Regularization After Marginal Learning for Ultra-High Dimensional Regression Models

Yang Feng and Mengjia Yu

Abstract Regularization is a popular variable selection technique for high dimensional regression models. However, under the ultra-high dimensional setting, a direct application of the regularization methods tends to fail in terms of model selection consistency due to the possible spurious correlations among predictors. Motivated by the ideas of screening (Fan and Lv, J R Stat Soc Ser B Stat Methodol 70:849–911, 2008) and retention (Weng et al, Manuscript, 2013), we propose a new two-step framework for variable selection, where in the first step, marginal learning techniques are utilized to partition variables into different categories, and the regularization methods can be applied afterwards. The technical conditions of model selection consistency for this broad framework relax those for the one-step regularization methods. Extensive simulations show the competitive performance of the new method.

Keywords Independence screening • Lasso • Marginal learning • Retention • Selection • Sign consistency

1 Introduction

With the booming of information and vast improvement for computation speed, we are able to collect large amount of data in terms of a large collections of n observations and p predictors, where $p \gg n$. Recently, model selection gains increasing attention especially for ultra-high dimensional regression problems. Theoretically, the accuracy and interpretability of selected model are crucial in variable selection. Practically, algorithm feasibility and efficiency are vital in applications.

A great variety of penalized methods have been proposed in recent years. The regularization techniques for simultaneous variable selection and estimation are particularly useful to obtain sparse models compared to simply apply traditional criteria such as Akaike's information criterion [1] and Bayesian information

Y. Feng (✉) • M. Yu
Department of Statistics, Columbia University, New York, NY 10027, USA
e-mail: yangfeng@stat.columbia.edu

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_1

criterion [18]. The least absolute shrinkage and selection operator (Lasso) [19] have been widely used as the l_1 penalty shrinks most coefficients to 0 and fulfills the task of variable selection. Many other regularization methods have been developed; including bridge regression [13], the smoothly clipped absolute deviation method [5], the elastic net [26], adaptive Lasso [25], LAMP [11], among others. Asymptotic analysis for the sign consistency in model selection [20, 24] has been introduced to provide theoretical support for various methods. Some other results such as parameter estimation [17], prediction [15], and oracle properties [5] have been introduced under different model contexts.

However, in ultra-high dimensional space where the dimension $p = \exp(n^a)$ (where $a > 0$), the conditions for sign consistency are easily violated as a consequence of large correlations among variables. To deal with such challenges, Fan and Lv [6] proposed the sure independence screening (SIS) method which is based on correlation learning to screen out irrelevant variables efficiently. Further analysis and generalization can be found in Fan and Song [7] and Fan et al. [8]. From the idea of retaining important variables rather than screening out irrelevant variables, Weng et al. [21] proposed the regularization after retention (RAR) method. The major differences between SIS and RAR can be summarized as follows. SIS makes use of marginal correlations between variables and response to screen noises out, while RAR tries to retain signals after acquiring these coefficients. Both of them relax the irrepresentable-type conditions [20] and achieve sign consistency.

In this paper, we would like to introduce a general multi-step estimation framework that integrates the idea of screening and retention in the first step to learn the importance of the features using the marginal information during the first step, and then impose regularization using corresponding weights. The main contribution of the paper is two-fold. First, the new framework is able to utilize the marginal information adaptively in two different directions, which will relax the conditions for sign consistency. Second, the idea of the framework is very general and covers the one-step regularization methods, the regularization after screening method, and the regularization after retention method as special cases.

The rest of this paper is organized as follows. In Sect. 2, we introduce the model setup and the relevant techniques. The new variable selection framework is elaborated in Sect. 3 with connections to existing methods explained. Section 4 develops the sign consistency result for the proposed estimators. Extensive simulations are conducted in Sect. 5 to compare the performance of the new method with the existing approaches. We conclude with a short discussion in Sect. 6. All the technical proofs are relegated to the appendix.

2 Model Setup and Several Methods in Variable Selection

2.1 Model Setup and Notations

Let (X_i, Y_i) be i.i.d. random pairs following the linear regression model:

$$Y_i = X_i\beta + \varepsilon_i, \quad i = 1, \dots, n,$$

where $X_i = (X_i^1, \dots, X_i^p)^T$ is p_n-dimensional vector distributed as $N(0, \Sigma)$, $\beta = (\beta_1, \dots, \beta_p)^T$ is the true coefficient vector, $\varepsilon_1, \dots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$, and $\{X_i\}_{i=1}^n$ are independent of $\{\varepsilon_i\}_{i=1}^n$. Note here, we sometimes use p_n to emphasize the dimension p is diverging with the sample size n. Denote the support index set of β by $S = \{j : \beta_j \neq 0\}$ and the cardinality of S by s_n, and $\Sigma_{S^c|S} = \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}$. Both p_n and s_n are allowed to increase as n increases. For conciseness, we sometimes use signals and noises to represent relevant predictors S and irrelevant predictors S^c (or their corresponding coefficients) respectively.

For any set A, let A^c be its complement set. For any k dimensional vector w and any subset $K \subseteq \{1, \dots, k\}$, w_K denotes the subvector of w indexed by K, and let $\|w\|_1 = \sum_{i=1}^k |w_i|$, $\|w\|_2 = (\sum_{i=1}^k w_i^2)^{1/2}$, $\|w\|_\infty = \max_{i=1,\dots,k} |w_i|$. For any $k_1 \times k_2$ matrix M, any subsets $K_1 \subseteq \{1, \dots, k_1\}$, $K_2 \subseteq \{1, \dots, k_2\}$, $M_{K_1K_2}$ 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 $\|M\|_2 = \{\Lambda_{\max}(M^TM)\}^{1/2}$ and $\|M\|_\infty = \max_{i=1,\dots,k} \sum_{j=1}^k |M_{ij}|$. When $k_1 = k_2 = k$, let $\rho(M) = \max_{i=1,\dots,k} M_{ii}$, $\Lambda_{\min}(M)$ and $\Lambda_{\max}(M)$ be the minimum and maximum eigenvalues of M, respectively.

2.2 Regularization Techniques

The Lasso [19] defined as

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

is a popular variable selection method. Thanks to the invention of efficient algorithms including LARS [4] and the coordinate descent algorithm [14], Lasso and its variants are applied to a wide range of different scenarios in this big data era. There is a large amount of research related to the theoretical properties of Lasso. Zhao and Yu [24] proposed almost necessary and sufficient conditions for the sign consistency for Lasso to select true model in the large p_n setting as n increases. Considering the sensitivity of tuning parameter λ_n and consistency for model selection, Wainwright

[20] has identified precise conditions of achieving sparsity recovery with a family of regularization parameters λ_n under deterministic design.

Another effective approach to the penalization problem is adaptive Lasso (AdaLasso) [25], which uses an adaptively weighted l_1-penalty term, defined as

$$\hat{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p_n} \omega_j |\beta_j| \right\}, \quad \lambda_n \geq 0. \tag{2}$$

where $\omega_j = 1/|\hat{\beta}_{\text{init}}|^{\gamma}$ for some $\gamma \geq 0$, in which $\hat{\beta}_{\text{init}}$ is some initial estimator. When signals are weakly correlated to noises, Huang et al. [16] proved AdaLasso is sign consistent with $\omega_j = 1/|\hat{\beta}_j^M| \equiv 1/|(\tilde{X^j})^T Y|$, where $\tilde{X}$ is the centered and scaled data matrix. One potential issue of this weighting choice is that when the correlations between some signals and response are too small, those signals would be severely penalized and may be estimated as noises. We will use numeric examples to demonstrate this point in the simulation section.

2.3 *Sure Independence Screening*

To reduce dimension from ultra-high to a moderate level, Fan and Lv [6] proposed a sure independence screening (SIS) method, which makes use of marginal correlations as a measure of importance in first step and then utilizes other operators such as Lasso to fulfill the target of variable selection. In particular, first we calculate the component-wise regression coefficients for each variable, i.e., $\hat{\beta}_j^M = (\tilde{X^j})^T \tilde{Y}$, $j = 1, \ldots, p_n$, where $\tilde{X^j}$ is the standardized j-th column of data X and $\tilde{Y}$ is the standardized response. Second, we define a sub-model with respect to the largest coefficients

$$\mathcal{M}_\gamma = \{1 \leq j \leq p_n : |\hat{\beta}_j^M| \text{ is among the first } \lfloor \gamma n \rfloor \text{ of all}\}.$$

Predictors that are not in $\mathcal{M}_\gamma$ are regarded as noise and therefore discarded for further analysis. SIS reduces the number of candidate covariates to a moderate level for the subsequent analysis. Combining SIS and Lasso, Fan and Lv [6] introduced SIS-Lasso estimator,

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta \in \mathcal{M}_\gamma} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda_n \sum_{j \in \mathcal{M}_\gamma} |\beta_j| \right\} \\ &= \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda_n \sum_{j \in \mathcal{M}_\gamma} |\beta_j| + \infty \sum_{j \in \mathcal{M}_\gamma^c} |\beta_j| \right\}. \end{aligned} \tag{3}$$

Clearly, γ should be chosen carefully to avoid screening out signals. To deal with the issue that signals may be marginally uncorrelated with the response in some cases, iterative-SIS was introduced [6] as a practical procedure but without rigorous theoretical support for the sign consistency. As a result, solely relying on marginal information is sometimes a bit too risky, or greedy, for model selection purpose.

3 Regularization After Marginal Learning

3.1 Algorithm

From Sect. 2, one potential drawback shared between AdaLasso and SIS-Lasso is that they may miss important covariates that are marginally weakly correlated with the response.

Now, we introduce a new algorithm, regularization after marginal (RAM) learning, to solve the issue. It utilizes marginal correlation to divide all variables into three candidate sets: a retention set, a noise set, and an undetermined set. Then regularization is imposed to find signals in the uncertainty set as well as to identify falsely retention signals and falsely screened noises.

