Electronic Journal of Statistics
Vol. 11 (2017) 3226–3250
ISSN: 1935-7524
DOI: 10.1214/17-EJS1320

Recovery of weak signal in high dimensional linear regression by data perturbation

Yongli Zhang

Lunquist College of Business
University of Oregon
Eugene, OR 97403
e-mail: yongli@uoregon.edu

Abstract: How to recover weak signals (i.e., small nonzero regression coefficients) is a difficult task in high dimensional feature selection problems. Both convex and nonconvex regularization methods fail to fully recover the true model whenever there exist strong columnwise correlations in design matrices or small nonzero coefficients below some threshold. To address the two challenges, we propose a procedure, Perturbed LASSO (PLA), that weakens correlations in the design matrix and strengthens signals by adding random perturbations to the design matrix. Moreover, a quantitative relationship between the selection accuracy and computing cost of PLA is derived. We theoretically prove and demonstrate using simulations that PLA substantially improves the chance of recovering weak signals and outperforms comparable methods at a limited cost of computation.

MSC 2010 subject classifications: Primary 62J07.
Keywords and phrases: Beta-min condition, data perturbation, high dimensional data, irrepresentable condition, LASSO, weak signal.

Received November 2016.

Contents

1. Introduction

To achieve high accuracy in feature selection performed by a regularization procedure, as proved in [10] and others, the informative (or useful) features whose corresponding regression coefficients are nonzero must be well-separated from the uninformative (or useless) features whose coefficients are zeros. However, strong columnwise correlations in the design matrix and small nonzero regression coefficients make informative and uninformative features inseparable and result in poor selection accuracy. Theoretically, it has been shown that the Least Absolute Shrinkage and Selection Operator (LASSO) [8], a convex regularization procedure, needs to meet the Irrepresentable condition [13] and the Beta-min condition [4] to achieve high selection accuracy. The Irrepresentable condition requires the true model comprising of all informative features be weakly correlated with each uninformative feature, while the Beta-min Condition requires all nonzero regression coefficients be sufficiently large versus some threshold. Similar "weak correlation and strong signal" conditions on the design matrix and the nonzero regression coefficients are needed for nonconvex regularization procedures such as Minimax Concave Penalty (MCP) [9] and Multi-stage Convex Relaxation (MCR) [10]. In summary, "weak correlation and strong signal" are prerequisites for recovery of the true model for both convex and nonconvex regularization procedures. However, they are unlikely to hold and hard to check in practice [4]. In this article, we aim to develop a computationally feasible procedure that is able to recover the true model when there exist strong correlation and/or weak signal.

Many efforts have been directed towards overcoming the Irrepresentable condition and the Beta-min condition. They fall into two categories: decorrelation and resampling. The elastic net [14] weakens columnwise correlation by utilizing the squared second norm of penalty (i.e., the ridge penalty). Other studies show that resampling often leads to higher selection accuracy [1]. Two representative approaches are Stability Selection (SS) [7] and Bootstrapped Enhanced LASSO (BoLASSO) [1].

In this work we develop a procedure, Perturbed LASSO (PLA), to improve selection accuracy by combining the power of decorrelation and resampling. The procedure PLA is implemented in two steps. First, we generate H pseudo-samples from the original data by adding random perturbations to the design matrix repeatedly and create a model subspace by applying LASSO with a set of D predefined regularization parameters to each pseudo-sample; consequently, H pseudo-samples produce H model subspaces and each model subspace consists of D models. Thus, the union of the H subspaces include no more than DH unique models. After the union model space is created, the next step is to perform model selection based on the original data by an information criterion. As shown in theory and numerical studies the random perturbation does not only weaken correlation in the design matrix but also strengths the signal. As a consequence, PLA overcomes both the Irrepresentable condition and the Beta-min condition.

As for the computing cost of PLA, we provide a quantitative relationship between the number of perturbations H and the probability (lower bound) of

selecting the true model, which we believe is a key contribution to the field. In a nutshell, the increase in computation always improves selection accuracy while smallish nonzero regression coefficients always necessitate heavy computation. We run a series of simulations to demonstrate the gains of PLA in selection accuracy against its computing cost and testify that PLA obtains substantially higher selection accuracy than its competitors. Moreover, most these competitors deal with the Irrepresentable condition, but our method addresses both the Irrepresentable condition and the Beta-min condition.

In the past two decades the trade-off between goodness-of-fit and parsimony is a central topic in literature on model selection. However, the trade-off between computing cost and selection accuracy imposes great challenges and is worth our inputs with the advent of high dimensional data era. This work explores the second type of tradeoff from both theoretical and numerical perspectives.

The paper is organized as follows. In Section 2, we set up the problem. PLA is proposed and its properties are studied in Section 3. In Sections 4 and 5, simulation results and a real data example are presented, respectively. Concluding remarks are in Section 6. The proofs of the main results are in the Appendix.

2. Model setting

As is typical in high-dimensional regression analysis, throughout this article we only consider linear models of the form

$$Y = \sum_{j=1}^{p} \beta_j x_j + \epsilon, \tag{2.1}$$

which includes a response variable Y and p deterministic features $(x_1, \cdots, x_p)$. Here we use the standard notation

$$\operatorname{supp}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\} \qquad \|\boldsymbol{\beta}\|_0 = |\operatorname{supp}(\boldsymbol{\beta})| \qquad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

for $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$. All nonempty subsets of $(x_1, ..., x_p)$ constitute the model space $\mathcal{M}$ and each element in $\mathcal{M}$ defines a model M with cardinality $|M|$.

Assume the data are generated by

$$Y = \sum_{j=1}^{p} \beta_j^0 x_j + \epsilon, \tag{2.2}$$

where ϵ is distributed as $N(0, \sigma^2)$. Then we define the true model M_0 as the set of informative or useful features whose regression coefficient $\beta_j^0 \neq 0$. The other features corresponding to zero coefficients are defined as uninformative or useless. Let $q = \|\boldsymbol{\beta}^0\|_0$. Moreover, assume that the first q coefficients $\boldsymbol{\beta}_1^0 = (\beta_1^0, \cdots, \beta_q^0)'$ are nonzero and all other $(p - q)$ coefficients $\boldsymbol{\beta}_2^0 = (\beta_{q+1}^0, \cdots, \beta_p^0)'$ are zeros. Thus, $\operatorname{supp}(\boldsymbol{\beta}^0) = \{1, \cdots, q\}$. Define $\beta_{\min}^0 = \min(|\beta_1^0|, \cdots, |\beta_q^0|)$. The primary

goal of model selection is to recover these nonzero coefficients or the informative features $(x_1, \cdots, x_q)$. In this work we investigate feature selection in the context of high dimensional data $(p > n)$ and sparse true models $(n/(\log p) \gg q)$.

The original sample $(\mathbf{X}, \mathbf{Y})$ include an $n \times p$ feature matrix $\mathbf{X}$ and an n-dimensional vector $\mathbf{Y}$ representing n independent observations on the response variable Y. Let $\boldsymbol{\mu} = E(\mathbf{Y})$ and $\boldsymbol{\varepsilon} = \mathbf{Y} - \boldsymbol{\mu}$. Across this article we only consider the deterministic design. Let $\mathbf{X}_1$ be the submatrix spanned by the first q columns of $\mathbf{X}$ and $\mathbf{X}_2$ be the submatrix by the remaining $p-q$ columns. Let $\mathbf{X}_{[,j]}$ denote the j-th column of $\mathbf{X}$ and suppose $\|\mathbf{X}_{[,j]}\|_2^2 = n$ $(j = 1, ..., p)$. The $p \times p$ matrix $\mathbf{X}'\mathbf{X}$ can be expressed in a block-wise form,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}.$$

Suppose $\boldsymbol{\Sigma}_{11}^n = n^{-1}\mathbf{X}_1'\mathbf{X}_1 \to \boldsymbol{\Sigma}_{11}$ elementwise as $n \to \infty$ and the two $q \times q$ matrices $\boldsymbol{\Sigma}_{11}^n$ and $\boldsymbol{\Sigma}_{11}$ are both positive definite. The eigenvalues of $\boldsymbol{\Sigma}_{11}^n$ and $\boldsymbol{\Sigma}_{11}$ are $0 < \Lambda_1^n \le \cdots \le \Lambda_q^n < q$ and $0 < \Lambda_1 \le \cdots \le \Lambda_q < q$, respectively. Denote $n^{-1}\mathbf{X}_1'\mathbf{X}_2$ by $\boldsymbol{\Sigma}_{12}^n$, which converges to $\boldsymbol{\Sigma}_{12}$ elementwise as $n \to \infty$.

