of \(X_{Sc}\) given \(X_S\). This condition imposes another kind of eigenvalue-type dependency constraint. In addition to the dependency conditions between signals and noises, the signals should be linearly independent and the minimum signal can not decay too fast as shown by Conditions 1 and 4, respectively.

**Proposition 1.** Under the scaling specified in Condition 0, if the covariance matrix \(\Sigma\) and the true parameter \(\beta\) satisfy Conditions 1–4, and \(p_n \to \infty, p_n - s_n \to \infty, \lambda_n \asymp n^{(\delta + a_1 - 1)/2}\), we have sign consistency

\[
\Pr(\hat{\beta} \text{ is unique, and } \text{sign}(\hat{\beta}) = \text{sign}(\beta)) \to 1 \quad \text{as } n \to \infty.
\]

### 2.3 Independence screening

Sure independence screening was proposed by Fan and Lv (2008) in the linear regression model framework. It conducts variable selection according to magnitude of marginal correlations. Specifically, assume that the columns in the design matrix \(X = (X^1, \ldots, X^{p_n})\) have been standardized with mean zero and variance one. Denote the response vector \(Y = (Y_1, \ldots, Y_n)^T\) and the rescaled sample correlation between each predictor \(X^j\) and \(Y\) by \(\hat{\beta}_j^M = Y^T X^j\) \((1 \leq j \leq p_n)\). Then the selected submodel...
by sure independence screening is

\[ \hat{M}_{dn} = \{ 1 \leq j \leq p_n : |\hat{\beta}_j^M| \text{ belongs to the } d_n \text{ largest values} \} , \]

where \( d_n \) is a positive integer smaller than \( n \). This simple procedure turns out to enjoy the sure screening property as reviewed in the following. Consider \( \log p_n = O(n^a), a \in (0, 1 - 2\kappa) \), where \( 0 < \kappa < 1/2 \). Under the conditions

\[
\begin{align*}
\text{var}(Y_i) &= O(1), \quad \Lambda_{\max}(\Sigma) = O(n^\tau), \quad \min_{j \in S} |\beta_j| \geq cn^{-\kappa}, \quad \tau \geq 0, \\
\min_{j \in S} |\text{cov}(\beta_j^{-1}Y_1, X_j^1)| &\geq c > 0, \\
\end{align*}
\]

(1)

Fan and Lv (2008) showed that if \( 2\kappa + \tau < 1 \), then there exists some \( \theta \in (2\kappa + \tau, 1) \) such that for \( d_n \approx n^{\theta} \), we have for some \( C > 0 \),

\[
\text{pr}(S \subseteq \hat{M}_{dn}) = 1 - O(p_n \exp(-Cn^{1-2\kappa}/\log n))
\]

The condition in (1) imposes a lower bound for magnitudes of the marginal correlations between response and signals. In some cases, signals are marginally uncorrelated with the response, then this condition is not satisfied. Although Fan and Lv (2008) introduced an iterative version to overcome this issue, the associated theoretical property is still unknown. We will drop this assumption and focus on the situation where the marginal correlations between noises and the response are not large.
3 Method and theory

3.1 The new two-step estimator

In this section, we propose a two-step method named regularization after retention (RAR). In the first step, we use marginal information to retain important signals, and in the second step, we conduct a penalized least square with penalty only on the variables not retained in the first step.

*Step 1. (Retention)* Calculate the marginal regression coefficient estimate for each predictor,

\[
\hat{\beta}_j^M = \frac{\sum_{i=1}^n (X^j_i - \bar{X}^j)Y_i}{\sum_{i=1}^n (X^j_i - \bar{X}^j)^2} \quad (1 \leq j \leq p),
\]

where \( \bar{X}^j = n^{-1} \sum_{i=1}^n X^j_i \). Then define a retention set by \( \hat{R} = \{1 \leq j \leq p : |\hat{\beta}_j^M| \geq \gamma_n \} \), for a positive constant \( \gamma_n \).

*Step 2. (Regularization)* The final estimator is

\[
\hat{\beta} = \arg \min_\beta \left\{ \frac{2}{n} \sum_{i=1}^n (Y_i - X^T_i \beta)^2 + \lambda_n \sum_{j \in \hat{R}^c} |\beta_j| \right\}.
\]

The retention step is very similar to independence screening. Technically, they are both based on marginal utility to infer about the joint information. Conceptually, independence screening aims at screening out as many noises as possible, while our method tries to detect and retain as many signals as possible. The threshold \( \gamma_n \) needs to be chosen carefully so that no noise is retained. In the desired situation when \( \hat{R} \subseteq S \), meaning all the variables in \( \hat{R} \) are signals, one only needs to impose sparsity on \( \hat{R}^c \) to recover the entire sparsity pattern. The advantage is that the estimation accuracy of \( \beta_{\hat{R}} \) is not compromised due to regularization.
Moreover, it turns out that this well-learned information can relax the consistency conditions of lasso. On the other hand, we need extra regularity conditions to guarantee $\hat{R} \subseteq S$ with high probability. We will show that under the scaling $\log p_n = O(n^\xi)$ ($\xi > 0$), our estimator $\hat{\beta}$ achieves sign consistency. The two steps will be studied separately in Section 3.2 and Section 3.3.

### 3.2 Asymptotics in the retention step

Let the marginal regression coefficients $\beta_j^M = \text{cov}(X_j^1, Y_1)$. For simplicity, we assume the covariance matrix $\Sigma$ for $X_1$ has unit diagonal elements and the variance of random error is $\sigma^2 = 1$. We first present several conditions.

**Condition 5.** $\|\Sigma \beta\|_\infty = O(n^{(1-2\kappa)/8})$, where $0 < \kappa < \frac{1}{2}$ is a constant.

**Condition 6.** $\beta^T \Sigma S \beta_s = O(1)$.

**Proposition 2.** Under Conditions 5 and 6, we have for any $c_\ast > 0$, there exists $c_2 > 0$,

$$\text{pr}(\max_{1 \leq j \leq p_n} |\hat{\beta}_j^M - \beta_j^M| > c_\ast n^{-\kappa}) = O(p_n \exp(-c_2 n^{(1-2\kappa)/4})).$$

(2)

**Corollary 1.** Let $\zeta_n = \|\Sigma S \beta_s\|_\infty$ and $c_1$ be a positive constant. Under Conditions 5-6, and when the threshold $\gamma_n = \zeta_n + c_1 n^{-\kappa}$, we have the following sure retention property,

$$\text{pr}(\hat{R} \subseteq S) = 1 - O(p_n \exp(-c_2 n^{(1-2\kappa)/4})).$$

(3)

The essential part of proof for Proposition 2 follows an exponential inequality for the quasi-maximum likelihood estimator in Fan and Song (2010). Condition 5 puts an upper bound on the maximum marginal correlation between covariates and the response, and is a technical condition required to achieve the convergence rate in (2).
Condition 6 bounds $\text{var}(Y_1)$ as in Fan and Lv (2008). $\zeta_n$ is the maximum correlation magnitude between noises and response. The choice of the threshold $\gamma_n$ is essential for sure retention.

Equation (3) may not be informative if the threshold $\gamma_n$ is set too high so that $\hat{R}$ is an empty set. Before quantifying how large $\hat{R}$ is, define the marginal strong signal set $R = \{j \in S : |\beta^M_j| > \zeta_n + 2c_1n^{-\kappa}\}$. On the set $\{\max_{1 \leq j \leq p_n} |\hat{\beta}^M_j - \beta^M_j| \leq c_1n^{-\kappa}\}$, we have $\{|\hat{\beta}^M_j| > \zeta_n + 2c_1n^{-\kappa}\} \subseteq \{|\hat{\beta}^M_j| > \zeta_n + c_1n^{-\kappa}\}$ holds for any $j$. Thus,

$$\text{pr}(R \subseteq \hat{R}) \geq 1 - O(p_n \exp(-c_2n^{(1-2\kappa)/4})).$$

Equation (4) indicates that our retention set $\hat{R}$ contains the marginal strong signal set $R$ with high probability when the dimensionality $p_n$ satisfying $\log p_n = o(n^{(1-2\kappa)/4})$. It will be clear from the conditions in the next subsection that the size of $R$ plays an important role in achieving sign consistency for $\hat{\beta}$.

### 3.3 Sign consistency in the regularization step

In the retention step, we can detect part of signals with high probability, including the marginal strong signal set $R$. Incorporating this information into the regularization step, namely not penalizing the retained signals, we can show that the sign consistency of $\ell_1$ regularized least square holds in weaker conditions.

**Condition 7.** $\log p_n = O(n^{a_1})$, $s_n = O(n^{a_2})$, where $0 < a_1 < (1 - 2\kappa)/4$ with $\kappa$ the same as in Condition 5, $a_2 > 0$, and $\max(a_1, a_2) + a_2 < 1$.

**Condition 8.** $\Lambda_{\text{min}}(\Sigma_{SS}) \geq C_{\text{min}} > 0$.

**Condition 9.** $\|\{\Sigma_{SS}^{-1}\}^S \cap R\|\infty \leq 1 - \gamma$, $\gamma \in (0, 1]$. 


**Condition 10.** \( \min_{j \in S} |\beta_j| \geq Cn^{-\delta+\alpha_2/2} \) for a sufficient large \( C \), where \( 0 < \delta < \{1 - \max(a_1, a_2)\}/2 \).