A detailed description of the algorithm is as follows:

Step 0 (Marginal Learning) *Calculate the marginal regression coefficients after standardizing each predictor, i.e.,*

$$\hat{\beta}_j^{\mathcal{M}} = \sum_{i=1}^{n} \frac{(X_i^j - \bar{X}^j)}{\hat{\sigma}_j} Y_i, \quad 1 \le j \le p_n, \tag{4}$$

where $\bar{X}^j = \frac{1}{n}\sum_{i=1}^n X_i^j$ *and* $\hat{\sigma}_j^2 = \sqrt{\frac{\sum_{i=1}^n (X_i^j - \bar{X}^j)^2}{n-1}}$.

Define a retention set by $\hat{\mathcal{R}} = \{1 \le j \le p : |\hat{\beta}_j^{\mathcal{M}}| \ge \gamma_n\}$, *for a positive constant* γ_n; *a noise set by* $\hat{\mathcal{N}} = \{1 \le j \le p : |\hat{\beta}_j^{\mathcal{M}}| \le \tilde{\gamma}_n\}$, *for a positive constant* $\tilde{\gamma}_n < \gamma_n$; *and an undetermined set by* $\hat{\mathcal{U}} = (\hat{\mathcal{R}} \cup \hat{\mathcal{N}})^c$.

Step 1 (Regularization After Screening Noises Out) *Search for signals in* $\hat{\mathcal{U}}$ *by solving*

$$\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1} = \underset{\beta_{\hat{\mathcal{N}}}=0}{\arg\min} \left\{ (2n)^{-1} \sum_{i=1}^{n} \Big(Y_i - \sum_{j \in \hat{\mathcal{U}}} X_{ij}\beta_j - \sum_{k \in \hat{\mathcal{R}}} X_{ik}\beta_k \Big)^2 + \lambda_n \sum_{j \in \hat{\mathcal{U}}} |\beta_j| \right\}, \tag{5}$$

where the index $\hat{\mathcal{U}}_1$ *is denoted as the set of variables that are estimated as signals in* $\hat{\mathcal{U}}$, *namely* $\hat{\mathcal{U}}_1 = \{j \in \hat{\mathcal{U}} | (\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1})_j \ne 0\}$. *After Step 1, the selected variable set is* $\hat{\mathcal{R}} \cup \hat{\mathcal{U}}_1$.

Step 2 (Retrieve Falsely Discarded Signals) *Reevaluate the set $\hat{\mathcal{N}}$ to check whether it contains any signals. Solve*

$$\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} = \underset{\beta_{\hat{\mathcal{U}}_2}=0}{\arg\min} \left\{ (2n)^{-1} \sum_{i=1}^{n} \Big(Y_i - \sum_{j\in\hat{\mathcal{N}}} X_{ij}\beta_j - \sum_{k\in\hat{\mathcal{R}}\cup\hat{\mathcal{U}}_1} X_{ik}\beta_k \Big)^2 + \lambda_n^{\star} \sum_{j\in\hat{\mathcal{N}}} |\beta_j| \right\}, \tag{6}$$

where $\hat{\mathcal{U}}_2 = \hat{\mathcal{U}}\backslash\hat{\mathcal{U}}_1$.

This step is used to retrieve important variables which are weakly correlated to response marginally. This step can be omitted if we are sure about the noise set $\hat{\mathcal{N}}$. The selected variable set is now $\hat{\mathcal{R}} \cup \hat{\mathcal{U}}_1 \cup \hat{\mathcal{N}}_1$.

Step 3 (Remove Falsely Retained Signals) *Inspect the retention set $\hat{\mathcal{R}}$ to check whether it contains any noises. Solve*

$$\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} = \underset{\beta_{\hat{\mathcal{U}}_2\cup\hat{\mathcal{N}}_2}=0}{\arg\min} \left\{ (2n)^{-1} \sum_{i=1}^{n} \Big(Y_i - \sum_{j\in\hat{\mathcal{R}}} X_{ij}\beta_j - \sum_{k\in\hat{\mathcal{U}}_1\cup\hat{\mathcal{N}}_1} X_{ik}\beta_k \Big)^2 + \lambda_n^{\star\star} \sum_{j\in\hat{\mathcal{R}}} |\beta_j| \right\}, \tag{7}$$

where $\hat{\mathcal{N}}_2 = \hat{\mathcal{N}}\backslash\hat{\mathcal{N}}_1$.

This step is used to remove noises which are highly correlated with the response marginally. This step can be omitted if we are sure about the retention set $\hat{\mathcal{R}}$. The final selected variable set is $\hat{\mathcal{R}}_1 \cup \hat{\mathcal{U}}_1 \cup \hat{\mathcal{N}}_1$.

The final estimator $\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}$ is called the regularization after marginal (RAM) learning estimator. Note that the optimization problem described Step 2 in the RAM algorithm is of the same complexity as the original Lasso problem. A more efficient version of the algorithm where we remove Step 2 is called RAM-2. The corresponding selected variable set of RAM-2 is $\hat{\mathcal{R}}_1 \cup \hat{\mathcal{U}}_1$ as $\hat{\mathcal{N}}_1 = \emptyset$.

3.2 Connections to SIS and RAR

In the preparation Step 0 of RAM, marginal correlation provides us with a first evaluation of the importance for all variables. Usually, we expect that the variables with high marginal correlations are likely to be signals, while noises tend to have low marginal correlations. The choice of the thresholds γ_n and $\tilde{\gamma}_n$ are critical to ensure the accuracy of the retention set and the noise set. We follow Weng et al. [21] to select γ_n using a permutation-based approach. In particular, denote $Y_{(1)}, \ldots, Y_{(n)}$ as randomly permuted responses. Let γ_n be the largest marginal regression coefficient between permuted response and original data, i.e.,

$$\gamma_n = \max_{1\le j\le p_n} \left\{ |D_j| \Big| D_j = \sum_{i=1}^{n} \frac{(X_i^j - \bar{X}^j)}{\hat{\sigma}_j} Y_{(i)} \right\}. \tag{8}$$

In practice, we may adjust the threshold to ensure at most $\lceil n^{1/2} \rceil$ variables are included in the retention set, considering the root n consistency of classical least square estimators as well as SIS-based models. For $\tilde{\gamma}_n$, we can set it as the n-th largest coefficient in magnitude so that the cardinality of $\hat{\mathcal{R}} \cup \hat{\mathcal{U}}$ is $n - 1$.

RAM-2 is closely connected to SIS. Technically, it utilizes marginal information to remove as many noises as possible. In addition, RAM-2 can be viewed as a greedy implementation of RAR+ [21], which is summarized in the following.

- (Retention) Define a retention set $\hat{R}$ which represents the coefficients strongly correlated to response marginally.
- (Regularization) Apply penalization on $\hat{R}^c$ to recover signals

$$\check{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + 0 \sum_{j \in \hat{R}} |\beta_j| + \lambda_n \sum_{j \in \hat{R}^c} |\beta_j| \right\}. \tag{9}$$

- (Redemption) Denote $Q = \{j \in \hat{R}^c : \check{\beta}_j \neq 0\}$, additional signals detected from the second step. Calculate the following penalized least square problem:

$$\tilde{\beta} = \arg\min_{\beta_{(\hat{R} \cup Q)^c} = 0} \left\{ (2n)^{-1} \sum_{i=1}^{n} \left(Y_i - \sum_{j \in \hat{R}} X_{ij} \beta_j - \sum_{k \in Q} X_{ik} \beta_k \right)^2 + \lambda_n^* \sum_{j \in \hat{R}} |\beta_j| \right\}, \tag{10}$$

  where $\lambda_n^\star$ is the penalty parameter and is in general different from λ_n in the previous step.

The regularization step only imposes penalty to variables that are not in $\hat{R}$. When all covariates in $\hat{R}$ are signals, we need only to recover the sparsity in $\hat{R}^c$. Although RAR performs well when the retention set $\hat{R} \subseteq S$, it could fail to recover the true sparsity pattern when $\hat{R}$ contains noises. Hence, the redemption step is necessary to rule out falsely selected noises.

As the intrinsic idea for RAR is retention, RAR+ can be regarded as a bidirectional and self-corrected version of RAR. Motivated by SIS-Lasso (3) and RAR+ (10), RAM first explores data by dividing variables into three sets in which one contains signal-like variables, one contains noise-like variables, and one contains the remaining undetermined variables. In Steps 1 and 3, RAM-2 combines advantages of SIS and RAR: on one hand, in terms of computational efficiency, like SIS, it is very efficient, thanks to the many noises screened out in the first step; on the other hand, RAM-2 could relax the regularity condition for sign consistency due to the retention set.

3.3 From RAM-2 to RAM

Though RAM-2 takes advantages of both SIS and RAR+, it shares the same drawback as SIS since signals that are marginally uncorrelated with the response could be removed during Step 1. To avoid fully replying on marginal correlation, RAM adds Step 2 to recover such signals.

Instead of re-examining $\hat{\mathcal{R}}$ immediately, the optional Step 2 is designed to reexamine the "noise" set $\hat{\mathcal{N}}$ and find signals in it. Intuitively, the retention of signals in $\hat{\mathcal{U}}_1 \cup \hat{\mathcal{R}}$ gives "weak" signals in $\hat{\mathcal{N}}_1$ an opportunity to show their significance in regression. Furthermore, noises in $\hat{\mathcal{R}}$ will also be weakly correlated with the residues $Y - X_{\mathcal{U}_1 \cup \mathcal{N}_1} \beta_{\mathcal{U}_1 \cup \mathcal{N}_1}$ in Step 3. Thus, we do not start to eliminate unnecessary variables in $\hat{\mathcal{R}}$ until all the other signals have been identified. Step 2 in RAM reduces the risk of signal losses, and increases the reliability of the model selection process.

We provide a brief comparison in Table 1 to show the similarities as well as differences among SIS-Lasso, RAR/RAR+, and RAM-2/RAM. The last row of Table 1 shows the final variable selection result. Note that, though some of the notations for different methods are same in Table 1, they are not necessarily identical since different procedures may lead to different results. Among these methods, RAM-2 and SIS-Lasso remove the variables in the noise set detected via marginal learning; RAR retains all variables in $\mathcal{R}$; RAM and RAR+ perform a recheck on all the candidate sets.