3. Proposed method

This paper focuses on the feature selection problem (i.e., recovery of nonzero coefficients) by a regularization criterion in the form

$$\widehat{\boldsymbol{\beta}}^{p_\lambda} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + p_\lambda(\boldsymbol{\beta}), \tag{3.1}$$

where $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2$ is the Euclidean distance between $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$, and $p_\lambda(\boldsymbol{\beta})$ works as a penalty of controlling model dimension. The family of regularization criteria (3.1) include the L_1 regularization method (i.e., LASSO) where $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ $(\lambda > 0)$; the L_0 regularization method (i.e., information criteria) where $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_0$ $(\lambda > 0)$; and other methods such as MCP and MCR, both of which belong to nonconvex regularizations.

In this article, we are mainly interested in feature selection by LASSO,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{3.2}$$

in which the L_1 penalty $\lambda\|\boldsymbol{\beta}\|_1$ imposes sparsity by shrinking some coefficients to zero and the regularization parameter λ controls sparsity level in that a large λ always leads to large bias and great sparsity. The set of features corresponding to the nonzero entries in $\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda}$ constitute the model selected by LASSO with the regularization parameter λ. As a convex regularization criterion, LASSO is more computationally efficient than nonconvex regularization criteria such as MCP, which can achieve unbiased feature selection by solving a nonconvex optimization problem.

3.1. The Beta-min conditions of convex and nonconvex regularizations

Both convex and nonconvex regularization criteria need to meet their own "weak correlation and strong signal" conditions to achieve exact recovery of the true model [10].

Concerning feature selection by LASSO, the "weak correlation and strong signal" conditions refer to the Irrepresentable condition and the Beta-min condition. In the language of [13] the Irrepresentable condition is

$$|\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0)| <_c 1-\delta \tag{3.3}$$

where δ is a scalar between 0 and 1, the left-hand expression is a $(p-q)$ dimensional vector and $<_c$ denotes that each element of a vector is less than a scalar. The Irrepresentation condition (3.3) works for ruling out the strong correlation in the design matrix. The Beta-min Condition for LASSO is formulated in [4] as

$$\min_{1\le j\le q}|\beta_j^0| \ge C\sigma\sqrt{\frac{q\log p}{n\phi_0^2}} \tag{3.4}$$

where C is a generic constant and ϕ_0^2 is the so-called compatibility constant determined by the design matrix $\mathbf{X}$. Consequently, small nonzero regression coefficients below the threshold $C\sigma\sqrt{q\log p/(n\phi_0^2)}$ cannot be detected by LASSO (in a consistent way). In [10], the LASSO-version Beta-min condition is presented as

$$\min_{1\le j\le q}|\beta_j^0| \ge C_1\sigma\sqrt{\frac{q\log p}{n}}. \tag{3.5}$$

where C_1 corresponds to C/ϕ_0 in (3.4).

As for MCP, the sparse Riesz condition (SRC), which is in consonance with the Irrepresentable condition (3.3) of LASSO, is supposed and works for ruling out strong columnwise correlation in the design matrix [9]. Under SRC, the Beta-min condition for MCP requires that

$$\min_{1\le j\le q}|\beta_j^0| \ge C_2\sigma\sqrt{\frac{\log p}{n}}, \tag{3.6}$$

where C_2 is a constant depending on $\mathbf{X}$.

It is worth noting that the Beta-min condition of MCP (3.6) is independent of q, and consequently, weaker than that of LASSO (3.4). As discussed in [9] and [10], it is the bias of LASSO that causes the threshold value of LASSO to be "$\sqrt{q}$ times larger" than that of MCP in magnitude. We will confirm by simulations in Section 4 that MCP does outperform LASSO in terms of detecting small nonzero regression coefficients. Next, we develop a procedure based on LASSO to overcome both the Beta-min condition and the Irrepresentable condition, and this procedure outperforms MCP and other competing methods in terms of selection accuracy across various settings.

3.2. The algorithm: Perturbed LASSO (PLA)

As for any regularization procedures, strong correlation in the design matrix and weak signal (i.e., small nonzero regression coefficients) are two formidable hindrances to selection accuracy. An intuitive but effective approach of weakening correlations is to create a perturbed design matrix $\mathbf{Z}$ by adding an $n \times p$ random matrix composed of np *iid* random entries distributed as $N(0, \tau^2)$ onto $\mathbf{X}$. Then the columnwise correlation of the perturbed design matrix reduces to below $1/(1+\tau^2)$, and consequently, the Irrepresentable condition is overcome as τ is large enough. Moreover, as shown in Section 3.3, the random perturbation improves the chance of recovering weak signal through adding an amount to the regression coefficients while the threshold is kept unchanged.

Given a set of pre-defined regularization parameters $\Lambda = \{\lambda_1, \cdots, \lambda_D\}$, the above-described scheme is formalized in the following two-step procedure, which we call Perturbed LASSO (PLA).

1. ***Inclusion:*** *Get estimates* $\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0}$ *by applying LASSO with*

$$\lambda_0 = 2\sigma\sqrt{2n \log p} \tag{3.7}$$

[2] to the original sample $(\mathbf{X}, \mathbf{Y})$*. Generate* H *independent* $n \times p$ *random matrices* $\boldsymbol{\Xi}_h$*'s* $(h = 1, \cdots, H)$ *and each of them is composed of* np *entries identically and independently distributed as* $N(0, 1)$*. Let*

$$\mathbf{Z}_h = \mathbf{X} + \tau\boldsymbol{\Xi}_h \tag{3.8}$$

$$\mathbf{W}_h = \mathbf{Y} + \tau\boldsymbol{\Xi}_h\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} \tag{3.9}$$

where τ *is the perturbation size. Create a model subspace* $\widehat{\mathcal{M}}_h^{\mathrm{PLA},\Lambda}$ $(h = 1, \cdots, H)$ *by applying LASSO to the perturbed sample* $(\mathbf{Z}_h, \mathbf{W}_h)$ *for each* $\lambda \in \Lambda$*. Thus, we get a union model space,* $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H} = \cup_{h=1}^{H}\widehat{\mathcal{M}}_h^{\mathrm{PLA},\Lambda}$.

2. ***Selection:*** *Perform model selection within* $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$ *by an information criterion based on* $(\mathbf{X}, \mathbf{Y})$.

In summary, PLA proceeds in two steps: 1) Create a model space $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$ by repeating perturbations and a set of preselected penalty coefficients; 2) Perform model selection within this space by an information criterion. In this article, we adopt RIC_c, whose advantage over other information criteria has been testified in [11]. The model subspace $\widehat{\mathcal{M}}_h^{\mathrm{PLA},\Lambda}$ $(h = 1, \cdots, H)$ generated by each perturbation includes D models and the union space $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$ includes no more than DH unique models. Therefore, in the following simulations three indices are adopted to assess the performance: the *inclusion accuracy* (P_1) to measure whether the true model is included by $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$, the *selection accuracy* (P_2) to measure whether the true model is ultimately selected and the *size* (N) to measure the number of unique models in $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$. Increasing D and H will improve inclusion accuracy in the cost of computing load.

Obviously, a large τ can remove correlation (almost) entirely, but an overly large τ will blur the relationship between Y and $(x_1, \cdots, x_p)$. Hence, the perturbation size τ needs to balance the above two ends. Next we conduct theoretical analyses on our proposed method PLA and provide theoretical guidance on the choice of the perturbation size τ and the perturbation number H.