**Theorem 1.** Under Conditions 5-10, if \( s_n \to \infty \) and \( \lambda_n \asymp n^{-\delta} \), our two-step estimator \( \hat{\beta} \) achieves sign consistency,

\[
\Pr(\hat{\beta} \text{ is unique and } \text{sign}(\hat{\beta}) = \text{sign}(\beta)) \to 1, \quad \text{as } n \to \infty.
\]

Our proof follows the essential techniques of the proof in Wainwright (2009). That is why Conditions 8–10 share similarity with Conditions 1–4. The key difference is to prove that with high probability, the estimator in the second step recovers the signs when \( S_1 \) is not penalized, uniformly for all sets \( S_1 \) satisfying \( R \subseteq S_1 \subseteq S \). Since the retention set \( \hat{R} \) in the first step satisfies \( R \subseteq \hat{R} \subseteq S \) with high probability from Corollary 1, the final two-step estimator achieves sign consistency.

Condition 9 is a weaker version of irrepresentable Condition 2. Each row of \( \Sigma_{S^cS} \Sigma_{SS}^{-1} \) can be regarded as the regression coefficients (population version) by regressing the corresponding noise on signals. Thus, Condition 2 requires that for each noise, the sum of the absolute values of its regression coefficients is less than \( 1 - \gamma \). In contrast, the corresponding sum in Condition 9 excludes coefficients corresponding to the retained signals. As a result, we allow larger regression coefficients for the retained signals. Note that regression coefficients measure the dependency between response and regressors. In this sense, our method allows stronger dependency between noises and the retained signals. How much we gain by conducting the first step largely depends on the size of the strong signal set \( R \). The larger \( R \) is, the greater improvement our method can make over lasso.
3.4 The redemption in a third step

In Theorem 1, the threshold $\gamma_n$ is required to be higher than the maximum marginal magnitude between response and noise. In practice, we propose to select $\gamma_n$ by a permutation-based method. Denote $m$ randomly permuted response vectors by $Y^{(1)}, \ldots, Y^{(m)}$. Let the marginal regression coefficients from the permuted data be

$$D^j_k = \frac{\sum_{i=1}^n (X^j_i - \bar{X}^j)Y^{(k)}_i}{\sum_{i=1}^n (X^j_i - \bar{X}^j)^2}, 1 \leq j \leq p, 1 \leq k \leq m.$$ 

Then we set the tentative threshold $\gamma_n = \max_{k,j} |D^j_k|$. Intuitively, if noises are not strongly correlated with response, the maximum absolute value of marginal regression coefficients from permutation should be a reasonable threshold. If this tentative threshold leads to a retention set with size larger than $\lceil n^{1/2} \rceil$, we then retain only the top $\lceil n^{1/2} \rceil$ variables with the largest magnitudes of the marginal coefficients $|\hat{\beta}_j^M|$. This ensures that there are at most $\lceil n^{1/2} \rceil$ variables not penalized in the second step. The choice of $\lceil n^{1/2} \rceil$ is inspired by the $n^{1/2}$-consistency of the classical least square estimators.

In case there are falsely retained variables in the first step, we propose to add one extra step to remove them. Denote by $Q$ the additional signals detected in the regularization step, that is $Q = \{ j \in \hat{R}^c : \tilde{\beta}_j \neq 0 \}$.

**Step 3** Calculate the following penalized least square problem

$$\tilde{\beta} = \arg \min_{\beta \in (\hat{R}^c \cup Q)^c} \left\{ (2n)^{-1} \sum_{i=1}^n (Y_i - \sum_{j \in \hat{R}} X_{ij} \beta_j - \sum_{k \in Q} X_{ik} \hat{\beta}_k)^2 + \lambda_* \sum_{j \in \hat{R}} |\beta_j| \right\},$$

where $\lambda_* \lambda_n$ is the penalty parameter, which is in general different from $\lambda_n$ in the second step. The idea is to regularize only the coefficients in the retained set $\hat{R}$ while keeping
the signals identified in $Q$. The three-step estimator $\tilde{\beta}$ is called modified regularization after retention (MRAR). It turns out that under certain regularity conditions, the three-step estimator $\tilde{\beta}$ achieves sign consistency. To this end, we define a strong noise set $Z = \{ j \in S^c : |\beta^M_j| \geq \gamma_n - c_1 n^{-\kappa} \}$ with its cardinality $z_n$. Recall the strong signal set $R = \{ j \in S : |\beta_j^M| \geq \gamma_n + c_1 n^{-\kappa} \}$. The new regularity conditions are as follows.

**Condition 11.** $\Lambda_{\min}(\Sigma_{S \cup Z, S \cup Z}) \geq C_{\min} > 0$.

**Condition 12.** $\max_{S \subset Q \subset S \cup Z} \| \Sigma_{Q^C} (\Sigma_{QQ}^{-1})_{S \cap R^c} \|_{\infty} \leq 1 - \gamma$, where $\gamma > 0$.

**Condition 13.** $\| \Sigma_{ZS} \Sigma_{SS}^{-1} \|_{\infty} \leq 1 - \alpha$, where $\alpha > 0$.

**Theorem 2.** Under Conditions 5-7 and 10-13, if $z_n/s_n \to 0$, $s_n \to \infty$ and $\lambda_n \asymp n^{-\delta}$, $\lambda^*_n \asymp n^{-\delta}$, our three-step estimator $\tilde{\beta}$ achieves sign consistency,

$$\text{pr}(\tilde{\beta} \text{ is unique and } \text{sign}(\tilde{\beta}) = \text{sign}(\beta)) \to 1, \text{ as } n \to \infty.$$ 

The proof is an extension of the proof for Theorem 1. Conditions 11-13 are generalizations of Conditions 8-9. They essentially require the possible noises selected in the retention step, denoted by the set $Z$, cannot be highly correlated with the signals. Theorem 2 provides insight into the robustness of the three-step approach. Compared to RAR, to achieve sign consistency, MRAR is able to tolerate false retention at a level quantified by Conditions 11-13.

### 3.5 Connections to SIS-lasso and adaptive lasso

In this section, we highlight the connections of RAR with sure independence screening followed by lasso (SIS-lasso) and adaptive lasso (Ada-lasso). In the first step, both RAR and SIS-lasso calculate and rank the marginal regression coefficient estimates.
In the second step, the estimator for RAR can be written as

$$
\hat{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + 0 \sum_{j \in \hat{R}} |\beta_j| + \lambda_n \sum_{j \in \hat{R}^c} |\beta_j| \right\},
$$

while the estimator for SIS-lasso is

$$
\arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j \in \hat{S}} |\beta_j| + \infty \sum_{j \in \hat{S}^c} |\beta_j| \right\},
$$

where \( \hat{S}^c \) is the set of the screened-out variables in Step 1 of SIS-lasso.

Both methods relax the consistency condition of lasso \( \| \Sigma_{S^c S} \Sigma_{SS}^{-1} \|_\infty \leq 1 - \gamma \). SIS-lasso reduces \( \| \Sigma_{S^c S} \Sigma_{SS}^{-1} \|_\infty \) by removing rows of \( \Sigma_{S^c S} \Sigma_{SS}^{-1} \) corresponding to the screened-out noises. RAR reduces \( \| \Sigma_{S^c S} \Sigma_{SS}^{-1} \|_\infty \) by removing columns of \( \Sigma_{S^c S} \Sigma_{SS}^{-1} \) corresponding to the retained signals. Although the number of removed rows by SIS is typically larger than that of removed columns by RAR, it does not necessarily mean that the amount of reduction by SIS will be greater than that by RAR. For example, if there exist signals highly correlated to noises (i.e., scenario 1(A) in Section 4.1), retaining the signal with the largest marginal correlation will substantially decrease \( \| \Sigma_{S^c S} \Sigma_{SS}^{-1} \|_\infty \), while removing noises with small marginal correlations does not change \( \| \Sigma_{S^c S} \Sigma_{SS}^{-1} \|_\infty \) at all.

(5) and (6) lead to a natural comparison with the adaptive lasso (Zou, 2006) estimator:

$$
\arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\},
$$

where the weight \( w_j \) is usually chosen as \( 1/|\beta_{j,\text{init}}|^\gamma \) for some \( \gamma > 0 \) using an initial estimator \( \beta_{j,\text{init}} \). For fixed design, Zou (2006) proved that the adaptive lasso estimator
achieves variable selection consistency under very mild conditions when $p$ is fixed. In the high dimensional regime, Huang et al. (2008) showed variable selection consistency with $w_j = 1/|\hat{\beta}_j^M|$ under the partial orthogonality condition (i.e., signals are weakly correlated to noises). A more general theoretical treatment is given in Zhou et al. (2009) under the restricted eigenvalue conditions (Bickel et al., 2009) for both fixed and random designs.

All of (5)-(7) aim at improving lasso by adaptively adjusting the penalty level for each predictor. The major difference between (5)-(6) and (7) is that (7) uses “soft” weights while both (5) and (6) use “thresholded” weights. For (7), it is possible that there exists $\beta_{j,\text{init}} \approx 0$ for some signal $j$ with small marginal correlation, leading to a very large weight for that variable, which makes the consistent selection difficult. Due to the specific thresholding choices, a similar observation can be found for (6). In contrast, (5) can still succeed in sparse recovery.