Table 1 Differences among 5 regularization methods using marginal information

RAM-2	RAM	SIS-Lasso	RAR	RAR+
$\mathcal{R}$: Retention set		$\mathcal{N}^c$: Candidates	$\mathcal{R}$: Retention set	
$\mathcal{U}$: Undetermined set			$\mathcal{R}^c$: Candidates	
$\mathcal{N}$: Noise set		$\mathcal{N}$: Noise set		
Retain $\mathcal{R}$	Retain $\mathcal{R}$	Check $\mathcal{N}^c$		
Check $\mathcal{U}$	Check $\mathcal{U}$	Remove $\mathcal{N}$		
Remove $\mathcal{N}$				
	Retain $\mathcal{R} \cup \mathcal{U}_1$		Retain $\mathcal{R}$	
	Check $\mathcal{N}$		Check $\mathcal{R}^c$	
Retain $\mathcal{U}_1 \cup \mathcal{N}_1$				Retain $(\mathcal{R}^c)_1$
Check $\mathcal{R}$				Check $\mathcal{R}$
$\mathcal{R}_1 \cup \mathcal{U}_1$	$\mathcal{R}_1 \cup \mathcal{U}_1 \cup \mathcal{N}_1$	$(\mathcal{N}^c)_1$	$\mathcal{R} \cup (\mathcal{R}^c)_1$	$\mathcal{R}_1 \cup (\mathcal{R}^c)_1$

The subscript 1 for each set denotes the signals recovered from the corresponding sets

4 Asymptotic Analysis

4.1 *Sure Independence Screening Property*

Considering the linear regression model under the scaling $\log p_n = O(n^{a_1})$, $s_n = O(n^{a_2}), a_1 > 0, a_2 > 0, a_1 + 2a_2 < 1$, which is also required for achieving Strong Irrepresentable Condition in Zhao and Yu [24]. Under the conditions below, Fan and Lv [6] showed that SIS asymptotically achieves to screen only noises out. This result is necessary for the consistency in SIS-Lasso (3) as well as in RAM-2.

Condition 1 $var(Y_1) = \beta_S^T \Sigma_{SS} \beta_S = O(1)$.

Condition 2 $\Lambda_{max}(\Sigma) \leq Cn^{\tau}$ *for a sufficiently large* $C, \tau \geq 0$.

Condition 3 $min_{j \in S} |cov(\beta_j^{-1} Y_1, X_1^j)| \geq c$ *for some positive constant* c.

Corollary 1 *Under Conditions 1–3, if* $min_{j \in S} |\beta_j| \geq Cn^{-a_3/2}$ *for some* $a_3 \in (0, 1 - a_1)$, *then there exists some* $\theta < 1 - a_1 - a_3$ *such that when* $\gamma \asymp n^{-\theta}$, *we have*

$$pr(S \subset \mathcal{M}_{\gamma}) \to 1 \quad as\ n \to \infty.$$

Condition 1 implies that there cannot be too many variables that have marginal regression coefficients exceeding certain thresholding level as in Fan and Song [7]. When Condition 2 fails, there is heavy collinearity in X, which leads to difficulty for differentiating signals from linearly correlated noises. Condition 3 rules out the situation that signals are jointly correlated with Y but their marginal correlations are relatively weak.

4.2 *Sign Consistency for RAM-2*

Given the success of screening in the first step, the following conditions are necessary to achieve sign consistency for RAM-2.

Condition 4 $\|\Sigma\beta\|_\infty = O(n^{(1-2\kappa)/8})$, *where* $0 < \kappa < \frac{1}{2}$ *is a constant.*

Condition 5 $min_{j \in S} |\beta_j| \geq Cn^{-\delta + a_2/2}$ *for a sufficiently large* C, *where* $0 < \delta < \{1 - max(a_1, a_2)\}/2$.

Condition 6 $\Lambda_{min}(\Sigma_{S \cup Z, S \cup Z}) \geq C_{min} > 0$, *where the strong noise set is defined as* $Z = \{j \in S^c : |\beta_j^M| \geq \gamma_n - c_1 n^{-\kappa}\}$ *with cardinality* z_n.

Condition 7 $\|\Sigma_{ZS} \Sigma_{SS}^{-1}\|_\infty \leq 1 - \alpha$, *where* $\alpha > 0$.

Condition 8 $max_{S \subset Q \subset S \cup Z} \|\{\Sigma_{Q^c Q} (\Sigma_{QQ})^{-1}\}_{S \cap R^c}\|_\infty \leq 1 - \gamma_1$, *where the strong signal set is defined as* $R = \{j \in S : |\beta_j^{\mathcal{M}}| > \gamma_n + c_1 n^{-\kappa}\}$ *and* $\gamma_1 > 0$.

Theorem 1 *If Conditions 1–8 are satisfied and $\gamma = \tilde{\gamma}_n$ holds for Corollary 1, then when $z_n/s_n \to 0, s_n \to \infty$ and $\lambda_n \asymp n^{-\delta}, \lambda_n^{**} \asymp n^{-\delta}$, RAM-2 achieves sign consistency*

$$pr(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} \textit{ is unique and } sign(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}) = sign(\beta)) \to 1, \textit{ as } n \to \infty.$$

Under the scaling conditions described in Theorem 1, Conditions 1 and 4 are required for establishing the uniform deviation results for marginal regression coefficients. Condition 5, which is a similar condition as that in Corollary 1, imposed a lower bound for magnitudes of the marginal regression coefficients. When strong noises in Z are not highly correlated to the signals, the probability of sign consistency converges to 1 as $n \to \infty$. In fact, when Z is an empty set, Conditions 6–8 are generalizations of some key conditions in Wainwright [20]. They relax the irrepresentable condition in Zhao and Yu [24] and give a toleration level on Z.

4.3 Sign Consistency for RAM

The key point for achieving sign consistency is the restriction for $\hat{\mathcal{U}}_1$ in Condition 8. In Step 2, we require similar restrictions on $\hat{\mathcal{N}}_1$ to guarantee the sign consistency of RAM. Different with RAM-2, we will take a second look on $\hat{\mathcal{N}}$ so that the success of RAM does not heavily depend on the screening step. We still control the scale as $\log p_n = O(n^{a_1}), s_n = O(n^{a_2}), a_1 > 0, a_2 > 0, a_1 + 2a_2 < 1$.

Theorem 2 *Under Conditions 4–8, when $z_n/s_n \to 0, s_n \to \infty$ and $\lambda_n, \lambda_n^*, \lambda_n^{**} \asymp n^{-\delta}$, RAM achieves sign consistency*

$$pr(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} \textit{ is unique and } sign(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}) = sign(\beta)) \to 1, \textit{ as } n \to \infty.$$

5 Numerical Study

5.1 Tuning Parameter Selection

In Weng et al. [21], the reports of successes are with respect to the oracle performance, namely the existence of an estimator that recovers the true model on the solution path. When comes to practice, it is necessary to choose an effective criterion for assessment of models under different tuning parameters λ_n. Chen and Chen [3] proposed an extended Bayesian information criterion (EBIC),

$$\mathrm{BIC}_\gamma = \mathrm{BIC} + 2\gamma \log \tau(\mathcal{S}_k), \quad 0 \le \gamma \le 1, \tag{11}$$

where $\mathcal{S}_k$ is the collection of all models with k covariates, and $\tau(\mathcal{S}_k)$ is the size of $\mathcal{S}_k$. Clearly, in our linear model, $\tau(\mathcal{S}_k) = \binom{p_n}{k}$. EBIC (BIC$_\gamma$) usually leads to a model with smaller size than BIC, since the additional term penalizes heavily on the model size. Therefore it is suitable for the ultra-high dimensional scenario we are considering. Chen and Chen [3] also established EBIC's consistency property. For all the penalize solution path calculation in the numeric studies, we apply EBIC for choosing the penalty parameter. Note that beside using a criterion function to select tuning parameter, another popular way is to use cross-validation-based approaches including Friedman et al. [14], Feng and Yu [12], and Yu and Feng [23].

5.2 *Simulations*

Note that in the RAM algorithm, we can replace the Lasso penalty with the adaptive Lasso penalty for all regularization steps. We implement both versions and call the corresponding estimators RAM-2-Lasso, RAM-2-AdaLasso, RAM-Lasso, and RAM-AdaLasso.

We compare the performances of model selection and parameter estimation under various ultra-high dimensional linear regression settings. The methods included for comparison are Lasso, AdaLasso, SIS-Lasso, RAR, RAR+, RAM-2-Lasso, RAM-2-AdaLasso, RAM-Lasso, and RAM-AdaLasso. 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 and compare the MSE of the coefficient estimates, i.e., $\|\hat{\beta} - \beta\|_2^2$. All the penalization steps are implemented by using the R package glmnet [14] with corresponding weights. Note that other solution path calculation methods can also be used, including LARS [4] and APPLE [22]. The following scenarios are considered.

(1) The covariance matrix Σ is

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

Set $r = 0.6, \sigma = 3.5, s_n = 4, \beta_S = (3, -2, 2, -2)^T, \beta = (\beta_S^T, 0^T)^T$. After calculation, the absolute value of correlations between response and predictors are $(0.390, 0.043, 0.304, 0.043, 0.130, 0.130, 0.130, 0.130, 0, 0, \dots)^T$.

(2) The covariance matrix Σ is

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & I \end{bmatrix}, \text{ where } \Sigma_{11} = \begin{bmatrix} 1 & \dots & r & 0 \\ \vdots & \ddots & \vdots & \vdots \\ r & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix}_{(2s_n-1)\times(2s_n-1)} .$$

(2a) Set $r = 0.5, \sigma = 2.5, s_n = 5, \beta_S = (3, 2, 1, -1, 0.75)^T, \beta = (0, 0, 0, 0, \beta_S^T, 0^T)^T$. After calculation, the absolute value of correlations between response and predictors are $(0.483, 0.483, 0.483, 0.483, 0.772, 0.676, 0.579, 0.386, 0.145, 0, 0, \dots)^T$.

(2b) Set $r = 0.5, \sigma = 2, s_n = 5, \beta_S = (2.5, 2, 1, -1, 0.5)^T, \beta = (0, 0, 0, 0, \beta_S^T, 0^T)^T$. After calculation, the absolute value of correlations between response and predictors are $(0.497, 0.497, 0.497, 0.497, 0.773, 0.718, 0.607, 0.387, 0.110, 0, 0, \dots)^T$.