3.3. Theory

For a perturbed sample $(\mathbf{Z}_h, \mathbf{W}_h)$ generated by (3.8) and (3.9), let $\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda} = \mathrm{argmin}_\theta(\|\mathbf{W}_h - \mathbf{Z}_h\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1)$. Split the perturbed feature matrix $\mathbf{Z}_h$ into two submatrices $\mathbf{Z}_{h,1}$ and $\mathbf{Z}_{h,2}$, which are spanned by the first q and the remaining $(p-q)$ columns, respectively. Accordingly, the matrix $\boldsymbol{\Xi}_h$ is split into $\boldsymbol{\Xi}_{h,1}$ and $\boldsymbol{\Xi}_{h,2}$. Using KKT condition [3] we derive four sufficient conditions for $\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda}$ and $\boldsymbol{\beta}^0$ to have the same sign, which are summarized in the following proposition. Let $*$ denote the elementwise product of two vectors.

Proposition 3.1. *If the following conditions hold for a scalar $\eta \in (-\infty, \infty)$ and a scalar $\delta \in (0, 1)$,*

$$|\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\,\mathrm{sign}(\boldsymbol{\beta}_1^0)| \le_c 1-\delta \quad \textit{(p-Irrepresentable)} \tag{3.10}$$

$$|\mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1}\mathbf{W}_h| \le_c 2^{-1}\delta\lambda \quad \textit{(p-Exclusion)} \tag{3.11}$$

$$\mathrm{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}(\tau\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon}) + |\boldsymbol{\beta}_1^0| >_c \eta \quad \textit{(p-Beta-min)} \tag{3.12}$$

$$\mathrm{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\big(\frac{\lambda}{2}\,\mathrm{sign}(\boldsymbol{\beta}_1^0) - \mathbf{X}'_1\boldsymbol{\varepsilon} - \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}^0)\big) <_c \eta \quad \textit{(p-Inclusion)} \tag{3.13}$$

then $\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda}$ and $\boldsymbol{\beta}^0$ have the same sign.

In the above proposition, Conditions (3.10) and (3.11) each consisting of $(p-q)$ inequalities function for excluding all uninformative features, whereas Conditions (3.12) and (3.13) each consisting of q inequalities function for including all informative features. The proof of Proposition 3.1 is deferred to Appendix. To distinguish from (3.3) and (3.4), these four conditions are named p-Irrepresentable, p-Exclusion, p-Beta-min and p-Inclusion conditions in order where p- stands for perturbed. Here it is worth pointing out that Conditions (3.12) and (3.13) are derived from $\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} * \boldsymbol{\beta}_1^0 >_c 0$ instead of $|\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} - \boldsymbol{\beta}_1^0| >_c |\boldsymbol{\beta}_1^0|$ [13], where the former is a sufficient and necessary condition for $\mathrm{sign}(\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}) = \mathrm{sign}(\boldsymbol{\beta}_1^0)$, whereas the latter is a sufficient but unnecessary condition.

Next we briefly discuss how PLA overcomes the Irrepresentable condition and the Beta-min condition. The columnwise correlation of $\mathbf{Z}_h$ will decrease to zero as τ goes up to ∞ such that all columns are uncorrelated and the Irrepresent condition is met (in asymptotic sense). In more detail, each element of the left-hand side of (3.10) is bounded by $q\tau^{-2}$ in probability (the proof is provided

in Appendix: Lemmas 1 and 2). Thus, as the perturbation size $\tau > \sqrt{q}$, the Irrepresentable condition is satisfied (asymptotically).

The perturbed Beta-min condition (3.12) differs from the Beta-min condition (3.4) by the q-dimensional random vector $\text{sign}(\boldsymbol{\beta}_1^0) * \tau(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon}$ (referred to $\mathbf{b}_h = (b_{h,1}, \cdots, b_{h,q})'$ next) while η is the same. Compare the two Beta-min conditions

$$\begin{aligned} \min_{1\le j\le q}\left(|\beta_j^0|\right) &\ge \eta \qquad \text{LASSO;} \\ \min_{1\le j\le q}\left(b_{h,j}+|\beta_j^0|\right) &\ge \eta \qquad \text{PLA} \end{aligned}$$

where $b_{h,j}$ is asymptotically distributed as $N(0, \sigma_{b_j}^2)$ with σ_{b_j} is inversely proportional to τ (more details are provided in the proof of Theorem 1). Consequently, when the Beta-min condition (3.4) is violated (i.e., $|\beta_j^0| \le \eta$ for some j), the random term $b_{h,j}$ will push $|\beta_j^0|$ above η, and a small τ is preferred. Theoretically, no matter how small a nonzero regression coefficient, it can always be recovered by performing perturbations many times (i.e., large H).

Let κ_0^2 be the Restricted Eigenvalue constant as defined in Eq (16) of [2]. Concerning the inclusion consistency of PLA we have the following conclusion.

Theorem 1. *Let $\eta = 4\sigma\sqrt{2q(\log p)/(n\kappa_0^2)}$. Suppose $\tau \ge \sqrt{8}q^{3/4}$, then the probability that the true model M_0 is in the union model subspace $\widehat{\mathcal{M}}^{\text{PLA},\Lambda,H}$ is bounded below by*

$$Pr(M_0 \in \widehat{\mathcal{M}}^{\text{PLA},\Lambda,H}) \ge 1 - \left((1-\prod_{j=1}^{q} p_j)^H + C_0\Phi^c(\sqrt{2\log p})\right), \tag{3.14}$$

where $p_j = \Phi^c\left(\tau\sqrt{n}(\eta - |\beta_j^0|)/\sigma\right)$, Φ^c is the upper tail probability of standard normal distributions and C_0 is a generic constant.

Theorem 1 is about inclusion consistency, i.e., the probability that the true model is included by the subspace $\widehat{\mathcal{M}}^{\text{PLA},\Lambda,H}$ goes to 1. The proof of Theorem 1 is delayed to the appendix.

There are several consequences of Theorems 1.

1. As indicated by Theorem 1 and discussed above, a large τ is desired for overcoming the Irrepresentable condition, while a small τ is preferred for overcoming the Beta-min condition. We adopt

$$\tau_0 = \sqrt{\frac{n}{2\log p}}, \tag{3.15}$$

   which overcomes the Irrepresentable condition entirely under the sparsity assumption, i.e., $q \ll n/\log p$.
2. The inclusion probability (lower bound) by PLA (3.14) decreases in (σ, p and q) but increases in (n, κ_0^2, $|\boldsymbol{\beta}_1^0|$ and H). Hence, if there exist small $|\beta_j^0|$; $j = 1, \cdots, q$, strong correlation in $\mathbf{X}$ (small κ_0^2), large p and/or large q, then a large H is needed to achieve high selection accuracy.

3. In the context of "strong correlation and strong signal" where the Beta-min condition (3.4) is met but the Irrepresentable condition (3.3) is not, PLA can obtain 100% inclusion accuracy by one perturbation, while the inclusion accuracy of LASSO is 0. This is to be testified by simulations in Section 4.

The computing cost and selective performances of PLA will be examined by simulations in Section 4 where we assume that the true model is linear, sparse and included in the candidate model space. However, it is impractical to suppose the existence of a sparse and linear true model in real data, and this will be investigated in Section 5.

4. Simulations

For the data-generating processes, each row of the $n\times p$ design matrix $\mathbf{X} = (X_{i,j})$ is independently generated from $N(\mathbf{0}, \mathbf{\Sigma}_{p\times p})$, where $\mathbf{0}$ denotes the p-dimensional vector of 0's and $\mathbf{\Sigma}_{p\times p} = (\rho^{|j_1-j_2|})$ denotes a $p \times p$ covariance matrix with j_1 and j_2 $(j_1, j_2 = 1, \cdots, p)$ denoting the row and column indices, respectively. Two values of ρ (0.6 and 0.9) are examined. The responses are generated from the true model (2.2) with $\sigma = 1$. Among $\boldsymbol{\beta}^0$, q randomly selected coefficients $(\beta^0_{j_1}, \cdots, \beta^0_{j_q})$ ($q = 9$ or 12) are assigned nonzero values by the following rule:

$$\beta^0_{j_1} = \cdots = \beta^0_{j_k} = 2\alpha; \qquad \beta^0_{j_{k+1}} = \cdots = \beta^0_{j_q} = 2 \tag{4.1}$$

where the shrinkage factor $\alpha = 0.6$ or 0.3, and all other β_j's are assigned 0. Thus, the minimal nonzero regression coefficient $\beta^0_{\min} = 2\alpha$. Four values of k (1, 3, 5, 7) are examined. As ρ and k increase and α decreases, the Beta-min and the Irrepresentable conditions are more likely to be violated and the performances of all procedures will deteriorate.