4 Numerical studies

4.1 Simulations

We compare the variable selection performances of lasso, Ada-lasso, SIS-lasso, iterative sure independence screening (ISIS-lasso), RAR and MRAR in the ultrahigh dimensional linear regression setting. We set $n = 100, 200, 300, 400, 500$ and $p_n = \lfloor 100 \exp(n^{0.2}) \rfloor$, where $\lfloor k \rfloor$ is the largest integer not exceeding $k$. The number of repetitions is 200 for each triplet $(n, s_n, p_n)$. We calculate the proportion of exact sign recovery. All the lasso procedures are implemented using the R package glmnet (Friedman et al., 2010).
Scenario 1. The covariance matrix $\Sigma$ is

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & I \end{bmatrix}, \text{ where } \Sigma_{11} = \begin{bmatrix} 1 & \ldots & r \\ \vdots & \ddots & \vdots \\ r & \ldots & 1 \end{bmatrix}_{2s \times 2s}$$

(A). $r = 0.6, \sigma = 3.5, s_n = 4, \beta_S = (3, -2, 2, -2), \beta = (\beta_S, 0)$. The absolute correlations between response and predictors are (0.390, 0.043, 0.304, 0.043, 0.130, 0.130, 0.130, 0.130, 0, 0, \ldots).

(B). $r = 0.6, \sigma = 1.2, s_n = 5, \beta_S = (1, 1, -1, 1, -1), \beta = (\beta_S, 0)$. The absolute correlations between response and predictors are (0.498, 0.498, 0.100, 0.498, 0.100, 0.296, 0.296, 0.296, 0.296, 0, 0, \ldots).

Scenario 2. The covariance matrix $\Sigma$ is

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & I \end{bmatrix}, \text{ where } \Sigma_{11} = \begin{bmatrix} 1 & r_0 & r_1 & r_3 \\ r_0 & 1 & r_2 & r_4 \\ r_1 & r_2 & 1 & 0 \\ r_3 & r_4 & 0 & 1 \end{bmatrix}$$

(C). $r_0 = 0.8, r_1 = -r_2 = r_3 = -r_4 = -0.1, \sigma = 2.5, s_n = 2, \beta_S = (2.5, -2), \beta = (\beta_S, 0)$. The absolute correlations between response and predictors are (0.309, 0.000, 0.154, 0.154, 0, 0, \ldots).

(D). $r_0 = 0.75, r_1 = r_2 = r_3 = -r_4 = 0.2, \sigma = 2.5, s_n = 2, \beta_S = (2.5, -2), \beta = (\beta_S, 0)$. The absolute correlations between response and predictors are (0.333, 0.0417, 0.033, 0.300, 0, 0, \ldots).

Since data driven methods for tuning parameter selection introduce extra random-
ness into the entire variable selection process, we report the oracle performance of each method for fair comparison. Specifically, for lasso, Ada-lasso, the regularization steps of SIS-lasso, RAR and MRAR, we check if there exists at least one estimator on the solution path recovering sign sparsity pattern. For SIS-lasso and ISIS-lasso, we select the top $\lfloor n/\log n \rfloor$ variables with the largest absolute marginal correlation in the first step. For Ada-lasso, following Huang et al. (2008), we choose the weights $w_j = 1/|\hat{\beta}_j^M|$. 

For Scenario 1, SIS-lasso fails to recover the sparsity pattern in both 1(A) and 1(B), due to that some signals and noises have correlations in similar magnitude with the response. ISIS-lasso substantially improves the performance of SIS-lasso and has similar performance as lasso. The possible reason why it does not show clear advantage over lasso is that the discrete stochastic process of the iterative algorithm may induce too much randomness. Ada-lasso is outperformed by lasso for both 1(A) and 1(B), with the possible reason being that the weights are close to infinity for signals with small marginal correlation. Both RAR and MRAR work very well in 1(A). For 1(B), RAR fails due to that there are noises with very large marginal correlation while MRAR still has competitive performance. The performance for both RAR and MRAR is not much affected by the number of permutations. Note that MRAR with any number of permutations provides better performance than any non-RAR methods in both 1(A) and 1(B) across all $(n, p_n)$ pairs.

We design the challenging Scenario 2 for lasso. For both 2(C) and 2(D), lasso, SIS-lasso, ISIS-lasso and Ada-lasso all perform poorly. In contrast, RAR and MRAR have similar performances as Scenario 1.
Table 1: Sign recovery proportion over 200 simulation rounds for the oracle performance of each method. The subscript for RAR and MRAR denotes the number of permutations in the retention step.

\[(n, p_n)\] \hspace{1cm} (100, 1232) \hspace{1cm} (200, 1791) \hspace{1cm} (300, 2285) \hspace{1cm} (400, 2750) \hspace{1cm} (500, 3199)

| Scenario 1 (A) | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} |
|----------------|----------------|----------------|----------------|----------------|----------------|
| Lasso          | 0.000          | 0.000          | 0.040          | 0.240          | 0.485          |
| SIS-lasso      | 0.000          | 0.000          | 0.005          | 0.000          | 0.045          |
| ISIS-lasso     | 0.000          | 0.000          | 0.060          | 0.180          | 0.500          |
| Ada-lasso      | 0.000          | 0.000          | 0.005          | 0.015          | 0.050          |
| RAR\(_1\)      | 0.025          | 0.295          | 0.385          | 0.365          | 0.320          |
| RAR\(_5\)      | 0.010          | 0.405          | 0.615          | 0.685          | 0.690          |
| RAR\(_{10}\)   | 0.005          | 0.340          | 0.685          | 0.780          | 0.740          |
| RAR\(_{15}\)   | 0.000          | 0.325          | 0.745          | 0.785          | 0.800          |
| RAR\(_{30}\)   | 0.000          | 0.295          | 0.715          | 0.845          | 0.835          |
| MRAR\(_1\)     | 0.035          | 0.560          | 0.890          | 0.995          | 1.000          |
| MRAR\(_5\)     | 0.010          | 0.470          | 0.845          | 0.985          | 0.995          |
| MRAR\(_{10}\)  | 0.010          | 0.380          | 0.835          | 0.985          | 0.995          |
| MRAR\(_{15}\)  | 0.000          | 0.370          | 0.825          | 0.975          | 0.995          |
| MRAR\(_{30}\)  | 0.000          | 0.310          | 0.775          | 0.950          | 0.985          |

| Scenario 1 (B) | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} | \hspace{1cm} |
|----------------|----------------|----------------|----------------|----------------|----------------|
| Lasso          | 0.000          | 0.005          | 0.175          | 0.640          | 0.870          |
| SIS-lasso      | 0.000          | 0.000          | 0.000          | 0.100          | 0.255          |
| ISIS-lasso     | 0.000          | 0.005          | 0.155          | 0.565          | 0.885          |
| Ada-lasso      | 0.000          | 0.030          | 0.175          | 0.350          | 0.420          |
| RAR\(_1\)      | 0.105          | 0.025          | 0.005          | 0.000          | 0.000          |
| RAR\(_5\)      | 0.180          | 0.100          | 0.010          | 0.000          | 0.000          |
| RAR\(_{10}\)   | 0.165          | 0.190          | 0.030          | 0.000          | 0.000          |
| RAR\(_{15}\)   | 0.185          | 0.180          | 0.025          | 0.000          | 0.000          |
| RAR\(_{30}\)   | 0.175          | 0.225          | 0.040          | 0.000          | 0.000          |
| MRAR\(_1\)     | 0.265          | 0.850          | 0.940          | 0.985          | 0.985          |
| MRAR\(_5\)     | 0.265          | 0.900          | 0.985          | 0.995          | 0.985          |
| MRAR\(_{10}\)  | 0.240          | 0.910          | 0.990          | 1.000          | 0.995          |
| MRAR\(_{15}\)  | 0.215          | 0.915          | 0.995          | 1.000          | 0.995          |
| MRAR\(_{30}\)  | 0.205          | 0.935          | 0.990          | 1.000          | 1.000          |
Table 2: Sign recovery proportion over 200 simulation rounds for the oracle performance of each method. The subscript for RAR and MRAR denotes the number of permutations in the retention step.

\[(n, p_n)\] \hspace{1cm} (100, 1232) \hspace{1cm} (200, 1791) \hspace{1cm} (300, 2285) \hspace{1cm} (400, 2750) \hspace{1cm} (500, 3199)

| Method  | Scenario 2 (C) | Scenario 2 (D) |
|---------|----------------|----------------|
| Lasso   |                |                |
| SIS-lasso |                |                |
| ISIS-lasso |                |                |
| Ada-lasso |                |                |
| RAR_1   | 0.000          | 0.000          |
| RAR_5   | 0.000          | 0.000          |
| RAR_10  | 0.000          | 0.000          |
| RAR_15  | 0.000          | 0.000          |
| RAR_30  | 0.000          | 0.000          |
| MRAR_1  | 0.005          | 0.005          |
| MRAR_5  | 0.000          | 0.000          |
| MRAR_10 | 0.000          | 0.000          |
| MRAR_15 | 0.000          | 0.000          |
| MRAR_30 | 0.000          | 0.000          |
| RAR_1   | 0.150          | 0.160          |
| RAR_5   | 0.180          | 0.185          |
| RAR_10  | 0.210          | 0.205          |
| RAR_15  | 0.235          | 0.220          |
| RAR_30  | 0.255          | 0.260          |
| MRAR_1  | 0.285          | 0.260          |
| MRAR_5  | 0.785          | 0.530          |
| MRAR_10 | 0.870          | 0.505          |
| MRAR_15 | 0.910          | 0.860          |
| MRAR_30 | 0.945          | 0.845          |