For SIS-Lasso, we select the top $n - 1$ variables with largest absolute value of marginal correlations for fair comparison with RAMs. For AdaLasso, the weights are $\omega_j = 1/|\hat{\beta}_j^{\mathcal{M}}|$ as shown in (4). According to Weng et al. [21], the threshold γ_n for RAR/RAR+ is determined by one time permuted data,

$$\gamma_n = \max_{1\le j\le p} \left\{ |D_j^*| \middle| D_j^* = \sum_{i=1}^{n} \frac{(X_i^j - \bar{X}^j)}{\sum_{i=1}^{n}(X_i^j - \bar{X}^j)^2} Y_{(i)} \right\} .$$

For all penalized estimators, EBIC is used to select the tuning parameter. Tables 2 and 3 show the sign recovery proportion and MSE for each method.

In Scenario 1, only RAR+ and RAM-Lasso perform well especially when the dimension p_n becomes large. As the consequence of small marginal correlation coefficients β_2 and β_4, the two corresponding signals are screened out at the beginning, leading to the failure of SIS-Lasso and RAM-2. Their weak marginal correlations also lead to heavy penalties in regularization, which leads to the low sign recovery proportion and large MSE of AdaLasso as well as RAM-AdaLasso. In this scenario, RAR+ and RAM-Lasso perform the best in terms of both sign recovery proportion and the MSE.

In Scenario 2, an independent signal is included in both Scenario 2a and Scenario 2b, which leads to some interesting findings. For Scenario 2a, RAM-2-AdaLasso has impressive high success rates as RAM-AdaLasso does. This emphasizes the important role of marginal learning (RAM-AdaLasso v.s. AdaLasso) and the advantage from screening (RAM-2-AdaLasso v.s. AdaLasso). Noteworthy, RAM-2-Lasso is also comparable to RAR+ and RAM-Lasso, so it indicates that the more efficient version RAM-2 is a worthwhile alternative for variable selection. In Scenario 2b, with respect to the sign recovery proportion and the MSE criteria, RAR+ takes the lead while RAM-Lasso and RAM-2-Lasso follow closely.

Table 2 Sign recovery proportion over 200 simulation rounds of each method

n	100	200	300	400	500
Scenario 1					
Lasso	0.000	0.000	0.045	0.235	0.450
SIS-Lasso	0.000	0.000	0.015	0.065	0.095
AdaLasso	0.000	0.000	0.010	0.025	0.035
RAR	0.015	0.245	0.370	0.320	0.360
RAR+	**0.025**	**0.515**	**0.870**	**0.900**	**0.935**
RAM-2-Lasso	0.000	0.040	0.125	0.130	0.145
RAM-Lasso	**0.090**	**0.630**	**0.890**	**0.880**	**0.870**
RAM-2-AdaLasso	0.000	0.015	0.050	0.065	0.090
RAM-AdaLasso	0.000	0.050	0.190	0.290	0.330
Scenario 2a					
Lasso	0.000	0.000	0.005	0.010	0.035
SIS-Lasso	0.000	0.000	0.000	0.005	0.030
AdaLasso	0.000	0.125	0.380	0.625	0.675
RAR	0.000	0.000	0.000	0.000	0.000
RAR+	0.000	0.095	0.295	0.550	0.665
RAM-2-Lasso	0.000	0.080	0.300	0.530	0.675
RAM-Lasso	0.000	0.100	0.300	0.505	0.645
RAM-2-AdaLasso	**0.000**	**0.105**	**0.420**	**0.680**	**0.835**
RAM-AdaLasso	**0.000**	**0.125**	**0.425**	**0.710**	**0.850**
Scenario 2b					
Lasso	0.000	0.000	0.005	0.105	0.300
SIS-Lasso	0.000	0.000	0.005	0.090	0.255
AdaLasso	0.000	0.075	0.200	0.335	0.390
RAR	0.000	0.000	0.000	0.000	0.000
RAR+	**0.000**	**0.160**	**0.330**	**0.560**	**0.720**
RAM-2-Lasso	**0.000**	**0.110**	**0.315**	**0.495**	**0.630**
RAM-Lasso	**0.005**	**0.125**	**0.350**	**0.535**	**0.680**
RAM-2-AdaLasso	0.000	0.100	0.300	0.445	0.575
RAM-AdaLasso	0.000	0.120	0.315	0.470	0.645

Note: By setting $p_n = \lfloor 100\exp(n^{0.2}) \rfloor$, the number of variables are 1232, 1791, 2285, 2750, and 3199, respectively. The bold values represent the best performing methods under each scenario

6 Discussion

In this work, we propose a general framework for variable selection in ultra-high dimensional linear regression model by incorporating marginal information before regularization. It is shown to have sign consistency under a weaker condition compared with the one-step procedure if the marginal information is helpful.

The framework is quite general and can be easily extended to the case of generalized linear models as well as any other penalty form. Another important

Table 3 Mean square error $\|\hat{\beta} - \beta\|_2^2$ over 200 simulation rounds of each method

n	100	200	300	400	500
Scenario 1					
Lasso	4.0218	3.5174	2.3559	1.3320	0.8017
SIS-Lasso	4.0606	3.5962	3.1029	2.9109	2.7388
AdaLasso	3.9857	3.6522	3.2897	3.0739	2.7821
RAR	3.3556	1.6786	0.9485	0.7303	0.6733
RAR+	**3.4226**	**1.5673**	**0.7130**	**0.5413**	**0.4585**
RAM-2-Lasso	3.9420	3.3433	2.9404	2.8482	2.7078
RAM-Lasso	**3.0516**	**1.3942**	**0.7336**	**0.6030**	**0.5093**
RAM-2-AdaLasso	3.9469	3.4311	3.0538	2.9292	2.7711
RAM-AdaLasso	3.7514	2.9639	2.0931	1.7351	1.5575
Scenario 2a					
Lasso	1.7728	1.5525	1.3311	1.1789	1.0086
SIS-Lasso	1.7684	1.5488	1.3385	1.1724	0.9764
AdaLasso	1.7296	1.3663	0.8032	0.5767	0.4101
RAR	1.6421	0.9908	0.7189	0.6048	0.5104
RAR+	1.8041	1.3893	0.9026	0.6154	0.4390
RAM-2-Lasso	1.8471	1.4442	1.0207	0.6850	0.5125
RAM-Lasso	1.8900	1.4271	1.0111	0.6608	0.4937
RAM-2-AdaLasso	**1.8362**	**1.4144**	**0.8883**	**0.5946**	**0.4340**
RAM-AdaLasso	**1.8279**	**1.4105**	**0.8534**	**0.5873**	**0.4351**
Scenario 2b					
Lasso	1.5437	1.4189	1.2009	0.8720	0.5943
SIS-Lasso	1.5344	1.4152	1.1709	0.8118	0.5503
AdaLasso	1.5372	0.8574	0.5612	0.4743	0.4279
RAR	1.2475	0.7910	0.6334	0.5047	0.4399
RAR+	**1.5458**	**0.9677**	**0.6210**	**0.4027**	**0.3242**
RAM-2-Lasso	**1.5932**	**1.0729**	**0.6631**	**0.4806**	**0.4106**
RAM-Lasso	**1.6032**	**1.0598**	**0.6585**	**0.4759**	**0.3867**
RAM-2-AdaLasso	1.5786	0.9240	0.5973	0.4957	0.4237
RAM-AdaLasso	1.5781	0.9262	0.5999	0.4880	0.4026

The bold values represent the best performing methods under each scenario

extension would be the high dimensional classification [9, 10]. How to develop the parallel theory for those extensions would be an interesting future work.

Acknowledgements This work was partially supported by NSF grant DMS-1308566. The authors would like to thank the Editor and the referee for constructive comments which greatly improved the paper. The majority of the work was done when Mengjia Yu was an M.A. student at Columbia University.

Appendix

Proof of Theorem 1 Denote the design matrix by X, response vector by Y, and error vector by ε. The scale condition is $\log p_n = O(n^{a_1})$, $s_n = O(n^{a_2}), a_1 > 0, a_2 > 0, a_1 + 2a_2 < 1$.

Step I: Recall the index of variables with large coefficients

$$\mathcal{M}_{\tilde{\gamma}_n} = \{1 \le j \le p : |\hat{\beta}^M_j| \text{ is among the first} \lfloor \tilde{\gamma}_n \rfloor \text{ of all } \} = \mathcal{N}^c.$$

Under Corollary 1,

$$\mathrm{pr}(S \subset \mathcal{M}_{\tilde{\gamma}_n} = \mathcal{N}^c = \mathcal{R} \cup \mathcal{U}) \to 1 \quad \text{as } n \to \infty.$$

Hence with high probability the set $\hat{\mathcal{N}}$ contains only noises.

Step II: Next we will show that RAM-2 succeeds in detecting signals in $\hat{\mathcal{N}}^c$. Let $S = \{1 \le j \le p : \beta_j \neq 0\}$. Denote the compositions $S = \hat{\mathcal{R}}_1 \cup \hat{\mathcal{U}}_1$ and define the set of noises left in $\hat{\mathcal{N}}^c$ as $(\hat{\mathcal{R}} \setminus \hat{\mathcal{R}}_1) \cup (\hat{\mathcal{U}} \setminus \hat{\mathcal{U}}_1) \doteq \hat{\mathcal{R}}_2 \cup \hat{\mathcal{U}}_2$, where $\hat{\mathcal{R}}_1$ and $\hat{\mathcal{U}}_1$ are signals from $\hat{\mathcal{R}}$ and $\hat{\mathcal{U}}$, respectively.