In all simulations, σ^2 is assumed unknown and the penalty coefficients, $\Lambda = \{\lambda_1, \cdots, \lambda_{100}\}$ that are the default candidates in the R package *ncvreg* are utilized. For each setting, 100 replications (realizations of sample data) are performed. In PLA and BoLASSO (BLA), H perturbations or bootstrappings ($H = 10$ or 1000) are performed.

The following six procedures are compared.

- **PLA-RIC$_c$:** Create a union model space $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H} = \cup_{h=1}^{H}\widehat{\mathcal{M}}_h^{\mathrm{PLA},\Lambda}$ by employing the procedure PLA introduced in Section 3.2. At each perturbation, the model subspace $\widehat{\mathcal{M}}_h^{\mathrm{PLA},\Lambda}$ is created by applying LASSO with each $\lambda \in \Lambda$, as implemented by the R package *ncvreg*, to the perturbed data $(\mathbf{Z}_h, \mathbf{W}_h)$. Perform model selection within $\widehat{\mathcal{M}}^{\mathrm{PLA},\Lambda,H}$ that includes no more than DH unique models by RIC$_c$ based on the original data $(\mathbf{X}, \mathbf{Y})$. The variance σ^2 is estimated by Least Squares based on the model selected by LASSO, in which the regularization parameter λ is tuned by 5-fold cross-validations from Λ.
- **BLA-RIC$_c$:** The model subspace $\widehat{\mathcal{M}}^{\mathrm{BLA},\Lambda,H}$ is created by BoLASSO with each $\lambda \in \Lambda$, as implemented by the package *mht*. In BoLASSO, the selection frequency of each feature is calculated based on H bootstrapping

samples and the threshold frequency is set at the default value 100%. Then, RIC_c is employed to perform model selection within $\widehat{\mathcal{M}}^{\text{BLA},\Lambda,H}$, which includes no more than D unique models. Different from PLA, the size of $\widehat{\mathcal{M}}^{\text{BLA},\Lambda,H}$ does not increase with H.

- **LAS-RIC$_c$:** Apply LASSO with each $\lambda \in \Lambda$ to the original data $(\mathbf{X}, \mathbf{Y})$ to produce $D = 100$ models, which compose a model subspace $\widehat{\mathcal{M}}^{\text{LA},\Lambda}$, as implemented by the package *ncvreg*. Then, RIC_c is employed to perform model selection within $\widehat{\mathcal{M}}^{\text{LA},\Lambda}$ that includes no more than D unique models.
- **MCP-RIC$_c$:** Apply MCP with each $\lambda \in \Lambda$ to the original data $(\mathbf{X}, \mathbf{Y})$ to produce $D = 100$ models, which compose a model subspace $\widehat{\mathcal{M}}^{\text{MCP},\Lambda}$, as implemented by the package *ncvreg*. Then, RIC_c is employed to perform model selection within $\widehat{\mathcal{M}}^{\text{MCP},\Lambda}$ that includes no more than D unique models.
- **CV-LAS:** Model selection is done in two steps: (1) tune the best λ from Λ by 5-fold cross-validations; (2) apply LASSO with the tuned penalty to $(\mathbf{X}, \mathbf{Y})$.
- **CV-MCP:** Model selection is done in two steps: (1) tune the best λ from Λ by 5-fold cross-validations; (2) apply MCP with the tuned penalty to $(\mathbf{X}, \mathbf{Y})$.

The first four procedures (PLA-RIC_c, BLA-RIC_c, LAS-RIC_c and MCP-RIC_c) perform model selection in two steps (1) *Inclusion:* create a model subspace; (2) *Selection:* model selection within the subspace by an information criterion. Thus, their performances are assessed by *inclusion accuracy* (P_1) that is the proportion of replications where the method includes the true model M_0, and *selection accuracy* (P_2) that is the proportion of replications where the method selects the true model M_0. Undoubtedly, $P_2 \leq P_1$ for all procedures. Though our ultimate goal is model selection, we want to emphasize that, rather than the selection accuracy, the inclusion accuracy is more adequate for assessing whether a procedure is able to overcome the Beta-min Condition because a misused information criterion may result in misidentifying the true model. Therefore, in the following table the inclusion accuracy is highlighted by bold fonts and square brackets. Additionally, we report the size of model subspace (N) that is the number of unique models, the system time consumed for creating model subspace (T_1) and the system time for model selection within the model subspace (T_2). Consequently, the sum of T_1 and T_2 measures the total computational time of each procedure. The unit of T_1 and T_2 is second.

As a reference, we also examine the selective performance of LASSO and MCP equipped with Cross-validations (CV-LAS and CV-MCP). Though 5-fold or 10-fold CV is often adopted to tune the penalty level, neither of them can guarantee that optimal λ is tuned [12]. Therefore, as LASSO or MCP misidentifies the true model, it may be due to mistuning of the optimal penalty. Even worse, the bias of LASSO often leads to the choice of an overly small λ, which causes overly large false positives, and consequently, extremely low selection accuracy. Therefore, we strongly recommend the strategy of "screen by LASSO (i.e., creating a model

space by a set of regularization parameters Λ) and select by information criteria" [11] to replace the more routine procedure "tune by CV and select by LASSO", and the latter is more computationally burdensome than the former as shown in the following simulations. For CV-LAS and CV-MCP, only selection accuracy (P) and total computational time (T) are reported.

The results are as follows:

As shown in Table 1 and Table 2, LAS-RIC$_c$ beats CV-LAS in both computational efficiency and selection accuracy across various settings. The same pattern is observed between MCP-RIC$_c$ and CV-MCP, though the gap is not so big as LASSO.

It is worth noting that in the case of moderate correlation $\rho = 0.6$, the inclusion accuracy (P_1) of MCP-RIC$_c$ does not change with q, but the inclusion accuracy of LAS-RIC$_c$ goes down greatly as q goes up from 9 to 12. This is caused by a fact that the threshold in the Beta-min condition of LASSO (3.4) increases with q, but the threshold of MCP (3.6) does not. However, in the strong correlation case ($\rho = 0.9$), the performance of MCP-RIC$_c$ worsens as the true model size q increases from 9 to 12 because a large q brings up the chance of violating the Irrepresentable condition or the sparse Riesz condition, especially when strong correlations exist in $\mathbf{X}$.

As shown in Table 1 and Table 2, in the nicest case where the columnwise correlation is moderate ($\rho = 0.6$) and the nonzero coefficients are "sufficiently large" ($\alpha = 0.6$ such that $\beta^0_{\min} = 1.2$), the inclusion accuracy of LAS-RIC$_c$ is above 88% while all other two-step procedures (i.e., MCP-RIC$_c$, BLA-RIC$_c$ and PLA-RIC$_c$) achieve above 99% inclusion accuracy. However, all six procedures' performances deteriorate gradually as ρ and q increase and the nonzero coefficients decrease in size, which cause the Irrepresentable condition and the Beta-min condition less likely to hold. Overall, PLA with $H = 1000$ outperforms all other competitors in all cases. In the worst case ($q = 12$, $\rho = 0.9$, $q_s = 7$ and $\beta^0_{\min} = 0.6$), the inclusion accuracy (P_1) of LAS-RIC$_c$ and MCP-RIC$_c$ are 0 and 11%, respectively. Our procedure PLA elevates the inclusion accuracy up to 34% as H=10 and up to 61% as $H = 1000$. Another case worthy of attention is that PLA achieves almost 100% inclusion and selection accuracy by performing only $H = 10$ perturbations when both signal and correlation are strong ($\rho = 0.9$ and $\alpha = 0.6$), whereas the inclusion accuracy of LASSO is below 5%. This demonstrates the power of PLA on overcoming the Irrepresentale condition. This confirmed a conclusion implied by Theorem 1: the computing load is mainly driven by small nonzero β^0_j when the true model is sparse. In conclusion, LASSO requires stronger Beta-min condition than MCP, and consequently, suffers lower selection accuracy. However, PLA lowers the threshold down to 0 (as H is large enough) and accomplishes higher selection accuracy than MCP.