19
4.2 Real data application

We compare the performances of lasso, SIS-lasso, ISIS-lasso, Ada-lasso, RAR, and MRAR on the data set reported by Scheetz et al. (2006). For this data set, 120 twelve-week old male rats were selected for tissue harvesting from the eyes. The microarrays used to analyze the RNA from the eyes of these rats contain over 31,042 different probes (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multi-chip averaging method (Irizarry et al., 2003) to obtain summary expression values for each probe. Gene expression levels were analyzed on a logarithmic scale. Following Fan et al. (2011), we only focus on the 18,975 probes that are expressed in the eye tissue. We are interested in finding the genes that are related to the gene TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang et al., 2006), and is a genetically heterogeneous disease of multiple organ systems including the retina. The dataset includes \( n = 120 \) samples and \( p = 18,975 \) variables. We randomly partitioned the data into a training set of 96 observations and a test set of 24 observations. We use 5-fold cross validation for tuning parameter selection on the training set for the last regularization step of each method and calculate the prediction mean square error on the test set. For the second step of MRAR, generalized information criterion is employed (Fan and Tang, 2012). The whole procedure is repeated 200 times.

As shown in Table 3, RAR performs the best in terms of prediction error and selects fewer variables than lasso does. MRAR selects the most parsimonious model with slightly larger prediction error than RAR and Ada-lasso. Independence screening-based methods lead to sparser models than RAR, but they have the largest prediction errors. For SIS-lasso and ISIS-lasso, we select the top 60 variables in the screening step. We also try other thresholds and they lead to similar results. The reason may be that there exist signals that are weakly correlated with the response so that even
Table 3: Average prediction mean square error and the average model size over 200 repetitions. The superscript * indicates the performance of the optimal estimator in terms of prediction on the entire solution path. The standard deviations of the error or model size are enclosed in parentheses.

|               | Lasso      | SIS-lasso  | ISIS-lasso | Ada-lasso | RAR       | MRAR       |
|---------------|------------|------------|------------|-----------|-----------|------------|
| Error (%)     | 0.72 (0.336)| 0.83 (0.411)| 0.82 (0.369)| 0.67 (0.262)| **0.66** (0.239)| 0.69 (0.261) |
| Error* (%)    | 0.62 (0.273)| 0.68 (0.289)| 0.70 (0.301)| **0.60** (0.232)| 0.61 (0.223) | 0.61 (0.194) |
| Fitted model size | 63.0 (19.18) | 24.0 (6.98) | 20.3 (8.44) | 67.1 (10.68) | 39.9 (4.41) | **6.3** (0.94) |

ISIS-lasso misses them.

To show the potential of each method, we also report the optimal prediction error on the entire solution path. Ada-lasso performs the best, at the cost of selecting the most variables. RAR and MRAR have very close performance to Ada-lasso.

5 Discussion

The proposed regularization after retention method is a general framework for model selection and estimation. In the retention step, there exist alternatives to obtain the retention set beyond those using marginal information. For example, we can use forward regression with early stopping. In the regularization step, it is straightforward to replace lasso with other penalized methods such as smoothly clipped absolute deviation (Fan and Li, 2001) and adaptive lasso (Zou, 2006).

We conjecture that the permutation approach for choosing the threshold $\gamma_n$ can effectively control the false retention rate to close to zero. Theoretical extension for sub-Gaussian distributions of $X$ and $\varepsilon$ is possible. Similar results may also hold in more general models including generalized linear models, additive models, and semi-parametric models.
Appendix

Proof of Proposition 1. We refer to the general result of Theorem 3 in Wainwright (2009). By Condition 3, we have as \( n \to \infty \),

\[
\lambda_n^2 n / \{ \rho(\Sigma_{S|S}) \log p_n \} > n^\delta / \rho(\Sigma_{S|S}) = 1/o(1) \to \infty, \tag{8}
\]

\[
\frac{n}{\{ \rho(\Sigma_{S|S}) s_n \log (p_n - s_n) \}} > n / \{ \rho(\Sigma_{S|S}) s_n \log p_n \} > n^{1-a_1^{-2a_2-\delta}} / o(1) \to \infty. \tag{9}
\]

With the same notations as in Wainwright (2009), (8) implies \( \phi_p \to \infty \), and (9) shows that equation (34) in Wainwright (2009) holds. Then we need to show \( \min_{j \in S} |\beta_j| > g(\lambda_n) \), for sufficient large \( n \), where \( g(\lambda_n) \) is defined by equation (33) in Wainwright (2009),

\[
g(\lambda_n) \leq c_3 \lambda_n s_n \Lambda_{\max}^2 (\Sigma_{SS}^{-1/2}) + O((\log n / n)^{1/2}) = O(n^{(d+a_1+2a_2-1)/2}) < \min_{j \in S} |\beta_j|,
\]

due to \( \Lambda_{\max}^2 (\Sigma_{SS}^{-1/2}) = \Lambda_{\max}(\Sigma_{SS}^{-1}) = 1 / \Lambda_{\min}(\Sigma_{SS}) \) and Conditions 1, 3, 4. \( \square \)

Proof of Proposition 2. We refer to Theorem 4 in Fan and Song (2010) about uniform convergence of maximum marginal likelihood estimator in the generalized linear model. As in Fan and Song (2010), let \( \mathbf{\beta}_j = (\beta_{j,0}, \beta_j) \) and \( \mathbf{\beta} = (\beta_0, \beta_1) \) be two-dimensional vectors and \( \mathbf{X}_j = (1, X_j)^T \), where \( X_j \) is the \( j \)th predictor. Denote \( \mathcal{B} = \{ |\beta_{j,0}| \leq B, |\beta_j| \leq B \} \) and the true coefficient by \( \mathbf{\beta}^* \). We need to verify Conditions \( A', B', C' \) and \( D \) in Fan and Song (2010). Condition \( A' \) is trivial since \( b(\theta) = \theta^2 / 2 \).

The expected marginal loglikelihood \( E \{ l(\mathbf{X}_j^T \mathbf{\beta}_j, Y) \} \) is a quadratic function of \( \mathbf{\beta}_j \) with identity Hessian matrix. Thus, Condition \( C' \) is satisfied with \( V = 1/2 \). For Condition
$$|E\{b(X_j^T \beta)I(|X_j| > K_n)\}| = \frac{\beta_0^2}{2} P(|X_j| > K_n) + \frac{\beta_1^2}{2} E\{X_j^2 I(|X_j| > K_n)\}$$

$$\leq \beta_0^2 e^{-K_n^2/2} + \beta_1^2 (1 + K_n) e^{-K_n^2/2}.$$  

(11)

(10) holds because $X_j$ is symmetric. (11) follows from the standard normal distribution of $X_j$. As mentioned in Section 5.2 in Fan and Song (2010), the optimal order of $K_n = n^{(1-2\kappa)/8}$. Taking $B = O(n^{(1-2\kappa)/8})$ and using Condition 5, we get for any $\varepsilon > 0$,

$$\sup_{\beta \in B, \|\beta - \beta_0^2\| \leq \varepsilon} |E\{b(X_j^T \beta)I(|X_j| > K_n)\}| \leq o(1/n).$$

With moment generating function of $|X_j|$, we know $Pr(|X_j| > t) \leq 2 \exp(-t^2/2)$ and

$$E\{\exp(b(X_0^T \beta + s_0) - b(X_0^T \beta^*))\} + E\{\exp(b(X^T \beta^* - s_0) - b(X^T \beta^*))\} = 2 \exp(s_0^2(1 + (\beta^*)^T \Sigma \beta^*)/2).$$

(13)

Let positive constants $\alpha = 2, m_0 = 1/2, m_1 - s_1 = 2, s_0 = 1$. By Condition 6, there exists positive constant $s_1$ such that (12) $\leq s_1$. Therefore, Condition D holds.

Proof of Theorem

Denote the design matrix by $X$, response vector by $Y$, and error vector by $\varepsilon$. Define $\bar{S} = \tilde{R}e \setminus S^c$. Let

$$\hat{\beta} = \arg\min_{\beta} \left\{ (2n)^{-\frac{1}{2}} \|Y - X_{\tilde{R}} \beta_{\tilde{R}} - X_{\tilde{R}} \|_2^2 + \lambda_n \|\beta_{\tilde{R}}\|_1 \right\},$$

(14)

$$\tilde{\beta} = \arg\min_{\beta_{\tilde{S}} = 0} \left\{ (2n)^{-\frac{1}{2}} \|Y - X_S \beta_S\|_2^2 + \lambda_n \|\beta_{\tilde{S}}\|_1 \right\}.$$

(15)
Since $X_S^T X_S \sim W_{s_n}(\Sigma_{SS}, n)$, which is Wishart distribution, when the number of signals $s_n < n$, as in our scaling, $X_S$ is of full rank with probability one. Therefore, (15) is a strictly convex problem and $\beta$ is unique with probability one.