Firstly, we would like to introduce an important technique in RAR+. Define the set of true signals as S, and in an arbitrary regularization, define the set that is hold without penalty as H while the set that needs to be checked with penalty as C. Let

$$\check{\beta} = \arg\min_{\beta} \left\{ (2n)^{-1} \|Y - X\beta\|_2^2 + \lambda_n \|\beta_C\|_1 \right\}, \tag{12}$$

$$\bar{\beta} = \arg\min_{\beta_{(S\cup H)^c}=0} \left\{ (2n)^{-1} \|Y - X\beta\|_2^2 + \lambda_n \|\beta_{C\cap S}\|_1 \right\}. \tag{13}$$

Now we define $Q = S \cup H$ which are the variables we would like to retain, and then the variables that are supposed to be discarded are $Q^c = C \setminus S$.

By optimality conditions of convex problems [2], $\check{\beta}$ is a solution to (12) if and only if

$$n^{-1} X^T (Y - X\check{\beta}) = \lambda_n \partial \|\check{\beta}_C\|, \tag{14}$$

where $\partial\|\check{\beta}_C\|$ is the subgradient of $\|\beta_C\|_1$ at $\beta = \check{\beta}$. Namely, the ith $(1 \le i \le p_n)$ element of $\partial\|\check{\beta}_C\|$ is

$$(\partial\|\check{\beta}_C\|)_i = \begin{cases} 0 & \text{if } i \in C; \\ \mathrm{sign}(\check{\beta}_i) & \text{if } i \in C^c \text{ and } \check{\beta}_i \neq 0; \\ t & \text{otherwise,} \end{cases}$$

where t can be any real number with $|t| \leq 1$. Similarly, $\bar{\beta}$ is the unique solution to (13) if and only if

$$\bar{\beta}_{Q^c} = 0, \quad n^{-1}X_Q^T(Y - X_Q\bar{\beta}_Q) = \lambda_n \mathrm{sig}(\bar{\beta}_Q), \tag{15}$$

where $\mathrm{sig}(\bar{\beta}_Q)$, a vector of length $\mathrm{card}(Q)$, is the subgradient of $\|\beta_{\bar{Q}^c}\|_1$ at $\beta_Q = \bar{\beta}_Q$. Then it is not hard to see that the unique solution $\bar{\beta}$ is also a solution for (13) if

$$\|n^{-1}X_{Q^c}^T(Y - X_Q\bar{\beta}_Q)\|_\infty < \lambda_n, \tag{16}$$

simply because (15) and (16) imply $\bar{\beta}$ satisfies (14). Solving the equation in (15) gives

$$\bar{\beta}_Q = (X_Q^TX_Q)^{-1}\left[X_Q^TY - n\lambda_n\mathrm{sig}(\bar{\beta}_Q)\right]. \tag{17}$$

Using (17) and $Y = X_S\beta_S + \varepsilon$, (16) is equivalent to

$$\begin{aligned}&\|X_{Q^c}^TX_Q(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar{\beta}_Q)\\ &\quad + (n\lambda_n)^{-1}X_{Q^c}^T(I - X_Q(X_Q^TX_Q)^{-1}X_Q^T)(X_S\beta_S + \varepsilon)\|_\infty < 1\end{aligned} \tag{18}$$

Since $(I - X_Q(X_Q^TX_Q)^{-1}X_Q^T)X_Q = 0$, (18) can be simplified as

$$\|X_{Q^c}^TX_Q(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar{\beta}_Q) + (n\lambda_n)^{-1}X_{Q^c}^T(I - X_Q(X_Q^TX_Q)^{-1}X_Q^T)\varepsilon\|_\infty < 1. \tag{19}$$

Note that, if there is a unique solution for (12), say $\check{\beta}$, and $\bar{\beta}$ satisfies (19), then $\bar{\beta}$ is indeed the unique solution for (12). This is equivalent to $\check{\beta}_{Q^c} = 0$. Furthermore, if $\min_{j\in Q}|\beta_j| > \|\beta_j - \bar{\beta}_j\|_\infty$ also holds, we can conclude $\check{\beta}_Q \neq 0$. Thus (12) achieves sign recovery. In the following, we will make use of this idea repeatedly.

Secondly, consider the Step 1 (5),

$$\begin{aligned}\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1} &= \mathop{\arg\min}_{\beta_{\hat{\mathcal{N}}}=0}\left\{(2n)^{-1}\sum_{i=1}^n\left(Y_i - \sum_{j\in\hat{\mathcal{U}}}X_{ij}\beta_j - \sum_{k\in\hat{\mathcal{R}}}X_{ik}\beta_k\right)^2 + \lambda_n\sum_{j\in\hat{\mathcal{U}}}|\beta_j|\right\},\\ &= \mathop{\arg\min}_{\beta_{\hat{\mathcal{N}}}=0}\left\{(2n)^{-1}\|Y - X\beta\|_2^2 + \lambda_n\|\beta_{\hat{\mathcal{U}}}\|_1\right\}.\end{aligned} \tag{20}$$

Here, denote $\check{\beta} = \hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1}$. After this step, the ideal result is that with high probability,

$$\check{\beta}_{\hat{\mathcal{U}}_1} \neq 0 \ \text{ and } \ \check{\beta}_{\hat{\mathcal{U}}\backslash\hat{\mathcal{U}}_1} = 0. \tag{21}$$

Therefore, define an oracle estimator of (20),

$$\bar{\beta} = \underset{\beta_{(\hat{\mathcal{R}}\cup\hat{\mathcal{U}}_1)^c}=0}{\arg\min} \left\{ (2n)^{-1}\|Y - X_{\hat{Q}}\beta_{\hat{Q}}\|_2^2 + \lambda_n \|\beta_{\hat{\mathcal{U}}_1}\|_1 \right\}, \tag{22}$$

where $\hat{Q} = \hat{\mathcal{R}} \cup \hat{\mathcal{U}}_1 = S \cup \hat{\mathcal{R}}_2$. Now, we plug $\check{\beta}, \bar{\beta}$, and $\hat{Q}$ back to (12), (13), and (19), then it is sufficient to prove (20) has a unique solution and it achieves sign consistency with $Q = \hat{Q}$.

Let

$$\begin{aligned} F &= X_{\hat{Q}^c}^T - \Sigma_{\hat{Q}^c\hat{Q}}\Sigma_{\hat{Q}\hat{Q}}^{-1}X_{\hat{Q}}^T, \\ K_1 &= \Sigma_{\hat{Q}^c\hat{Q}}\Sigma_{\hat{Q}\hat{Q}}^{-1}\mathrm{sig}(\bar{\beta}_{\hat{Q}}), \\ K_2 &= FX_{\hat{Q}}(X_{\hat{Q}}^TX_{\hat{Q}})^{-1}\mathrm{sig}(\bar{\beta}_{\hat{Q}}) + (n\lambda_n)^{-1}F\{I - X_{\hat{Q}}(X_{\hat{Q}}^TX_{\hat{Q}})^{-1}X_{\hat{Q}}^T\}\varepsilon. \end{aligned}$$

Then, (19) is equivalent to

$$\|K_1 + K_2\|_\infty < 1.$$

To be more clear that, since we have already screen $\hat{\mathcal{N}}$ out, $\hat{Q}^c$ is in fact the complement of $\hat{Q}$ under the "universe" $\hat{\mathcal{R}}\cup\hat{\mathcal{U}}$. We write $\hat{Q}^c$ instead of $(\hat{\mathcal{R}}\cup\hat{\mathcal{U}})\backslash\hat{Q} = \hat{\mathcal{U}}_2$ to show a close connection with the analysis in first part above.

Now let

$$\begin{aligned} A &= \{R \subset \hat{\mathcal{R}}_1 \subset S, S \subset \hat{Q} \subset S \cup Z\}, \\ B &= \{S \subset \hat{Q} \subset S \cup Z\}, \\ \mathcal{T}_A &= \{(\mathcal{R}_1, Q) | R \subset \mathcal{R}_1 \subset S, S \subset Q \subset S \cup Z\}. \end{aligned}$$

From Conditions 1 and 4, $P(A) \to 1$ as a direct result of Proposition 2 in Weng et al. [21]. Since Condition 8 implies

$$\mathrm{pr}(\|K_1\|_\infty \leq 1 - \gamma_1) \geq \mathrm{pr}(\{\|K_1\|_\infty \leq 1 - \gamma_1\} \cap A) = \mathrm{pr}(A) \to 1 \tag{23}$$

as given A

$$\|K_1\|_\infty = \|\Sigma_{\hat{Q}^c\hat{Q}}\Sigma_{\hat{Q}\hat{Q}}^{-1}\mathrm{sig}(\bar{\beta}_{\hat{Q}})\|_\infty \leq \|\{\Sigma_{\hat{Q}^c\hat{Q}}\Sigma_{\hat{Q}\hat{Q}}^{-1}\}_{\hat{\mathcal{U}}_1}\|_\infty$$

is always less than $1 - \gamma_1$.

Denote $K_2(\mathcal{R}_1, Q)$ as the analogy of K_2 and $\bar{\beta}_Q$ as the analogy of $\bar{\beta}_{\hat{Q}}$ by replacing $\hat{\mathcal{R}}_1$ and $\hat{Q}$ in (22) with $\mathcal{R}_1$ and Q. Since given X_Q and ε, the j-th element of $K_2(\mathcal{R}_1, Q)$, namely

$$F(j)X_Q(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar{\beta}_Q) + (n\lambda_n)^{-1}F(j)\{I - X_Q(X_Q^TX_Q)^{-1}X_Q^T\}\varepsilon, \tag{24}$$

is normally distributed with mean 0 and variance V_j, where

$$
\begin{aligned}
V_j &\le (\Sigma_{Q^c|Q})_{jj}\Big[\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q)+(n\lambda_n)^{-2}\varepsilon^T\{I-X_Q(X_Q^TX_Q)^{-1}X_Q^T\}\varepsilon\Big]\\
&\le \mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q)+(n\lambda_n)^{-2}\|\varepsilon\|_2^2.
\end{aligned}
$$

Hence, we let

$$
\begin{aligned}
H = \bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\Big\{&\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q)+(n\lambda_n)^{-2}\|\varepsilon\|_2^2\\
&>\frac{s_n+z_n}{nC_{\min}}(8(s_n+z_n)^{1/2}n^{-1/2}+1)+(1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2)\Big\}.
\end{aligned}
$$