Between the two resampling-based methods BoLASSO and PLA, a notable difference is that the performance of PLA always improves with increasing H that supports the conclusion of Theorem 1, but BoLASSO does not. In particular, when there exist some "small" coefficients ($\alpha = 0.3$ such that $\beta^0_{\min} = 0.6$), the inclusion accuracy of BoLASSO worsens as H goes up because a large H increases the chance of misselecting relevant features in some perturbations.

TABLE 1. *Comparison of Inclusion Accuracy, Selection Accuracy and Computational Time*

q_s	CV-LAS		CV-MCP		LAS-RIC$_c$					MCP-RIC$_c$					BLA-RIC$_c$ (H=10)					BLA-RIC$_c$ (H=1000)					PLA-RIC$_c$ (H=10)					PLA-RIC$_c$ (H=1000)				
	P	T	P	T	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2
													$q = 9$, $\rho = 0.6$, $\alpha = 0.6$																					
1	0	7	0.65	6	30	**[0.94]**	0.69	0.94	0.08	23	**[1]**	0.61	1	0.05	14	**[0.98]**	17	0.98	0.03	8	**[1]**	1651	1	0.01	55	**[1]**	5	1	0.11	699	**[1]**	1284	1	1
3	0	7	0.63	7	31	**[0.97]**	0.71	0.97	0.09	24	**[1]**	0.67	1	0.06	15	[1]	18	1	0.03	9	**[1]**	1657	1	0.02	45	**[1]**	5	1	0.09	576	**[1]**	1323	1	1
5	0	8	0.66	8	34	**[0.88]**	0.81	0.88	0.1	25	**[1]**	0.75	1	0.06	17	**[0.99]**	19	0.99	0.04	9	**[1]**	1771	1	0.02	43	**[1]**	6	1	0.09	733	**[1]**	1446	1	1
7	0	7	0.69	8	35	**[0.93]**	0.72	0.93	0.11	27	**[1]**	0.76	1	0.07	18	**[0.99]**	17	0.99	0.04	9	**[1]**	1657	1	0.02	52	**[1]**	5	1	0.11	1058	**[1]**	1287	1	2
													$q = 9$, $\rho = 0.6$, $\alpha = 0.3$																					
1	0	7	0.52	6	30	**[0.80]**	0.68	0.79	0.08	22	**[1]**	0.63	0.99	0.05	13	**[0.98]**	18	0.97	0.03	7	**[1]**	1651	0.99	0.01	53	**[1]**	5	0.99	0.1	616	**[1]**	1296	0.99	1
3	0	7	0.23	7	31	**[0.79]**	0.69	0.79	0.09	24	**[1]**	0.69	1	0.06	14	**[0.94]**	18	0.94	0.03	8	**[1]**	1625	1	0.02	41	**[1]**	5	1	0.08	481	**[1]**	1299	1	0.9
5	0	8	0.15	8	34	**[0.71]**	0.76	0.71	0.1	26	**[1]**	0.79	1	0.07	16	**[0.95]**	19	0.95	0.03	9	**[0.99]**	1777	0.99	0.02	36	**[1]**	6	1	0.08	670	**[1]**	1406	1	1
7	0	7	0.06	9	36	**[0.65]**	0.72	0.65	0.11	28	**[1]**	0.8	1	0.07	19	**[0.93]**	18	0.93	0.04	10	**[0.99]**	1726	0.99	0.02	45	**[1]**	5	1	0.09	1020	**[1]**	1383	1	2
													$q = 9$, $\rho = 0.9$, $\alpha = 0.6$																					
1	0	26	0.3	7	32	**[0.05]**	2	0.05	0.08	19	**[0.82]**	0.67	0.82	0.04	13	**[0.73]**	21	0.73	0.03	9	**[1]**	1854	1	0.02	73	**[1]**	6	1	0.16	1784	**[1]**	1356	1	4
3	0	25	0.09	7	33	**[0.08]**	2	0.08	0.09	20	**[0.76]**	0.67	0.76	0.04	13	**[0.74]**	21	0.74	0.03	9	**[0.99]**	1822	0.99	0.02	63	**[1]**	6	0.99	0.14	1855	**[1]**	1333	0.99	4
5	0	26	0.12	8	35	**[0.06]**	2	0.06	0.1	22	**[0.60]**	0.75	0.6	0.05	14	**[0.74]**	21	0.74	0.03	9	**[1]**	1884	1	0.02	62	**[1]**	6	1	0.14	2175	**[1]**	1376	1	5
7	0	26	0.08	8	37	**[0.08]**	2	0.08	0.11	24	**[0.53]**	0.81	0.53	0.06	14	**[0.76]**	22	0.76	0.03	9	**[0.96]**	1838	0.96	0.02	77	**[1]**	6	1	0.17	2919	**[1]**	1355	1	6
													$q = 9$, $\rho = 0.9$, $\alpha = 0.3$																					
1	0	26	0.25	7	31	**[0.05]**	2	0.05	0.08	19	**[0.79]**	0.7	0.79	0.04	12	**[0.69]**	21	0.68	0.03	8	**[0.49]**	1813	0.48	0.01	75	**[0.92]**	6	0.9	0.16	1848	**[0.98]**	1337	0.94	4
3	0	26	0.12	7	34	**[0.03]**	2	0.03	0.09	20	**[0.59]**	0.71	0.59	0.05	13	**[0.46]**	22	0.46	0.03	8	**[0.12]**	1915	0.12	0.02	64	**[0.69]**	6	0.67	0.14	1843	**[0.89]**	1448	0.84	4
5	0	22	0.02	7	35	**[0.03]**	2	0	0.1	22	**[0.32]**	0.71	0.32	0.05	15	**[0.35]**	20	0.34	0.03	8	**[0.01]**	1815	0.01	0.01	61	**[0.61]**	5	0.57	0.13	2209	**[0.79]**	1305	0.71	5
7	0	22	0.01	8	37	**[0.04]**	2	0.04	0.11	24	**[0.23]**	0.81	0.23	0.06	16	**[0.31]**	21	0.3	0.03	9	**[0.03]**	1834	0.03	0.02	71	**[0.59]**	6	0.5	0.16	2794	**[0.78]**	1353	0.59	6

$n = 250$, $p = 2000$. The feature matrix is generated by $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \rho^{|j_1 - j_2|}$. Among the q nonzero $\beta_j's$, q_s coefficients are assigned value 2α and the remainings are assigned 2. The 1st column lists the value of q_s. The other columns show the computational time of creating model subspace (T_1) and the time for model selection within this subspace by RIC$_c$ (T_2), the size of model subspace (N); the proportion of including the true model by the model subspace (P_1), and the proportion of selecting the true model by RIC$_c$(P_2). As for PLA and BoLASSO, $H = 10$ or 1000 perturbations or bootstrappings are performed. The unit of T_1 and T_2 is second.

TABLE 2. *Comparison of Inclusion Accuracy, Selection Accuracy and Computational Time*