By optimality conditions of convex problems, $\tilde{\beta}$ is a solution to (14) if and only if

$$
n^{-1} X_T^T (Y - X \tilde{\beta}) = \lambda_n \partial \| \tilde{\beta}_{R^c} \|, \tag{16}$$

where $\partial \| \tilde{\beta}_{R^c} \|$ is the subgradient of $\| \tilde{\beta}_{R^c} \|_1$ at $\beta = \tilde{\beta}$. Namely, the $i$th $(1 \leq i \leq p_n)$ element of $\partial \| \tilde{\beta}_{R^c} \|$ is

$$
(\partial \| \tilde{\beta}_{R^c} \|)_i = \begin{cases} 
0 & \text{if } i \in \hat{R} \\
\text{sign}(\tilde{\beta}_i) & \text{if } i \in \hat{R}^c \text{ and } \tilde{\beta}_i \neq 0 \\
t & \text{otherwise}
\end{cases}
$$

where $t$ can be any real number with $|t| \leq 1$. Similarly, $\bar{\beta}$ is the unique solution to (15) if and only if

$$
\bar{\beta}_{S^c} = 0, \quad n^{-1} X_{S^c}^T (Y - X_S \bar{\beta}_S) = \lambda_n \text{sig}(\bar{\beta}_S), \tag{17}
$$

where $\text{sig}(\bar{\beta}_S)$, a vector of length $s_n$, is the subgradient of $\| \beta_{S^c} \|$ at $\beta_S = \bar{\beta}_S$. Then it is not hard to see that, the unique solution $\bar{\beta}$ is also a solution for (14) if

$$
\| n^{-1} X_{S^c}^T (Y - X_S \bar{\beta}_S) \|_\infty < \lambda_n, \tag{18}
$$

simply because (17) and (18) imply $\bar{\beta}$ satisfies (16). Solving the equation in (17) gives

$$
\bar{\beta}_S = (X_S^T X_S)^{-1} \left[ X_S^T Y - n \lambda_n \text{sig}(\bar{\beta}_S) \right]. \tag{19}
$$
Using (19) and $Y = X_S \beta_S + \varepsilon$, (18) is equivalent to

$$\|X^T_S X_S (X^T_S X_S)^{-1} \text{sign}(\bar{\beta}_S) + (n \lambda_n)^{-1} X^T_S I - X_S (X^T_S X_S)^{-1} X^T_S \|_\infty < 1. \quad (20)$$

Based on the optimality conditions of convex problem, we have showed that if the optimization problem (15)’s unique solution $\bar{\beta}$ satisfies (20), then $\bar{\beta}$ is also a solution to (14). On the other hand, it is easily seen that, for any solution $\tilde{\beta}$ to (14), $\text{sign}(\tilde{\beta}) = \text{sign}(\beta)$ only if $\tilde{\beta}$ is also a solution to (15). Therefore, If (14) has a unique solution and $\bar{\beta}$ satisfies (20), then $\bar{\beta}$ is that unique solution and $\text{Supp}\{\bar{\beta}\} \subseteq S$. Furthermore, if the maximum gap $\|\bar{\beta}_S - \beta_S\|_\infty$ is upper bounded by the minimum absolute magnitude of $\beta_S$, we can achieve sign recovery. In summary, let

$$W = \{\tilde{\beta} \text{ is unique and } \text{sign}(\tilde{\beta}) = \text{sign}(\beta)\},$$

$$W_1 = \{(14) \text{ has a unique solution and (20) holds}\},$$

$$W_2 = \{\min_{j \in S} |\hat{\beta}_j| > \|\bar{\beta}_S - \beta_S\|_\infty\}.$$  

Then, we have

$$\Pr(W) \geq \Pr(W_1 \cap W_2) \geq 1 - \Pr(W_1^c) - \Pr(W_2^c) = \Pr(W_1) - \Pr(W_2^c). \quad (21)$$

In the following, we will show $P(W_1) \to 1$ and $P(W_2^c) \to 0$ in two steps separately. Since (14) is similar to lasso in random design, our proof mainly follows the proof of Theorem 3 in [Wainwright (2009)]. The key difference is that the penalty term in (14) is random due to the retention step of our method. To take care of that part, we need
more notations. Let

\[ T = \{ S_* : R \subseteq S_* \subseteq S \}, \]
\[ A = \{ R \subseteq \hat{R} \subseteq S \}, \]
\[ B = \left\{ \max_{1 \leq j \leq p_n} |\hat{\beta}_j^M - \beta_j^M| \leq c_1 n^{-\kappa} \right\}. \]

Then \( B \subseteq A \) and \( \text{pr}(B) = 1 - O(p_n \exp(-c_2 n^{(1-2\kappa)/4})), \) as we discussed in Section 3.2.

**Step I.** Let \( F = X_S^T - \Sigma_{S^cS} \Sigma_{SS}^{-1} X_S^T, \) and \( F(j) \) be the \( j \)th row of \( F. \) By the property of conditional distribution of multivariate Gaussian, \( F^1, \ldots, F^n \) are independently and identically distributed as \( N(0, \Sigma_{S^c|S}) \), and \( F \) is independent of \( X_S. \) After simple algebra calculation using \( X_S^T = \Sigma_{S^cS} \Sigma_{SS}^{-1} X_S + F, \) we get

\[
X_S^T X_S (X_S^T X_S)^{-1} \text{sig}(\bar{\beta}_S) + (n \lambda_n)^{-1} X_S^T \{ I - X_S (X_S^T X_S)^{-1} X_S^T \} \varepsilon
\]
\[
= \Sigma_{S^cS} \Sigma_{SS}^{-1} \text{sig}(\bar{\beta}_S) + F X_S (X_S^T X_S)^{-1} \text{sig}(\bar{\beta}_S) + (n \lambda_n)^{-1} F \{ I - X_S (X_S^T X_S)^{-1} X_S^T \} \varepsilon.
\]

Let \( K_1 = \Sigma_{S^cS} \Sigma_{SS}^{-1} \text{sig}(\bar{\beta}_S) \) and \( K_2 = F X_S (X_S^T X_S)^{-1} \text{sig}(\bar{\beta}_S) + (n \lambda_n)^{-1} F \{ I - X_S (X_S^T X_S)^{-1} X_S^T \} \varepsilon. \) Then (20) is equivalent to \( \| K_1 + K_2 \|_\infty < 1. \) We analyze \( \| K_1 \|_\infty \) and \( \| K_2 \|_\infty \) on the high probability set \( A. \) Firstly, it is not hard to see,

\[
\text{pr}(\| K_1 \|_\infty \leq 1 - \gamma) = \text{pr}(\{ \| K_1 \|_\infty \leq 1 - \gamma \} \cap A) + \text{pr}(\{ \| K_1 \|_\infty \leq 1 - \gamma \} \cap A^c)
\]
\[
\overset{(1)}{=} \text{pr}(A) + \text{pr}(\{ \| K_1 \|_\infty \leq 1 - \gamma \} \cap A^c),
\]

where (1) holds since when \( A \) holds, by condition 10,

\[
\| K_1 \|_\infty \leq \| \Sigma_{S^cS} (\Sigma_{SS})^{-1} \|_{S \cap R^c} \|_\infty \leq 1 - \gamma.
\]
Under the scaling in theorem 1, \( \text{pr}(A) \to 1 \), \( \text{pr}(\{\|K_1\|_\infty \leq 1 - \gamma\} \cap A^c) \leq \text{pr}(A^c) \to 0 \), as \( n \to \infty \). Hence,

\[
\text{pr}(\|K_1\|_\infty \leq 1 - \gamma) \to 1, \quad \text{as} \quad n \to \infty.
\]

Similarly,

\[
\text{pr}(\|K_2\|_\infty > \frac{\gamma}{2}) = \text{pr}(\{\|K_2\|_\infty > \frac{\gamma}{2}\} \cap A) + \text{pr}(\{\|K_2\|_\infty > \frac{\gamma}{2}\} \cap A^c)
\leq \text{pr}\left(\bigcup_{S_1 \in T} \{\|K_2(S_1)\|_\infty > \frac{\gamma}{2}\} \cap A\right) + \text{pr}(A^c)
\leq \text{pr}\left(\bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty > \frac{\gamma}{2}\right) + \text{pr}(A^c),
\]

where \( K_2(S_1) \) is the analogy of \( K_2 \) in (22) by replacing \( \hat{R} \) with \( S_1 \) in (14) and (15). Denote the corresponding solution to (15) by \( \bar{\beta}(S_1) \). Then,

\[
K_2(S_1) = FX_S(X^T_SX_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) + (n\lambda_n)^{-1}F\{I - X_S(X^T_SX_S)^{-1}X^T_S\}\varepsilon.
\]