Next, we want to show

$$
\begin{aligned}
\mathrm{pr}\Big(\|K_2\|_\infty>\frac{\gamma_1}{2}\Big) &\le \mathrm{pr}\Big(\Big\{\|K_2\|_\infty>\frac{\gamma_1}{2}\Big\}\cap A\Big)+\mathrm{pr}(A^c)\\
&\le \mathrm{pr}\Big(\Big\{\bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\|K_2(\mathcal{R}_1,Q)\|_\infty>\frac{\gamma_1}{2}\Big\}\cap A\Big)+\mathrm{pr}(A^c)\\
&\le \mathrm{pr}\Big(\bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\|K_2(\mathcal{R}_1,Q)\|_\infty>\frac{\gamma_1}{2}\mid H^c\Big)+\mathrm{pr}(H)+\mathrm{pr}(A^c)\\
&\longrightarrow 0. \qquad (25)
\end{aligned}
$$

By the tail probability inequality of Gaussian distribution (inequality (48) in Wainwright [20]), it is not hard to see that

$$
\begin{aligned}
&\mathrm{pr}\Big(\bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\|K_2(\mathcal{R}_1,Q)\|_\infty>\frac{\gamma_1}{2}\mid H^c\Big)\\
&\le \sum_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\mathrm{pr}(\|K_2(\mathcal{R}_1,Q)\|_\infty>\frac{\gamma_1}{2}\mid H^c)\\
&\le 2^{s_n+z_n}\cdot\max_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\mathrm{pr}(\|K_2(\mathcal{R}_1,Q)\|_\infty>\frac{\gamma_1}{2}\mid H^c)\\
&\le 2^{s_n+z_n}\cdot 2(p_n-s_n)\exp(-\gamma_1^2/8V), \qquad (26)
\end{aligned}
$$

where $V=(1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2)+\frac{s_n+z_n}{nC_{\min}}(8(s_n+z_n)^{1/2}n^{-1/2}+1)\ge V_j$ under condition H^c. Since $\log[2^{s_n+z_n+1}(p_n-s_n)]=o(\gamma_1^2/8V)$ under our scaling, (26) $\to 0$.

To bound $\mathrm{pr}(H)$, note that

$$\begin{aligned}\mathrm{pr}(H) \le \mathrm{pr}\Big(\bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\Big\{&\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q)\\ &> \frac{s_n+z_n}{nC_{\min}}(8(s_n+z_n)^{1/2}n^{-1/2}+1)\Big\}\Big)\\ &+\mathrm{pr}\Big((n\lambda_n)^{-2}\|\varepsilon\|_2^2 > (1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2)\Big) \qquad (27)\end{aligned}$$

Since $\|\varepsilon\|_2^2 \sim \chi^2(n)$, using the inequality of (54a) in Wainwright [20], we get

$$\begin{aligned}\mathrm{pr}\Big((n\lambda_n)^{-2}\|\varepsilon\|_2^2 > (1+s_n^{1/2}n^{-1/2})/(n\lambda_n^2)\Big) &\le \mathrm{pr}\Big(\|\varepsilon\|_2^2 \ge (1+s_n^{1/2}n^{-1/2})n\Big)\\ &\le \exp(-\frac{3}{16}s_n), \qquad (28)\end{aligned}$$

whenever $s_n/n < 1/2$. For any given Q that satisfying $S \subset Q \subset S\cup Z$,

$$\begin{aligned}\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q) &\le (s_n+z_n)\|(X_Q^TX_Q)^{-1}\|_2\\ &\le (s_n+z_n)/n\Big(\|(X_Q^TX_Q/n)^{-1}-\Sigma_{QQ}^{-1}\|_2+\|\Sigma_{QQ}^{-1}\|_2\Big)\\ &\le (s_n+z_n)/n\Big(\|(X_Q^TX_Q/n)^{-1}-\Sigma_{QQ}^{-1}\|_2+1/C_{\min}\Big).\end{aligned}$$

holds for any $\mathcal{R}_1$ that satisfying $R \subset \mathcal{R}_1 \subset S$. Therefore, by the concentration inequality of (58b) in Wainwright [20],

$$\begin{aligned}&\mathrm{pr}\Big(\bigcup_{(\mathcal{R}_1,Q)\subset\mathcal{T}_A}\Big\{\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q) > \frac{s_n+z_n}{nC_{\min}}(8(s_n+z_n)^{1/2}n^{-1/2}+1)\Big\}\Big)\\ &\le \sum_{S\subset Q\subset S\cup Z}\mathrm{pr}\Big(\bigcup_{R\subset\mathcal{R}_1\subset S}\Big\{\mathrm{sig}(\bar\beta_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar\beta_Q) > \frac{s_n+z_n}{nC_{\min}}(8(s_n+z_n)^{1/2}n^{-1/2}+1)\Big\}\Big)\\ &\le \sum_{S\subset Q\subset S\cup Z}\mathrm{pr}\Big(\|(X_Q^TX_Q/n)^{-1}-\Sigma_{QQ}^{-1}\|_2 \ge \frac{8}{C_{\min}}(s_n+z_n)^{1/2}n^{-1/2}\Big)\\ &\le \sum_{S\subset Q\subset S\cup Z}\mathrm{pr}\Big(\|(X_Q^TX_Q/n)^{-1}-\Sigma_{QQ}^{-1}\|_2 \ge \frac{8}{C_{\min}}(\mathrm{Card}(Q))^{1/2}n^{-1/2}\Big)\\ &\le 2^{z_n+1}\exp\left(-\frac{s_n}{2}\right). \qquad (29)\end{aligned}$$

Hence, (28) and (29) imply $\mathrm{pr}(H) \le 2^{z_n+1}\exp(-\frac{s_n}{2}) + \exp(-\frac{3}{16}s_n) \to 0$.

Since $P(A^c) = 1 - P(A) \to 0$, the inequalities (26)–(29) imply (25) under the scaling in Theorem 1. Thus $\|K_1 + K_2\|_\infty < 1$ achieves with high probability, which also means $\breve{\beta}_{\hat{\mathcal{U}}\setminus\hat{\mathcal{U}}_1} = 0$ achieves asymptotically.

From our analysis in the first part, the following goal is the uniqueness of (20). If there is another solution, let's call it $\breve{\beta}'$. For any t such that $0 < t < 1$, the linear combination $\breve{\beta}(t) = t\breve{\beta} + (1-t)\breve{\beta}'$ is also a solution to (20) as a consequence of the convexity. Note that, the new solution point $\breve{\beta}(t)$ satisfies (16) and $\breve{\beta}(t)_{Q^c} = 0$, hence it is a solution to (13). From the uniqueness of (13), we conclude that $\breve{\beta} = \breve{\beta}'$.

The last part of this step is to prove $\bar{\beta}_{\mathcal{U}_1} \neq 0$ with high probability. By (17) and $Y = X_S\beta_S + \varepsilon = X_{\hat{Q}}\beta_{\hat{Q}} + \varepsilon$, we have

$$\begin{aligned}
\|\beta_{\hat{Q}} - \bar{\beta}_{\hat{Q}}\|_\infty &= \|\lambda_n (X_{\hat{Q}}^T X_{\hat{Q}}/n)^{-1}\mathrm{sig}(\bar{\beta}_{\hat{Q}}) - (X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty \\
&\le \lambda_n \|(X_{\hat{Q}}^T X_{\hat{Q}}/n)^{-1}\|_\infty + \|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty \\
&\le \lambda_n (s_n + z_n)^{1/2}\|(X_{\hat{Q}}^T X_{\hat{Q}}/n)^{-1}\|_2 + \|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty \\
&\le \lambda_n (s_n + z_n)^{1/2}(\|(X_{\hat{Q}}^T X_{\hat{Q}}/n)^{-1} - \Sigma_{\hat{Q}\hat{Q}}^{-1}\|_2 + 1/C_{\min}) \\
&\quad + \|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty
\end{aligned} \tag{30}$$

for any $\hat{Q}$ satisfying $S \subset \hat{Q} \subset S \cup Z$. In (29), we have already got

$$\mathrm{pr}\Big(\|(X_{\hat{Q}}^T X_{\hat{Q}}/n)^{-1} - \Sigma_{\hat{Q}\hat{Q}}^{-1}\|_2 \ge \frac{8}{C_{\min}}(s_n + z_n)^{1/2} n^{-1/2}\Big) \le 2\exp(-\frac{s_n}{2}) \tag{31}$$

Let $G = \Big\{\|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}\|_2 > 9/(nC_{\min})\Big\}$, by the inequality (60) in Wainwright [20],

$$\mathrm{pr}(G) \le \mathrm{pr}(\|(X^T X)^{-1}\|_2 > 9/(nC_{\min})) \le 2\exp(-n/2).$$

Since $(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon \mid X_{\hat{Q}} \sim N(0, (X_{\hat{Q}}^T X_{\hat{Q}})^{-1})$, then when we condition on G and achieve

$$\begin{aligned}
&\mathrm{pr}\Big(\|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty > \frac{(s_n + z_n)^{1/2}}{n^{1/2}C_{\min}^{1/2}}\Big) \\
&\quad \le \mathrm{pr}\Big(\|(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon\|_\infty > \frac{(s_n + z_n)^{1/2}}{n^{1/2}C_{\min}^{1/2}} \mid G^c\Big) + \mathrm{pr}(G) \\
&\quad \le 2(s_n + z_n)e^{-(s_n+z_n)/18} + 2e^{-n/2},
\end{aligned} \tag{32}$$

since under G^c, each component of $(X_{\hat{Q}}^T X_{\hat{Q}})^{-1}X_{\hat{Q}}^T\varepsilon \mid X_{\hat{Q}}$ is normally distributed with mean 0 and variance that is less than $9/(nC_{\min})$.