q_s	CV-LAS		CV-MCP		LAS-RIC$_c$					MCP-RIC$_c$					BLA-RIC$_c$ (H=10)					BLA-RIC$_c$ (H=1000)					PLA-RIC$_c$ (H=10)					PLA-RIC$_c$ (H=1000)				
	P	T	P	T	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2	N	P_1	T_1	P_2	T_2
$q = 12,\ \rho = 0.6,\ \alpha = 0.6$																																	
1	0	9	0.71	6	33	**[0.64]**	0.83	0.64	0.09	22	[1]	0.66	0.99	0.05	16	**[0.92]**	19	0.92	0.03	9	[1]	1779	0.99	0.02	90	[1]	6	1	0.2	2932	[1]	1431	0.99	6
3	0	8	0.61	6	34	**[0.65]**	0.78	0.65	0.1	23	[1]	0.62	1	0.06	17	**[0.97]**	18	0.97	0.03	11	**[0.99]**	1661	0.99	0.02	83	[1]	6	1	0.17	2104	[1]	1312	1	4
5	0	8	0.74	7	36	**[0.68]**	0.81	0.68	0.1	25	[1]	0.67	1	0.06	19	**[0.97]**	18	0.97	0.04	11	[1]	1639	1	0.02	84	[1]	6	1	0.18	2156	[1]	1318	1	4
7	0	9	0.69	7	37	**[0.72]**	0.83	0.72	0.11	26	[1]	0.68	0.98	0.07	19	**[0.91]**	19	0.9	0.04	11	[1]	1747	0.99	0.02	85	[1]	6	0.98	0.18	2348	[1]	1525	0.98	5
$q = 12,\ \rho = 0.6,\ \alpha = 0.3$																																	
1	0	9	0.46	7	33	**[0.5]**	0.83	0.5	0.09	22	[1]	0.68	0.98	0.05	16	**[0.91]**	19	0.9	0.03	9	[1]	1800	0.99	0.02	91	[1]	6	0.98	0.2	2921	[1]	1490	0.98	6
3	0	8	0.35	7	35	**[0.39]**	0.82	0.39	0.1	24	[1]	0.71	1	0.06	17	**[0.89]**	19	0.89	0.04	10	**[0.96]**	1755	0.96	0.02	80	[1]	6	1	0.17	1924	[1]	1443	1	4
5	0	8	0.14	8	35	**[0.47]**	0.8	0.47	0.1	25	[1]	0.75	1	0.06	18	**[0.9]**	19	0.9	0.04	11	**[0.98]**	1797	0.98	0.02	74	[1]	6	1	0.16	1852	[1]	1435	0.99	4
7	0	9	0.11	8	37	**[0.28]**	0.82	0.28	0.11	27	**[0.99]**	0.74	0.99	0.07	20	**[0.79]**	19	0.79	0.05	12	**[0.97]**	1818	0.96	0.03	76	[1]	6	0.99	0.17	2281	[1]	1462	0.99	5
$q = 12,\ \rho = 0.9,\ \alpha = 0.6$																																	
1	0	29	0.13	8	36	**[0.01]**	3	0.01	0.1	21	**[0.67]**	0.79	0.67	0.05	19	**[0.6]**	22	0.6	0.04	11	[1]	1858	1	0.02	125	[1]	6	1	0.28	5766	[1]	1353	1	13
3	0	29	0.11	8	37	**[0.01]**	3	0.01	0.1	22	**[0.68]**	0.73	0.68	0.05	19	**[0.5]**	22	0.5	0.04	12	**[0.98]**	1863	0.98	0.02	125	[1]	6	1	0.3	5519	[1]	1366	1	12
5	0	30	0.05	8	40	**[0]**	3	0	0.11	23	**[0.46]**	0.8	0.46	0.05	20	**[0.67]**	23	0.67	0.04	12	**[0.95]**	1920	0.95	0.02	127	[1]	6	1	0.3	5845	[1]	1408	1	14
7	0	29	0.03	8	40	**[0.01]**	3	0.01	0.11	24	**[0.45]**	0.79	0.45	0.05	20	**[0.58]**	22	0.58	0.04	12	**[0.98]**	1869	0.98	0.02	128	[1]	6	1	0.31	6172	[1]	1370	1	14
$q = 12,\ \rho = 0.9,\ \alpha = 0.3$																																	
1	0	28	0.14	8	36	**[0]**	3	0	0.1	22	**[0.69]**	0.75	0.69	0.05	18	**[0.44]**	22	0.44	0.04	10	**[0.47]**	1858	0.47	0.02	129	**[0.86]**	6	0.85	0.3	5846	**[0.99]**	1350	0.98	12
3	0	28	0.06	8	37	**[0]**	3	0	0.1	21	**[0.51]**	0.74	0.5	0.05	18	**[0.34]**	22	0.33	0.04	10	**[0.03]**	1866	0.03	0.02	122	**[0.62]**	6	0.58	0.29	5227	**[0.89]**	1334	0.78	11
5	0	28	0.01	8	39	**[0.01]**	3	0.01	0.11	23	**[0.23]**	0.81	0.23	0.05	19	**[0.16]**	22	0.15	0.04	10	**[0.02]**	1917	0.02	0.02	124	**[0.38]**	6	0.36	0.29	5773	**[0.77]**	1401	0.69	13
7	0	28	0	9	41	**[0]**	3	0	0.12	25	**[0.11]**	0.87	0.11	0.06	20	**[0.16]**	23	0.16	0.04	11	**[0.01]**	1971	0.01	0.02	126	**[0.34]**	6	0.31	0.31	6648	**[0.61]**	1461	0.52	16

$n = 250$, $p = 2000$. The feature matrix is generated by $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \rho^{|j_1 - j_2|}$. Among the q nonzero $\beta_j's$, q_s coefficients are assigned value 2α and the remainings are assigned 2. The 1st column lists the value of q_s. The other columns show the computational time of creating model subspace (T_1) and the time for model selection within this subspace by RIC$_c$ (T_2), the size of model subspace (N); the proportion of including the true model by the model subspace (P_1), and the proportion of selecting the true model by RIC$_c$(P_2). As for PLA and BoLASSO, $H = 10$ or 1000 perturbations or bootstrappings are performed. The unit of T_1 and T_2 is second.

Overall, PLA achieves higher inclusion (P_1) and selection accuracy (P_2) than BoLASSO while costing less computing time. Examining T_1 and T_2, it is clear that the first step of BoLASSO and PLA accounts for most computing load. For example, in the worst case ($q = 12$, $\rho = 0.9$, $q_s = 7$ and $\beta^0_{\min} = 0.6$), the first step of PLA used up 1461 seconds on creating a space including 6648 unique models and the second step, model selection by RIC_c, only used up 16 seconds.

Additional simulation results of various (n, p, ρ, $\boldsymbol{\beta}^0$, Λ) further supporting the robustness and flexibility of PLA are available upon request. Similar patterns to the ones presented in this section are observed, further demonstrating the success of our proposed method. The often sizable advantage of PLA over its competitors, especially when there exist small regression coefficients and strong correlation among features, makes PLA a powerful tool in high dimensional variable selection.

We also examined the performance as perturbation partners with MCP, which is referred to as "Perturbed MCP", and the simulation results are available upon request. As demonstrated by the simulation, Perturbed MCP improves the inclusion and selection accuracy of MCP but not so substantial as Perturbed LASSO. Furthermore, Perturbed LASSO outperforms Perturbed MCP across all settings.

Due to constraints of computing resource, it is hard to recover overly small regression coefficients, which may not contribute to prediction in the context of "large p small n". However, the harvest in selection and prediction is always directly proportional to pay in computation as shown in the following real data example.

5. Real data application

In this section we analyze the dataset, *riboflavin* that is about vitamin B2 production and publicly available through the R package, *hdi* (www.r-project.org) [5]. The data comprise of 71 observations on a single real-valued response variable that is the logarithm of the B2 production rate, and $p = 4088$ features measuring the logarithm of the expression level of 4088 genes.

First of all, we compare the predictive performances of the four two-step procedures (PLA-RIC_c, BLA-RIC_c, LAS-RIC_c, MCP-RIC_c) studied in Section 4. In PLA-RIC_c and BLA-RIC_c, $H = 1000$ bootstrappings or perturbations are performed. The comparison is done in three steps. First, the 71 observations are divided into an evaluation set of size N_e ($N_e = 2, \cdots, 5$) and an estimation set of size $n = 71 - N_e$. Second, the four competitors each develop predictive models from the estimation set, which are used to make predictions on the evaluation set. Finally, we perform the above two steps 100 times, selecting the evaluation sets at random each time, and get $N = 100$ average prediction errors for each method. The average mean squared error (MSE) for each of the four procedures are displayed in Table 3. From Table 3, PLA-RIC_c yields the best overall predictive accuracy.

Next, we apply PLA-RIC_c to the whole sample and select six genes: ARGF_at, XHLB_at, YDDK_at, YEBC_at, YOAB_at and YXLD_at. A similar analysis was

TABLE 3
Comparison of Prediction Errors for Riboflavin Data

N_e	LAS-RIC$_c$	MCP-RIC$_c$	BLA-RIC$_c$	PLA-RIC$_c$
2	0.333(0.044)	0.211(0.032)	0.568(0.073)	0.205(0.036)
3	0.299(0.030)	0.228(0.030)	0.547(0.056)	0.198(0.018)
4	0.248(0.023)	0.223(0.024)	0.486(0.048)	0.211(0.030)
5	0.311(0.027)	0.238(0.026)	0.520(0.049)	0.225(0.022)

For each method, the table shows the average of mean square error (MSE) and standard error (SE) (in parentheses). Note that the SEs reported are not the standard errors in the i.i.d. case.