By the definition of \( \bar{\beta}(S_1) \), \( \text{sig}(\bar{\beta}_S(S_1)) \) is a function of \( X_S \) and \( \varepsilon \), so

\[
F(j)X_S(X^T_SX_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) + (n\lambda_n)^{-1}F(j)\{I - X_S(X^T_SX_S)^{-1}X^T_S\}\varepsilon \mid (X_S, \varepsilon) \sim N(0, V_j),
\]

and

\[
V_j \leq (\Sigma_{S^c|S})_{jj}[\text{sig}(\bar{\beta}_S(S_1))^T(X^T_SX_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) + (n\lambda_n)^{-2}\varepsilon^T\{I - X_S(X^T_SX_S)^{-1}X^T_S\}\varepsilon]
\leq \text{sig}(\bar{\beta}_S(S_1))^T(X^T_SX_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) + (n\lambda_n)^{-2}\|\varepsilon\|_2^2,
\]

noticing that \( \Sigma_{jj} = 1 \) and \( I - X_S(X^T_SX_S)^{-1}X^T_S \) is an idempotent and symmetric
matrix. Let

\[
H = \bigcup_{S_1 \in T} \left\{ \text{sig}(\bar{\beta}_S(S_1))^{T}(X^T_S X_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) + (n\lambda_n)^{-2}\|\varepsilon\|_2^2 > \frac{s_n}{nC_{\min}} (8s_n^{1/2}n^{-1/2} + 1) + (1 + s_n^{1/2}n^{-1/2})/\lambda_n^2 \right\}.
\]

Then,

\[
\Pr\left( \bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty > \gamma/2 \right) \leq \Pr\left( \bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty > \gamma/2 \mid H^c \right) + \Pr(H). \quad (26)
\]

We first bound \( \Pr(H) \),

\[
\Pr(H) \leq \Pr\left( \bigcup_{S_1 \in T} \text{sig}(\bar{\beta}_S(S_1))^{T}(X^T_S X_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) > \frac{s_n}{nC_{\min}} (8s_n^{1/2}n^{-1/2} + 1) \right)
+ \Pr\left( (n\lambda_n)^{-2}\|\varepsilon\|_2^2 > (1 + s_n^{1/2}n^{-1/2})/\lambda_n^2 \right).
\]

For any \( S_1 \in T \),

\[
\text{sig}(\bar{\beta}_S(S_1))^{T}(X^T_S X_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) \leq s_n\|\left((X^T_S X_S)/n\right)^{-1} - \Sigma_S^{-1}\|_2 \leq \frac{s_n}{n\left(\|\left((X^T_S X_S)/n\right)^{-1} - \Sigma_S^{-1}\|_2 + 1/C_{\min}\right)}.
\]

Therefore,

\[
\Pr\left( \bigcup_{S_1 \in T} \text{sig}(\bar{\beta}_S(S_1))^{T}(X^T_S X_S)^{-1}\text{sig}(\bar{\beta}_S(S_1)) > \frac{s_n}{nC_{\min}} (8s_n^{1/2}n^{-1/2} + 1) \right)
\leq \Pr\left( \|\left((X^T_S X_S)/n\right)^{-1} - \Sigma_S^{-1}\|_2 \geq \frac{8}{C_{\min} s_n^{1/2}n^{-1/2}} \right) \leq 2\exp(-s_n/2), \quad (27)
\]

where we have used the concentration inequality of (58b) in [Wainwright, 2009]. Since
\[ \|\varepsilon\|^2_2 \sim \chi^2(n), \]

using the inequality of (54a) in Wainwright (2009), we get

\[
\Pr\left( (n\lambda_n)^{-2}\|\varepsilon\|^2_2 > (1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2) \right) \leq \Pr\left( \|\varepsilon\|^2_2 \geq (1+s_n^{1/2}n^{-1/2})n \right) \leq \exp(-3/16s_n),
\]

whenever \( s_n/n < 1/2 \). By the tail probability inequality of Gaussian distribution and (25),

\[
\Pr(\bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty \geq \gamma/2 \mid H^c) \leq 2^{s_n+1}(p_n - s_n) \exp(-\gamma^2/(8V)),
\]

where \( V = (1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2) + s_n n^{-1}C_{\min}^{-1}(8s_n^{1/2}n^{-1/2} + 1) \) and we used the cardinality of \( T \) is not larger than \( 2^{s_n} \). Under the scaling of Theorem 1, it is easy to verify that

\[
\log(p_n - s_n) + (s_n + 1) \log 2 = o(\gamma^2/(8V)).
\]

Hence, there exists \( c_1 > 0 \) so that

\[
\Pr(\bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty \geq \gamma/2 \mid H^c) \leq e^{-c_1s_n},
\]

for sufficiently large \( n \). Putting (26), (27), (28), and (29) together, we proved that there exist positive constants \( c_2, c_3, \)

\[
\Pr \left( \bigcup_{S_1 \in T} \|K_2(S_1)\|_\infty > \frac{\gamma}{2} \right) \leq c_2 e^{-c_3s_n}.
\]

(30) and (31) lead to

\[
\Pr(\|K_2\|_\infty > \frac{\gamma}{2}) \to 0, \quad \text{as} \quad n \to \infty.
\]
Then, (23) and (31) imply
\[
\Pr(\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}) \geq \Pr(\|K_1\|_\infty \leq 1 - \gamma) - \Pr(\|K_2\|_\infty > \frac{\gamma}{2}) \to 1, \text{ as } n \to \infty.
\] (32)

So,
\[
\Pr(W_1) \geq \Pr(A \cap \{\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}\} \text{ and (14) has a unique solution})
+ \Pr(A^c \cap \{\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}\} \text{ and (14) has a unique solution})
\overset{(2)}{=} \Pr(A^c \cap \{\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}\} \text{ and (14) has a unique solution})
+ \Pr(A \cap \{\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}\})
\to 1, \text{ as } n \to \infty,
\] (33)

where (2) is because when $A$ and $\|K_1 + K_2\|_\infty \leq 1 - \frac{\gamma}{2}$ hold, (14) always has a unique solution. If there exists another optimal solution to (14), say $\beta^*$. Let $\tilde{\beta}(\alpha) = \alpha \tilde{\beta} + (1 - \alpha) \beta^*, (0 < \alpha < 1)$. Convexity of (14) guarantees $\tilde{\beta}(\alpha)$ is also a solution to (14). By the optimality conditions and convexity, we have
\[
\|n^{-1}X_{\tilde{S}_c}^T(Y - X\tilde{\beta}(\alpha))\|_\infty \leq \alpha \|n^{-1}X_{\tilde{S}_c}^T(Y - X\tilde{\beta})\|_\infty + (1 - \alpha)\|n^{-1}X_{\tilde{S}_c}^T(Y - X\beta^*)\|_\infty,
\]
\[
< \alpha \lambda_n + (1 - \alpha) \lambda_n = \lambda_n,
\]

where we have used $\|n^{-1}X_{\tilde{S}_c}^T(Y - X\tilde{\beta})\|_\infty < \lambda_n$ and $\|n^{-1}X_{\tilde{S}_c}^T(Y - X\beta^*)\|_\infty \leq \lambda_n$. Therefore, $[\tilde{\beta}(\alpha)]_{\tilde{S}_c} = 0$. Then $\tilde{\beta}(\alpha)$ is also a solution to (15). The uniqueness of (15) leads to $\tilde{\beta} = \tilde{\beta}(\alpha)$, implying $\tilde{\beta} = \beta^*$. Hence the solution to (14) is also unique.
Step II. Plugging $Y = X_S \beta_S + \varepsilon$ into (19), we get,

$$
\|\beta - \hat{\beta}_S\|_{\infty} = \|\lambda_n(X_T^S X_S/n)^{-1}\text{sign}(\hat{\beta}_S) - (X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty}
\leq \lambda_n\|(X_T^S X_S/n)^{-1}\|_{\infty} + \|(X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty}
\leq \lambda_n s_n^{1/2}\|(X_T^S X_S/n)^{-1}\|_2 + \|(X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty}
\leq \lambda_n s_n^{1/2}(\|(X_T^S X_S/n)^{-1} - \Sigma_{SS}^{-1}\|_2 + 1/C_{\min}) + \|(X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty},
$$

Let $G = \left\{ \|(X_T^S X_S)^{-1}\|_2 > 9/(nC_{\min}) \right\}$, by the inequality (60) in Wainwright (2009),

$$
\text{pr}(G) \leq 2 \exp(-n/2).
$$

Since $(X_T^S X_S)^{-1}X_T^\varepsilon | X_S \sim N(0, (X_T^S X_S)^{-1})$, similarly we condition on $G$ to achieve,

$$
\text{pr}\left(\|(X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty} > \frac{s_n^{1/2}}{n^{1/2}C_{\min}^{1/2}}\right) \leq \text{pr}\left(\|(X_T^S X_S)^{-1}X_T^\varepsilon\|_{\infty} > \frac{s_n^{1/2}}{n^{1/2}C_{\min}^{1/2}} | G^c\right) + \text{pr}(G)
$$

$$
\leq 2s_n e^{-s_n/18} + 2e^{-n/2} \leq 2e^{-c_3 s_n},
$$

for some positive $c_3$. (27), (34), and (35) together imply that,

$$
\|\beta - \hat{\beta}_S\|_{\infty} \leq \lambda_n s_n^{1/2}\left(\frac{8}{C_{\min} s_n^{1/2}n^{-1/2}} + 1/C_{\min}\right) + \frac{s_n^{1/2}}{n^{1/2}C_{\min}^{1/2}}
$$

holds with probability larger than $1 - 2e^{-c_4 s_n}$ for a positive $c_4$. Under the scaling of Theorem 1 and Condition ??, it is easy to verify that

$$
\min_{j \in S} |\beta_j| > \lambda_n s_n^{1/2}\left(\frac{8}{C_{\min} s_n^{1/2}n^{-1/2}} + 1/C_{\min}\right) + \frac{s_n^{1/2}}{n^{1/2}C_{\min}^{1/2}},
$$

for sufficient large $n$. Thus,

$$
\text{pr}(W^c) = 1 - \text{pr}(W) \leq 1 - (1 - 2e^{-c_4 s_n}) \rightarrow 0, \quad \text{as } n \rightarrow \infty.
$$
Finally, (21), (33) and (37) together show that,

\[ \text{pr}(\hat{\beta} \text{ is unique and, } \text{sign}(\hat{\beta}) = \text{sign}(\beta)) \to 1, \quad \text{as } n \to \infty. \]