Hence (30)–(32) together imply that,

$$\|\bar{\beta}_{\hat{Q}}-\beta_{\hat{Q}}\|_\infty \le U_n \doteq \lambda_n(s_n+z_n)^{1/2}\Big(\frac{8}{C_{\min}}(s_n+z_n)^{1/2}n^{-1/2}+1/C_{\min}\Big)+\frac{(s_n+z_n)^{1/2}}{n^{1/2}C_{\min}^{1/2}}$$

holds with probability larger than $2(s_n+z_n)e^{-(s_n+z_n)/18}+2e^{-n/2}+2\exp^{-s_n/2}$. Therefore,

$$\begin{aligned}
&\mathrm{pr}(\|\bar{\beta}_{\hat{Q}}-\beta_{\hat{Q}}\|_\infty \ge U_n)\\
&\quad\le \mathrm{pr}\Big(\bigcup_{S\subset Q\subset S\cup Z}\{\|\bar{\beta}_Q-\beta_Q\|_\infty \ge U_n\}\cap B\Big)+\mathrm{pr}(B^c)\\
&\quad\le 2^{z_n}\Big(2(s_n+z_n)e^{-(s_n+z_n)/18}+2e^{-n/2}+2\exp^{-s_n/2}\Big)+\mathrm{pr}(B^c) \qquad (33)
\end{aligned}$$

Under the scaling of Theorem 1, we have $\mathrm{pr}(B)\ge\mathrm{pr}(A)\to 1$ and $2^{z_n}(2(s_n+z_n)e^{-(s_n+z_n)/18}+2e^{-n/2}+2\exp^{-s_n/2})\to 0$. From Condition 5, it is easy to verify that

$$\min_{j\in S}|\beta_j|>U_n,$$

for sufficiently large n. Thus with high probability $\min_{j\in S}|\beta_j|>\|\bar{\beta}_{\hat{Q}}-\beta_{\hat{Q}}\|_\infty$ as n increases, which also implies $\check{\beta}_{\hat{Q}}\neq 0$ with high probability.

Finally, $\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1}$ exactly recover signals with high probability as $n\to\infty$.

Step III: We need to prove that RAM-2 succeeds in detecting signals via Step 3. Similar to Step II, we need to define proper $\check{\beta}$ in (12) and $\bar{\beta}$ in (13). Since the main idea is the same as the procedure above, we only describe the key steps in the following proof. Recall the estimator (7),

$$\begin{aligned}
\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} &= \operatorname*{arg\,min}_{\beta_{\hat{\mathcal{U}}_2\cup\hat{\mathcal{N}}_2}=0}\left\{(2n)^{-1}\sum_{i=1}^n\Big(Y_i-\sum_{j\in\hat{\mathcal{R}}}X_{ij}\beta_j-\sum_{k\in\hat{\mathcal{U}}_1\cup\hat{\mathcal{N}}_1}X_{ik}\beta_k\Big)^2+\lambda_n^{\star\star}\sum_{j\in\hat{\mathcal{R}}}|\beta_j|\right\}\\
&= \operatorname*{arg\,min}_{\beta_{\hat{\mathcal{N}}\cup\hat{\mathcal{U}}\setminus\hat{\mathcal{U}}_1}=0}\left\{(2n)^{-1}\|Y-X\beta\|_2^2+\lambda_n^{\star\star}\|\beta_{\hat{\mathcal{R}}}\|_1\right\}. \qquad (34)
\end{aligned}$$

This is a new "$\check{\beta}$" in (12), and we denote it as $\tilde{\beta}$. After this step, the ideal result is that with high probability,

$$\tilde{\beta}_{\hat{\mathcal{R}}_1}\neq 0 \text{ and } \tilde{\beta}_{\hat{\mathcal{R}}\setminus\hat{\mathcal{R}}_1}=0. \qquad (35)$$

Therefore, define an oracle estimator of (34),

$$\mathring{\beta} = \mathop{\arg\min}_{\beta_{S^c}=0} \left\{ (2n)^{-1}\|Y - X_S\beta_S\|_2^2 + \lambda_n^{\star\star}\|\beta_{\hat{\mathcal{R}}_1}\|_1 \right\}. \tag{36}$$

Now, we plug $\tilde{\beta}$ and $\mathring{\beta}$ back to (12), (13), and (18), then it is sufficient to prove (34) has a unique solution and it achieves sign consistency with $Q = S$. Let

$$\begin{aligned} F' &= X_{\hat{\mathcal{R}}_2}^T - \Sigma_{\hat{\mathcal{R}}_2 S}\Sigma_S^{-1}X_S^T, \\ K_1' &= \Sigma_{\hat{\mathcal{R}}_2 S}\Sigma_{SS}^{-1}\mathrm{sig}(\mathring{\beta}_S), \\ K_2' &= F'X_S(X_S^TX_S)^{-1}\mathrm{sig}(\mathring{\beta}_S) + (n\lambda_n^{\star\star})^{-1}F'\{I - X_S(X_S^TX_S)^{-1}X_S^T\}\varepsilon. \end{aligned}$$

Similarly,

$$\mathrm{pr}\Big(\|K_1'\|_\infty \le 1-\alpha\Big) \ge \mathrm{pr}\Big(\big\{\|K_1'\|_\infty \le 1-\alpha\big\} \cap D\Big) \ge \mathrm{pr}(D) \ge \mathrm{pr}(A) \to 1, \tag{37}$$

where $D = \{\hat{\mathcal{R}}_2 \subset Z\}$ and it implies $\|K_1'\|_\infty \le 1-\alpha$ under Condition 7. Let

$$\begin{aligned} H' = \bigcup_{R\subset\mathcal{R}_2\subset S} \Big\{ & \mathrm{sig}(\mathring{\beta}_S)^T(X_S^TX_S)^{-1}\mathrm{sig}(\mathring{\beta}_S) + (n\lambda_n^{\star\star})^{-2}\|\varepsilon\|_2^2 > \frac{s_n}{nC_{\min}}\big(8s_n^{1/2}n^{-1/2}+1\big) \\ & + \big(1+s_n^{1/2}n^{-1/2}\big)/\big(n(\lambda_n^{\star\star})^2\big)\Big\}. \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{pr}\Big(\|K_2'\|_\infty > \frac{\alpha}{2}\Big) &\le \mathrm{pr}\Big(\Big\{\|K_2'\|_\infty > \frac{\alpha}{2}\Big\} \cap A\Big) + \mathrm{pr}(A^c) \\ &\le \mathrm{pr}\Big(\bigcup_{\substack{(\mathcal{R}_2,\mathcal{R}_1)\\ \mathcal{R}_2\subset Z\\ R\subset\mathcal{R}_1\subset S}} \Big\{\|\tilde{K}_2(\mathcal{R}_2,\mathcal{R}_1)\|_\infty > \frac{\alpha}{2}\Big\}\Big) + \mathrm{pr}(A^c) \\ &\le \mathrm{pr}\Big(\bigcup_{\substack{(\mathcal{R}_2,\mathcal{R}_1)\\ \mathcal{R}_2\subset Z\\ R\subset\mathcal{R}_1\subset S}} \Big\{\|\tilde{K}_2(\mathcal{R}_2,\mathcal{R}_1)\|_\infty > \frac{\alpha}{2}\Big\} \mid \tilde{H}^c\Big) + \mathrm{pr}(\tilde{H}) + \mathrm{pr}(A^c) \\ &\le 2^{z_n+s_n+1}z_ne^{-\alpha^2/8V'} + 2e^{-\frac{s_n}{2}} + e^{-\frac{3}{16}s_n} + \mathrm{pr}(A^c) \\ &\longrightarrow 0, \end{aligned} \tag{38}$$

where the last step of (38) follows from (26), (28), and (29) in the proof of Step II, and $V' = \frac{s_n}{nC_{\min}}(8s_n^{1/2}n^{-1/2}+1) + (1+s_n^{1/2}n^{-1/2})/(n(\lambda_n^{\star\star})^2)$.

Equations (37) and (38) indicate $\tilde{\beta}_{\hat{\mathcal{R}}\setminus\hat{\mathcal{R}}_1} = 0$. We skip the proof of uniqueness and move to the next step of proving $\tilde{\beta}_{\hat{\mathcal{R}}_1} \neq 0$.

$$\begin{aligned}\|\mathring{\beta}_S - \beta_S\|_\infty &= \|(X_S^TX_S)^{-1}(X_S^TY - n\lambda_n^{\star\star}\mathrm{sig}(\mathring{\beta}_S)) - \beta_S\|_\infty \\ &\le \|(X_S^TX_S)^{-1}X_S^T\varepsilon\|_\infty + \|\lambda_n^{\star\star}(X_S^TX_S/n)^{-1}\|_\infty.\end{aligned}$$

Let $W_n = \lambda_n^{\star\star}s_n^{1/2}\big(\frac{8}{C_{\min}}s_n^{1/2}n^{-\frac{1}{2}} + \frac{1}{C_{\min}}\big) + \frac{s_n^{1/2}}{n^{1/2}C_{\min}^{1/2}} = o(n^{a_2/2-\delta})$. In the same way, we can show that as $n \to \infty$

$$\mathrm{pr}\big(\|\mathring{\beta}_S - \beta_S\|_\infty \le W_n\big) \to 0$$

Hence, Condition 5 ensures that as $n \to \infty$,

$$\mathrm{pr}\big(\min_{j\in S}|\beta_j| > \|\mathring{\beta}_S - \beta_S\|_\infty\big) \to 1, \qquad (39)$$

which is equivalent to $\tilde{\beta}_{\hat{\mathcal{R}}_1} \neq 0$.

Finally, combining Step I, Step II, and Step III, we conclude that

$$P\big(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} \text{ is unique and } \mathrm{sign}(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}) = \mathrm{sign}(\beta)\big) \to 1, \quad \text{as } n \to \infty.$$

□

Proof of Theorem 2 Denote the compositions $S = \hat{\mathcal{R}}_1 \cup \hat{\mathcal{U}}_1 \cup \hat{\mathcal{N}}_1$ and define the set of noises left in $\hat{\mathcal{U}}^c$ as $(\hat{\mathcal{R}} \setminus \hat{\mathcal{R}}_1) \cup (\hat{\mathcal{N}} \setminus \hat{\mathcal{N}}_1) \doteq \hat{\mathcal{R}}_2 \cup \hat{\mathcal{N}}_2$, where $\hat{\mathcal{R}}_1$, $\hat{\mathcal{U}}_1$, and $\hat{\mathcal{N}}_1$ are signals from $\hat{\mathcal{R}}$, $\hat{\mathcal{U}}$, and $\hat{\mathcal{N}}$, respectively.