TABLE 4
Gene Selection for Riboflavin Data

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
The linear regression analysis output based on the model selected by PLA-RIC$_c$.				
Multiple R-squared: 0.8906, Adjusted R-squared: 0.8804				
(Intercept)	14.44803	2.14293	6.742	0.0000000052614127
ARGF_at	-0.31695	0.04075	-7.778	0.0000000000792608
XHLB_at	0.29052	0.05323	5.458	0.0000008343452216
YDDK_at	-0.29773	0.08510	-3.499	0.0008560000000000
YEBC_at	-0.83533	0.19959	-4.185	0.0000886548546881
YOAB_at	-1.22380	0.16385	-7.469	0.0000000002785447
YXLD_at	-0.37530	0.03865	-9.710	0.0000000000000332
The linear regression analysis output based on the model presented in [5].				
Multiple R-squared: 0.6763, Adjusted R-squared: 0.6618				
(Intercept)	16.32195	2.40206	6.795	0.000000003520
LYSC_at	-1.11713	0.31720	-3.522	0.000777000000
YOAB_at	-1.11722	0.25514	-4.379	0.000042885043
YXLD_at	-0.47970	0.06287	-7.630	0.000000000112

done by [5] using the R package *hdi* and three genes: LYSC_at, YOAB_at and YXLD_at are selected. The linear regression analysis based on the two models is done in R and the output is presented in Table 4. As shown Table 4, PLA-RIC$_c$ recovered regression coefficients whole absolute values are below 0.4, while the other approach only recovered coefficients whole absolute values are above 0.4.

6. Discussion

In high dimensional feature selection problems, strong correlation and weak signal are the two main hindrances to exact recovery of informative features. The method developed in this paper offers a solution by adding perturbations to the design matrix. As confirmed by the simulation, PLA achieves substantial advantage over other methods on selection accuracy. At the same time, our method performs well in prediction when the true model has unknown form or may be excluded from the candidate model space as demonstrated in the real data example.

Another great difficulty in high dimensional feature selection problems is the immense computing load caused by huge model space. However, literature has focused on the tradeoff between parsimony and goodness-of-fit, while the trade-off between selection accuracy and computing efficiency has been largely unex-

plored. From a theoretical perspective, we investigate this tradeoff by establishing a quantitative relationship between selection consistency and computation, and provide some guidance on how to balance selection with computation. Further investigation is necessary to generalize the proposed method to nonlinear regression models.

Appendix A: Proof

Firstly, we prove some inequalities used in the proof of Proposition 3.1 and Theorem 1.

A symmetric positive definite $q \times q$ matrix $\mathbf{\Sigma}_{11}$ can be decomposed as ([6] p449) $\mathbf{\Sigma}_{11} = \sum_{k=1}^{q} \Lambda_k \mathbf{A}_k$, where $\mathbf{A}_1, \cdots, \mathbf{A}_q$ represent $q \times q$ symmetric and idempotent matrices such that $\mathbf{A}'_{k_1}\mathbf{A}_{k_2} = \mathbf{0}_{qq}$ for $k_1 \neq k_2$.

Let $\mathbf{D}_{11} = \sum_{k=1}^{q} \frac{\Lambda_k}{(\tau^2+\Lambda_k)}\mathbf{A_k}$, then

$$(\mathbf{\Sigma}_{11} + \tau^2 \mathbf{I}_{q,q})^{-1} = \frac{1}{\tau^2}\big(\mathbf{I}_{q,q} - \mathbf{D}_{11}\big). \tag{A.1}$$

Similarly, the matrix $\mathbf{\Sigma}_{11}^n = \mathbf{X}_1'\mathbf{X}_1/n$ can be decomposed as $\mathbf{\Sigma}_{11}^n = \sum_{k=1}^{q} \Lambda_k^n \mathbf{A}_k^n$ where $\mathbf{A}_1^n, \cdots, \mathbf{A}_q^n$ represent $q \times q$ symmetric and idempotent matrices such that $(\mathbf{A}_{k_2}^n)'\mathbf{A}_{k_2}^n = \mathbf{0}_{qq}$ for $k_1 \neq k_2$. Let $\mathbf{D}_{11}^n = \sum_{k=1}^{q} \frac{\Lambda_k^n}{(\tau^2+\Lambda_k^n)}\mathbf{A}_k^n$ and $\mathbf{R}_{11}^n = \mathbf{X}_1'\mathbf{X}_1 + n\tau^2\mathbf{I}_{q,q}$, then

$$\big(E(\mathbf{Z_1'Z_1})\big)^{-1} = (\mathbf{R}_{11}^n)^{-1} = \frac{1}{n\tau^2}\big(\mathbf{I}_{q,q} - \mathbf{D}_{11}^n\big). \tag{A.2}$$

Lemma 1. *It holds that* $\frac{q}{\tau^2}(1-\frac{q}{\tau^2}) \leq \operatorname{sign}(\boldsymbol{\beta}_1')(\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1) \leq \frac{q}{\tau^2}$ *and* $\frac{1}{\tau^2}(1-\frac{q}{\tau^2}) \leq \mathbf{0}'_{(j)}(\mathbf{\Sigma_{11}} + \tau^2\mathbf{I_{qq}})^{-1}\mathbf{0}_{(j)} \leq \frac{1}{\tau^2}$.
Let $\boldsymbol{\gamma_1} = (\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1)$ *and* $\boldsymbol{\gamma_2} = (\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}\mathbf{0}_{(j)}$ *where* $\mathbf{0}_{(j)}$ *is* q*-dimensional vector of 0's except the* j*-th entry being 1, then*

$$\frac{q}{\tau^4}(1-\frac{2q}{\tau^2}) \leq \|\boldsymbol{\gamma}_1\|_2^2 \leq \frac{q}{\tau^4} \text{ and } \frac{1}{\tau^4}(1-\frac{2q}{\tau^2}) \leq \|\boldsymbol{\gamma}_2\|_2^2 \leq \frac{1}{\tau^4}. \tag{A.3}$$

The above conclusions continue to hold with $\mathbf{\Sigma}_{11}$ *replaced by* $\mathbf{\Sigma}_{11}^n$.