**Proof of Theorem 2**

Firstly, consider the second step,

\[ \hat{\beta} = \arg \min_{\beta} \left\{ \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda_n (\| \beta_{\hat{S}_1} \|_1 + \| \beta_{\hat{N}_1} \|_1) \right\}. \]  

(38)

We are going to show that with high probability,

\[ \hat{\beta}_{S_1} \neq 0 \quad \text{and} \quad \hat{\beta}_{N_1} = 0. \]  

(39)

Define an oracle estimator of (38),

\[ \bar{\beta} = \arg \min_{\beta_{N_1} = 0} \left\{ \frac{1}{2n} \| Y - X_{\hat{Q}}\beta_{\hat{Q}} \|_2^2 + \lambda_n \| \beta_{\bar{S}_1} \|_1 \right\}. \]  

(40)

where \( \hat{Q} = S \cup \hat{N}_2 \). Similar as in theorem 1 to show \( \hat{\beta}_{N_1} = 0 \), it is sufficient to prove,

\[ \| X_{\hat{Q}}^T X_{\hat{Q}}(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}\text{sign}(\hat{\beta}_{\hat{Q}}) + (n\lambda_n)^{-1}X_{\hat{Q}}^T (I - X_{\hat{Q}}(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T)(X_S \beta_S + \epsilon) \|_{\infty} < 1, \]  

(41)

and (38) has a unique solution. Since \( (I - X_{\hat{Q}}(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T)X_{\hat{Q}} = 0 \), (31) can be simplified as

\[ \| X_{\hat{Q}}^T X_{\hat{Q}}(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}\text{sign}(\hat{\beta}_{\hat{Q}}) + (n\lambda_n)^{-1}X_{\hat{Q}}^T (I - X_{\hat{Q}}(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T)\epsilon \|_{\infty} < 1. \]  

(42)
Let

\[ F = X^T - \Sigma_{\hat{Q}^*}^{-1}X^T, \]
\[ K_1 = \Sigma_{\hat{Q}^*}^{-1} \text{sig}(\hat{\beta}_{\hat{Q}}), \]
\[ K_2 = FX_{\hat{Q}}(X_{\hat{Q}}^TX_{\hat{Q}})^{-1}\text{sig}(\hat{\beta}_{\hat{Q}}) + (n\lambda_n)^{-1}F\{I - X_{\hat{Q}}(X_{\hat{Q}}^TX_{\hat{Q}})^{-1}X_{\hat{Q}}^T\} \varepsilon. \]

Then, (42) is equivalent to

\[ \|K_1 + K_2\|_\infty < 1. \]

Different from the proof in theorem 1, the subset \( \hat{Q} \) is random now. To this end, introduce

\[ A = \{R \subset \hat{S}_2 \subset S, S \subset \hat{Q} \subset S \cup Z\}, \]
\[ B = \{S \subset \hat{Q} \subset S \cup Z\}, \]
\[ C = \{N_2 \subset Z\}. \]

From proposition 2, It is not hard to show \( P(A) \to 1 \), under the scaling in theorem 2. Note that \( \text{sig}(\hat{\beta}_Q) \) only has \( \hat{S}_1 \) non-zero entries, hence

\[ \text{pr}(\|K_1\|_\infty \leq 1 - \gamma) \geq \text{pr}(\{\|K_1\|_\infty \leq 1 - \gamma\} \cap A) \]
\[ \stackrel{(a)}{=} \text{pr}(A), \]  \hspace{1cm} (43)

where \((a)\) holds because \( A \) and condition 12 imply \( \|K_1\|_\infty \leq 1 - \gamma \). To bound \( \|K_2 +\)
$K_3 \parallel_\infty$, let

$$H = \bigcup_{(Q, S_2) \subset Q \subset S \cup Z} \left\{ \text{sig}(\tilde{\beta}_Q)^T (X_Q^T X_Q)^{-1} \text{sig}(\tilde{\beta}_Q) + (n\lambda_n)^{-2} \|\epsilon\|^2 > \frac{s_n + z_n}{n C_{\min}} (s_n + z_n)^{1/2} n^{-1/2} + 1 \right\} + (1 + s_n^{1/2} n^{-1/2})/(n \lambda_n^2) \right\}. $$

Note that $\tilde{\beta}_Q$ is the analogy of $\hat{\beta}_Q$ by replacing $\hat{Q}$ and $\hat{S}_2$ in (40) with $Q$ and $S_2$. Then,

$$\text{pr}(\|K_2\|_\infty > \gamma \frac{3}{2}) \leq \text{pr}(\{\|K_2\|_\infty > \gamma \frac{3}{2}\} \cap A) + \text{pr}(A^c) \leq \text{pr}(\bigcup_{(Q, S_2) \subset Q \subset S \cup Z} \{\|K_2(Q, S_2)\|_\infty > \gamma \frac{3}{2}\} \cap A) + \text{pr}(A^c) \leq \text{pr}(\bigcup_{(Q, S_2) \subset Q \subset S \cup Z} \|K_2(Q, S_2)\|_\infty > \gamma \frac{3}{2} \mid H^c) + \text{pr}(H) + \text{pr}(A^c). \quad (44)$$

$$\text{pr}(H) \text{ can be bounded in the same way as in theorem 1,}$$

$$\text{pr}(H) \leq \text{pr}(\bigcup_{(Q, S_2) \subset Q \subset S \cup Z} \left\{ \text{sig}(\tilde{\beta}_Q)^T (X_Q^T X_Q)^{-1} \text{sig}(\tilde{\beta}_Q) > \frac{s_n + z_n}{n C_{\min}} (s_n + z_n)^{1/2} n^{-1/2} + 1 \right\}$$

$$+ \text{pr}(n \lambda_n)^{-2} \|\epsilon\|^2 > (1 + s_n^{1/2} n^{-1/2})/(n \lambda_n^2)) \leq \text{pr}(\bigcup_{s \subset Q \subset S \cup Z} \left\{ \|X_Q^T X_Q/n - \Sigma_Q^{-1}\|_2 \geq \frac{8}{C_{\min}} (s_n + z_n)^{1/2} n^{-1/2} \right\}) + e^{-\frac{3}{16} s_n} \leq \text{pr}(\bigcup_{s \subset Q \subset S \cup Z} \left\{ \|X_Q^T X_Q/n - \Sigma_Q^{-1}\|_2 \geq \frac{8}{C_{\min}} (\text{Card}(Q))^{1/2} n^{-1/2} \right\}) + e^{-\frac{3}{16} s_n} \leq 2z_n^1 \exp(-\frac{s_n}{2}) + \exp(-\frac{3}{16} s_n). \quad (45)$$
We use similar arguments as in theorem 1 for bounding the following,

\[
\begin{align*}
\Pr\left( \bigcup_{(Q,S) \in \mathcal{S} \subset \mathcal{Q} \subset \mathcal{S} \cup \mathcal{Z}} \| K_2(Q, S_2) \|_\infty > \frac{\gamma}{2} \mid H^c \right) & \leq 2^{s_n+1+z_n}(p_n - s_n) \exp(-\gamma^2/8V), \\
& \quad (46)
\end{align*}
\]

where \( V = \frac{s_n+z_n}{nC_{\min}} (8(s_n+z_n)^{1/2}n^{-1/2} + 1) + (1 + s_n^{1/2}n^{-1/2})/(n\lambda_n^2) \).