Step I: Consider the Step 1 in (5), which is exactly the same as (20). Since there is no difference from the Step II in the proof of Theorem 1, we skip the details here.

Step II: Let's consider the Step 2 in (6).

$$\begin{aligned}\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} &= \underset{\beta_{\hat{\mathcal{U}}\setminus\hat{\mathcal{U}}_1}=0}{\arg\min}\left\{(2n)^{-1}\sum_{i=1}^n\Big(Y_i - \sum_{j\in\hat{\mathcal{N}}}X_{ij}\beta_j - \sum_{k\in\hat{\mathcal{R}}\cup\hat{\mathcal{U}}_1}X_{ik}\beta_k\Big)^2 + \lambda_n^\star\sum_{j\in\hat{\mathcal{N}}}|\beta_j|\right\} \\ &= \underset{\beta_{\hat{\mathcal{U}}\setminus\hat{\mathcal{U}}_1}=0}{\arg\min}\left\{(2n)^{-1}\|Y - X\beta\|_2^2 + \lambda_n^\star\|\beta_{\hat{\mathcal{N}}}\|_1\right\}. \qquad (40)\end{aligned}$$

Here, denote $\check{\beta} = \hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}$. After this step, the ideal result is that with high probability,

$$\check{\beta}_{\hat{\mathcal{N}}_1} \neq 0 \;\text{ and }\; \check{\beta}_{\hat{\mathcal{N}}\setminus\hat{\mathcal{N}}_1} = 0. \qquad (41)$$

Then, define an oracle estimator of (20),

$$\bar{\beta} = \mathop{\arg\min}_{\beta_{(\hat{\mathcal{R}}\cup\hat{\mathcal{U}}_1\cup\hat{\mathcal{N}}_1)^c}=0} \left\{ (2n)^{-1}\|Y - X_{\hat{Q}}\beta_{\hat{Q}}\|_2^2 + \lambda_n^{\star}\|\beta_{\hat{\mathcal{N}}_1}\|_1 \right\}, \tag{42}$$

where $\hat{Q} = (\hat{\mathcal{R}}\cup\hat{\mathcal{U}}_1)\cup\hat{\mathcal{N}}_1 = S\cup\hat{\mathcal{R}}_2$. Similar to Step II in proof of Theorem 1, let

$$\begin{aligned}
F &= X^T_{\hat{\mathcal{N}}_2} - \Sigma_{\hat{Q}^c\hat{Q}}\Sigma^{-1}_{\hat{Q}\hat{Q}}X^T_{\hat{Q}},\\
K_1 &= \Sigma_{\hat{\mathcal{N}}_2\hat{Q}}\Sigma^{-1}_{\hat{Q}\hat{Q}}\mathrm{sig}(\bar{\beta}_{\hat{Q}}),\\
K_2 &= FX_{\hat{Q}}(X^T_{\hat{Q}}X_{\hat{Q}})^{-1}\mathrm{sig}(\bar{\beta}_{\hat{Q}}) + (n\lambda_n)^{-1}F\{I - X_{\hat{Q}}(X^T_{\hat{Q}}X_{\hat{Q}})^{-1}X^T_{\hat{Q}}\}\varepsilon,
\end{aligned}$$

and

$$\begin{aligned}
A &= \{R\subset\hat{\mathcal{L}}_1\doteq\hat{\mathcal{R}}_1\cup\hat{\mathcal{U}}_1\subset S, S\subset\hat{Q}\subset S\cup Z\},\\
B &= \{S\subset\hat{Q}\subset S\cup Z\},\\
\mathcal{T}_A &= \{(\mathcal{L}_1, Q)|R\subset\mathcal{L}_1\subset S, S\subset Q\subset S\cup Z\}.
\end{aligned}$$

Similarly, we get

$$\mathrm{pr}(\|K_1\|_\infty \le 1-\gamma_1) \ge \mathrm{pr}(\{\|K_1\|_\infty \le 1-\gamma_1\}\cap A) \ge \mathrm{pr}(A) \to 1. \tag{43}$$

To obtain $\mathrm{pr}(\|K_2\|_\infty > \frac{\gamma_1}{2}) \to 0$, we define event H as

$$\begin{aligned}
H = \bigcup_{(\mathcal{L}_1,Q)\subset\mathcal{T}_A} &\Big\{\mathrm{sig}(\bar{\beta}_Q)^T(X_Q^TX_Q)^{-1}\mathrm{sig}(\bar{\beta}_Q) + (n\lambda_n)^{-2}\|\varepsilon\|_2^2\\
&> \frac{s_n+z_n}{nC_{\min}}\big(8(s_n+z_n)^{1/2}n^{-1/2}+1\big) + \big(1+s_n^{1/2}n^{-1/2}\big)/(n\lambda_n^2)\Big\}.
\end{aligned}$$

Then, following (25)–(29),

$$\begin{aligned}
\mathrm{pr}\Big(\|K_2\|_\infty > \frac{\gamma_1}{2}\Big) &\le \mathrm{pr}\Big(\Big\{\|K_2\|_\infty > \frac{\gamma_1}{2}\Big\}\cap A\Big) + \mathrm{pr}(A^c)\\
&\le \mathrm{pr}\Big(\bigcup_{(\mathcal{L}_1,Q)\subset\mathcal{T}_A}\|K_2(\mathcal{L}_1,Q)\|_\infty > \frac{\gamma_1}{2}\mid H^c\Big) + \mathrm{pr}(H) + \mathrm{pr}(A^c)\\
&\le 2^{s_n+z_n}\cdot 2(p_n-s_n)\exp(-\gamma_1^2/8V) + 2^{z_n+1}\exp\Big(-\frac{s_n}{2}\Big)\\
&\quad + \exp\Big(-\frac{3}{16}s_n\Big)\\
&\longrightarrow 0,
\end{aligned} \tag{44}$$

where $V = (1+s_n^{1/2}n^{-1/2})/(n(\lambda_n^{\star})^2) + s_nn^{-1}C^{-1}_{\min}(8s_n^{1/2}n^{-1/2}+1)$.

Again, we skip the uniqueness of $\check{\beta}$ and move to bound $\|\bar{\beta}_{\hat{Q}} - \beta_{\bar{Q}}\|_\infty$. By (30)–(32) in the proof of Theorem 1, we have

$$\begin{aligned}
&\mathrm{pr}\big(\|\bar{\beta}_{\hat{Q}} - \beta_{\hat{Q}}\|_\infty \geq U_n\big) \\
&\quad \leq 2^{z_n}\big(2(s_n + z_n)e^{-(s_n+z_n)/18} + 2e^{-n/2} + 2\exp^{-s_n/2}\big) + \mathrm{pr}(B^c) \rightarrow 0,
\end{aligned}$$

where $U_n = \lambda_n(s_n + z_n)^{1/2}\Big(\frac{8}{C_{\min}}(s_n + z_n)^{1/2}n^{-1/2} + 1/C_{\min}\Big) + \frac{(s_n+z_n)^{1/2}}{n^{1/2}C_{\min}^{1/2}}$. As $\min_{j\in S}|\beta_j| \gg U_n$ with sufficiently large n, we conclude that with high probability $\min_{j\in S}|\beta_j| > \|\bar{\beta}_{\hat{Q}} - \beta_{\hat{Q}}\|_\infty$ as n increases, which also implies $\check{\beta}_{\hat{Q}} \neq 0$ with high probability.

Therefore, $\hat{\beta}_{\hat{\mathcal{R}},\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}$ successfully recover signals from $\hat{\mathcal{N}}$ with high probability when n is large enough.

Step III: Following the same steps as in Step III in the proof of Theorem 1, we have

$$P\big(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1} \text{ is unique and } \mathrm{sign}(\hat{\beta}_{\hat{\mathcal{R}}_1,\hat{\mathcal{U}}_1,\hat{\mathcal{N}}_1}) = \mathrm{sign}(\beta)\big) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

□

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. arXiv preprint arXiv:1108.0775 (2011)
3. Chen, J., Chen, Z.: Extended bayesian information criteria for model selection with large model spaces. Biometrika **95**, 759–771 (2008)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)
5. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
6. Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. **70**, 849–911 (2008)
7. Fan, J., Song, R.: Sure independence screening in generalized linear models with np-dimensionality. Ann. Stat. **38**, 3567–3604 (2010)
8. Fan, J., Feng, Y., Song, R.: Nonparametric independence screening in sparse ultra-high dimensional additive models. J. Am. Stat. Assoc. **106**, 544–557 (2011)
9. Fan, J., Feng, Y., Tong, X.: A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Stat. Soc. Ser. B. **74**, 745–771 (2012)
10. Fan, J., Feng, Y., Jiang, J., Tong, X.: Feature augmentation via nonparametrics and selection (fans) in high dimensional classification. J. Am. Stat. Assoc. (2014, to appear)
11. Feng, Y., Li, T., Ying, Z.: Likelihood adaptively modified penalties. arXiv preprint arXiv:1308.5036 (2013)

12. Feng, Y., Yu, Y.: Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. arXiv preprint arXiv:1308.5390 (2013)
13. Frank, l.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**, 109–135 (1993)
14. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)
15. Greenshtein, E., Ritov, Y.: Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli **10**, 971–988 (2004)
16. Huang, J., Ma, S., Zhang, C.-H.: Adaptive lasso for sparse high-dimensional regression models. Stat. Sin. **18**, 1603 (2008)
17. Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356–1378 (2000)
18. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
19. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **58**, 267–288 (1996)
20. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery. IEEE Trans. Inf. Theory **55**, 2183–2202 (2009)
21. Weng, H., Feng, Y., Qiao, X.: Regularization after retention in ultrahigh dimensional linear regression models. Manuscript (2013). Preprint, arXiv:1311.5625
22. Yu, Y., Feng, Y.: Apple: approximate path for penalized likelihood estimators. Stat. Comput. **24**, 803–819 (2014)
23. Yu, Y., Feng, Y.: Modified cross-validation for lasso penalized high-dimensional linear models. J. Comput. Graph. Stat. **23**, 1009–1027 (2014)
24. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)
25. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)
26. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**, 301–320 (2005)