Under the scaling in theorem 2, (43)(44)(45)(46) show that (41) holds with high probability. The uniqueness of (38) can be proved by the same arguments as in theorem 1. We skip the proof here for simplicity. Next, we bound \( \| \bar{\beta}_Q - \beta_Q \|_\infty \).

\[
\begin{align*}
\| \bar{\beta}_Q - \beta_Q \|_\infty & = \| (X^T_Q X_Q)^{-1}(X^T_Q Y - n\lambda_n\text{sign}(\bar{\beta}_Q)) - \beta_Q \|_\infty \\
& \underset{(b)}{=} \| (X^T_Q X_Q)^{-1}X^T_Q \varepsilon - \lambda_n(X^T_Q X_Q/n)^{-1}\text{sign}(\bar{\beta}_Q) \|_\infty \\
& \leq \| (X^T_Q X_Q)^{-1}X^T_Q \varepsilon \|_\infty + \| \lambda_n(X^T_Q X_Q/n)^{-1} \|_\infty,
\end{align*}
\]

where (b) holds because \((X^T_Q X_Q)^{-1}X^T_Q X_S \beta_S - \beta_Q = 0\). Let \( U_n = \lambda_n(s_n+z_n)^{1/2}\left(\frac{8}{C_{\min}}(s_n+z_n)^{1/2} + \frac{1}{C_{\min}} \right) + \frac{(s_n+z_n)^{1/2}}{n^{1/2}C_{\min}} \). Then,

\[
\begin{align*}
\Pr(\| \bar{\beta}_Q - \beta_Q \|_\infty \geq U_n) & \leq \Pr(\{ \| \bar{\beta}_Q - \beta_Q \|_\infty \geq U_n \} \cap B) + \Pr(B^c) \\
& \leq \Pr(\bigcup_{S \subset \mathcal{Q} \subset \mathcal{S} \cup \mathcal{Z}} \| (X^T_Q X_Q)^{-1}X^T_Q \varepsilon \|_\infty + \| \lambda_n(X^T_Q X_Q/n)^{-1} \|_\infty \geq U_n) + \Pr(B^c) \\
& \underset{(c)}{=} 2^{s_n}(2s_ne^{-\frac{n}{n^{1/2}}} + 2e^{-n/2} + 2e^{-\frac{n}{n^{1/2}}}) + \Pr(B^c),
\end{align*}
\]

where (c) follows from the bounds \( (34) \) and \( (35) \) in the proof of theorem 1. By condition 10, it is not hard to verify,
\[
\min_{j \in S} |\beta_j| \gg U_n,
\]

Thus,

\[
\Pr(\min_{j \in S} |\beta_j| > \|\hat{\beta}_{S_1} - \tilde{\beta}_{S_1}\|_\infty) \geq \Pr(B) - 2^{2n}(2s_n e^{-\frac{\lambda_n}{n}} + 2e^{-n/2} + 2e^{-\frac{\lambda_n}{n}}), \tag{48}
\]

Since \(P(B) \geq P(A) \to 1\) and \(2^{2n}(2s_n e^{-\frac{\lambda_n}{n}} + 2e^{-n/2} + 2e^{-\frac{\lambda_n}{n}}) \to 0\) under the scaling in theorem, \(48\) implies that \(\tilde{\beta}_{S_1} \neq 0\) with high probability.

Let us now consider the third step. We have shown that, with high probability, the third step takes the form

\[
\tilde{\beta} = \arg \min_{\hat{\beta}_{S_1}, \hat{\beta}_{N_2}} \left\{ \frac{1}{2n} \|Y - X_S \hat{\beta}_Q\|_2^2 + \lambda_n^*(\|\beta_{S_2}\|_1 + \|\beta_{N_2}\|_1) \right\}. \tag{49}
\]

To prove \(\text{sign}(\tilde{\beta}) = \text{sign}(\beta)\), it remains to show \(\text{sign}(\tilde{\beta}_S) = \text{sign}(\beta_S)\) and \(\tilde{\beta}_{N_2} = 0\). We use similar arguments as in the second step. Define the oracle estimator of \(49\),

\[
\hat{\beta} = \arg \min_{\hat{\beta}_{S_1}, \hat{\beta}_{N_2}} \left\{ \frac{1}{2n} \|Y - X_S \hat{\beta}_S\|_2^2 + \lambda_n^*(\|\beta_{S_2}\|_1 + \|\beta_{N_2}\|_1) \right\}. \tag{50}
\]

Let

\[
\tilde{F} = X_{\hat{N}_2}^T - \Sigma_{\hat{N}_2S}^{-1}X_S^T,
\]

\[
\tilde{K}_1 = \Sigma_{\hat{N}_2S}^{-1}X_S, \quad \tilde{K}_2 = \tilde{F}X_S(X_S^T X_S)^{-1}\text{sign}(\tilde{\beta}_S) + (n\lambda_n^*)^{-1}\tilde{F}\{I - X_S(X_S^T X_S)^{-1}X_S^T\} \epsilon.
\]

Then,
\[ P(\|\tilde{K}_1\|_\infty \leq 1 - \alpha) \geq P(\{\|\tilde{K}_1\|_\infty \leq 1 - \alpha\} \cap C) \]
\[ \geq P(C), \quad (51) \]

where \((d)\) holds because \(C\) and condition 13 imply \(\|\tilde{K}_1\|_\infty \leq 1 - \alpha\). Let

\[ \tilde{H} = \bigcup_{R \subseteq S_2 \subset S} \left\{ \text{sig}(\tilde{\beta}_S)^T (X^T_S X_S)^{-1} \text{sig}(\tilde{\beta}_S) + (n\lambda^*_n)^{-2} \|\varepsilon\|_2^2 \geq \frac{s_n}{nC_{\min}} (8s_{n}^{1/2}n^{-1/2} + 1) \right\} \]

Then,

\[ P(\|\tilde{K}_2\|_\infty > \frac{\alpha}{2}) \leq P(\{\|\tilde{K}_2\|_\infty > \frac{\alpha}{2}\} \cap A) + P(A^c) \]
\[ \leq P\left( \bigcup_{(N_2,S_2)} \left\{ \|\tilde{K}_2(N_2,S_2)\|_\infty > \frac{\alpha}{2} \right\} \right) + P(A^c) \]
\[ \leq P\left( \bigcup_{(N_2,S_2)} \left\{ \|\tilde{K}_2(N_2,S_2)\|_\infty > \frac{\alpha}{2} \right\} \right) P(\tilde{H}) + P(A^c) \]
\[ \leq 2^{s_n+s_n+1} z_n e^{-\alpha^2/8\tilde{V}} + 2e^{-\frac{s_n}{2}} + e^{-\frac{3}{8} s_n} + P(A^c), \quad (52) \]

where \((e)\) follows from \((27)\) and \((28)\) in the proof of theorem 1 and \(\tilde{V} = \frac{s_n}{nC_{\min}} (8s_{n}^{1/2}n^{-1/2} + 1) + (1 + s_{n}^{1/2}n^{-1/2})/(n(\lambda^*_n)^2)\). Again, we skip the proof of uniqueness of \((49)\). Now we have shown that \(\tilde{\beta}_{N_2} = 0\) with high probability. The final step is to bound \(\|\hat{\beta}_S - \beta_S\|_\infty\).

\[ \|\hat{\beta}_S - \beta_S\|_\infty = \| (X^T_S X_S)^{-1} (X^T_S Y - n\lambda^*_n \text{sig}(\hat{\beta}_S)) - \beta_S\|_\infty \]
\[ \leq \| (X^T_S X_S)^{-1} X^T_S \varepsilon\|_\infty + \|\lambda^*_n (X^T_S X_S/n)^{-1}\|_\infty. \]
Let $W_n = \lambda_n s_n^{1/2} \left( \frac{1}{\sigma_{\min}} s_n^{1/2} n^{-1/2} \frac{4}{\sqrt{n}} + \frac{1}{\sigma_{\min}} \right) + \frac{s_n^{1/2}}{n^{1/2} \sigma_{\min}}$. By (34) and (35) in the proof of theorem 1, we have

$$P(\|\hat{\beta}_S - \beta_S\|_\infty \leq W_n) \geq 1 - 2 \exp(-c_2 s_n),$$

for a positive $c_2$. Since $U_n \approx W_n$,

$$P(\min_{j \in S} |\beta_j| > \|\hat{\beta}_S - \beta_S\|_\infty) \geq 1 - 2 \exp(-c_2 s_n),$$

(53)

for sufficiently large $n$. Putting (51) (52) (53) together, we have shown

$$P(\tilde{\beta} \text{ is unique and } \text{sign}(\tilde{\beta}) = \text{sign}(\beta)) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

$\square$

References

AKAIKE, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19 716–723.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 1705–1732.

CHIANG, A. P., BECK, J. S., YEN, H.-J., TAYEH, M. K., SCHEETZ, T. E., SWIDERSKI, R., NISHIMURA, D., BRAUN, T. A., KIM, K.-Y., HUANG, J., ELBEDOUR, K., CARMI, R., SLUSARSKI, D. C., CASAVANT, T. L., STONE, E. M. and SHEFFIELD, V. C. (2006). Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardetcbiedl syndrome gene (bbs11). PNAS, 103 6287–6292.
FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106** 544–557.

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.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70** 849–911.

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, **10** 2013–2038.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, **38** 3567–3604.

FAN, Y. and TANG, C. Y. (2012). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75** 531–552.

FRANK, L. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35** 109–135.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1–22. URL [http://www.jstatsoft.org/v33/i01/](http://www.jstatsoft.org/v33/i01/)

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10** 971–988.
HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica, 18* 1603.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4* 249–264.

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics, 28* 1356–1378.

SCHEETZ, T. E., KIM, K.-Y. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A., KNUDTSON, K. L., DORRANCE, A. M., DiBONA, G. F., HUANG, J., CASAVANT, T. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *PNAS, 103* 14429–14434.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6* 461–464.

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

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery. *Information Theory, IEEE Transactions on, 55* 2183–2202.

WANG, H., LI, B. and LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71* 671–683.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38* 894–942.
Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research, 7* 2541–2563.

Zhou, S., van de Geer, S. and Bühlmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67* 301–320.