Proof. First of all, we have

$$\begin{aligned}
\operatorname{sign}(\boldsymbol{\beta}_1')\mathbf{D}_{11}\operatorname{sign}(\boldsymbol{\beta}_1) &\leq \sum_{k=1}^{q} \frac{\Lambda_k}{(\tau^2+\Lambda_1)}\operatorname{sign}(\boldsymbol{\beta}_1')\mathbf{A_k}\operatorname{sign}(\boldsymbol{\beta}_1) \\
&\leq \frac{q\Lambda_q}{(\tau^2+\Lambda_1)} \leq \frac{q^2}{\tau^2}; \\
\operatorname{sign}(\boldsymbol{\beta}_1')\mathbf{D}_{11}\operatorname{sign}(\boldsymbol{\beta}_1) &\geq \sum_{k=1}^{q} \frac{\Lambda_k}{(\tau^2+\Lambda_q)}\operatorname{sign}(\boldsymbol{\beta}_1')\mathbf{A_k}\operatorname{sign}(\boldsymbol{\beta}_1) \\
&\geq \frac{q\Lambda_1}{(\tau^2+\Lambda_q)} \geq 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbf{0}'_{(j)}\mathbf{D}_{11}\mathbf{0}_{(j)} &\leq \sum_{k=1}^{q}\frac{\Lambda_k}{(\tau^2+\Lambda_1)}\mathbf{0}'_{(j)}\mathbf{A_k}\mathbf{0}_{(j)} \leq \frac{\Lambda_q}{(\tau^2+\Lambda_1)} \leq \frac{q}{\tau^2};\\ \mathbf{0}'_{(j)}\mathbf{D}_{11}\mathbf{0}_{(j)} &\geq \sum_{k=1}^{q}\frac{\Lambda_k}{(\tau^2+\Lambda_q)}\mathbf{0}'_{(j)}\mathbf{A_k}\mathbf{0}_{(j)} \geq \frac{\Lambda_1}{(\tau^2+\Lambda_q)} \geq 0.\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{q}{\tau^2}(1-\frac{q}{\tau^2}) \leq \frac{q(\tau^2-q+\Lambda_1)}{\tau^4} \leq \frac{q}{\tau^2}(1-\frac{\Lambda_q}{\tau^2+\Lambda_1})\\ &\leq \operatorname{sign}(\boldsymbol{\beta}'_1)(\boldsymbol{\Sigma}_{11}+\tau^2\mathbf{I}_{q,q})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1) \leq \frac{q}{\tau^2}\frac{\tau^2+\Lambda_q-\Lambda_1}{\tau^2+\Lambda_q} \leq \frac{q}{\tau^2};\\ &\frac{1}{\tau^2}(1-\frac{q}{\tau^2}) \leq \frac{(\tau^2-q+\Lambda_1)}{\tau^4} \leq \frac{q}{\tau^2}(1-\frac{\Lambda_q}{\tau^2+\Lambda_1})\\ &\leq \mathbf{0}'_{(j)}(\boldsymbol{\Sigma}_{11}+\tau^2\mathbf{I}_{q,q})^{-1}\mathbf{0}_{(j)} \leq \frac{1}{\tau^2}\frac{\tau^2+\Lambda_q-\Lambda_1}{\tau^2+\Lambda_q} \leq \frac{1}{\tau^2}.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\operatorname{sign}(\boldsymbol{\beta}'_1)(\boldsymbol{\Sigma}_{11}+\tau^2\mathbf{I}_{q,q})^{-2}\operatorname{sign}(\boldsymbol{\beta}_1)\\ =\ &\frac{1}{\tau^4}\operatorname{sign}(\boldsymbol{\beta}'_1)\big((\mathbf{I}_{q,q}-\mathbf{D}_{11})-(\mathbf{D}_{11}-\mathbf{D}_{11}^2)\big)\operatorname{sign}(\boldsymbol{\beta}_1).\end{aligned}$$

The following two inequalities hold,

$$\begin{aligned}&\operatorname{sign}(\boldsymbol{\beta}'_1)(\boldsymbol{\Sigma_{11}}+\tau^2\mathbf{I}_{q,q})^{-2}\operatorname{sign}(\boldsymbol{\beta}_1)\\ &\quad\leq \frac{q}{\tau^4}-\frac{q\Lambda_1}{\tau^4(\tau^2+\Lambda_q)}-\frac{q\Lambda_1}{\tau^2(\tau^2+\Lambda_q)^2} \leq \frac{q}{\tau^4};\\ &\operatorname{sign}(\boldsymbol{\beta}'_1)(\boldsymbol{\Sigma_{11}}+\tau^2\mathbf{I}_{q,q})^{-2}\operatorname{sign}(\boldsymbol{\beta}_1)\\ &\quad\geq \frac{q}{\tau^4}-\frac{q\Lambda_q}{\tau^4(\tau^2+\Lambda_1)}-\frac{q\Lambda_q}{\tau^2(\tau^2+\Lambda_1)^2} \geq \frac{q}{\tau^4}\frac{\tau^2-2q+2\Lambda_1}{\tau^2}.\end{aligned}$$

Similarly,

$$\begin{aligned}&\frac{\tau^2-q+\Lambda_1}{\tau^4} \leq \frac{1}{\tau^2}\frac{\tau^2+\Lambda_1-\Lambda_q}{\tau^2+\Lambda_1} \leq \mathbf{0}'_{(\mathbf{j})}(\boldsymbol{\Sigma_{11}}+\tau^{\mathbf{2}}\mathbf{I_{qq}})^{-\mathbf{1}}\mathbf{0}_{(\mathbf{j})}\\ &\qquad\leq \frac{1}{\tau^2}(1-\frac{\Lambda_1}{\tau^2+\Lambda_q}) \leq \frac{1}{\tau^2};\\ &\frac{1}{\tau^4}\frac{\tau^2-2q+2\Lambda_1}{\tau^2} \leq \mathbf{0}'_{(\mathbf{j})}(\boldsymbol{\Sigma_{11}}+\tau^{\mathbf{2}}\mathbf{I_{qq}})^{-\mathbf{2}}\mathbf{0}_{(\mathbf{j})}\\ &\qquad\leq \frac{1}{\tau^4}(1-\frac{\Lambda_1}{\tau^2+\Lambda_q}-\frac{\tau^2\Lambda_1}{(\tau^2+\Lambda_q)^2}) \leq \frac{1}{\tau^4}.\end{aligned}$$

This completes the proof. □

Lemma 2. *Let $\mathbf{\Xi}$ be an $n \times p$ matrix comprising of np iid random entries distributed as $N(0,1)$ and split $\mathbf{\Xi}$ into $\mathbf{\Xi}_1$ and $\mathbf{\Xi}_2$ which are two submatrices spanned by the first q columns and the other $p-q$ columns of $\mathbf{\Xi}$, respectively. Let $\mathbf{Z}_1 = \mathbf{X}_1 + \tau\mathbf{\Xi}_1$ and $\mathbf{Z}_2 = \mathbf{X}_2 + \tau\mathbf{\Xi}_2$. Let $\mathbf{E^n_{11}} = (\frac{\mathbf{Z_1'Z_1}}{n})^{-1} - (\frac{E\mathbf{Z_1'Z_1}}{n})^{-1}$, then as $n \to \infty$, $\sqrt{n}\,\mathrm{sign}(\boldsymbol{\beta}_1')\mathbf{E^n_{11}}\,\mathrm{sign}(\boldsymbol{\beta}_1) \to_D N(0,\sigma_1^2)$ where $\frac{2q^2}{\tau^4}(1-\frac{2q}{\tau^2})^2 \le \sigma_1^2 \le \frac{2q^2}{\tau^4} + \frac{4q^3}{\tau^6}$, $\sqrt{n}\mathbf{0'_{(j)}}\mathbf{E^n_{11}}\mathbf{0_{(j)}} \to_D N(0,\sigma_2^2)$ where $\frac{2}{\tau^4}(1-\frac{2q}{\tau^2})^2 \le \sigma_2^2 \le \frac{2}{\tau^4} + \frac{4q}{\tau^6}$ and $\sqrt{n}\mathbf{0'_{(j)}}\mathbf{E^n_{11}}\,\mathrm{sign}(\boldsymbol{\beta}_1) \to_D N(0,\sigma_3^2)$ where $\frac{q}{\tau^4}(1-\frac{2q}{\tau^2})^2 \le \sigma_3^2 \le \frac{2q}{\tau^4} + \frac{4q^2}{\tau^6}$.*

Proof. Obviously,

$$E(\mathbf{Z_1'Z_1}) = \mathbf{X_1'X_1} + n\tau^2\mathbf{I}_{q,q} \tag{A.4}$$

Let $\mathbf{U^n} = \mathbf{Z_1'Z_1} - \mathbf{E}(\mathbf{Z_1'Z_1}) = \mathbf{X_1'\Xi_1} + \mathbf{\Xi_1'X_1} + \mathbf{\Xi_1'\Xi_1} - n\tau^2\mathbf{I}_{q,q}$ and by the formula of inverse of sum of matrices [6],

$$\begin{aligned} & (\mathbf{Z_1'Z_1})^{-1} - (E\mathbf{Z_1'Z_1})^{-1} \\ = \;& -(E\mathbf{Z_1'Z_1})^{-1}(\mathbf{I}_{q,q} + \mathbf{U^n}(E\mathbf{Z_1'Z_1})^{-1})^{-1}\mathbf{U^n}(E\mathbf{Z_1'Z_1})^{-1}. \end{aligned} \tag{A.5}$$

where $(\mathbf{I_{qq}} + \mathbf{U^n}(\mathbf{EZ_1'Z_1})^{-1})^{-1} \to_p \mathbf{I}_{q,q}$ and $n(E\mathbf{Z_1'Z_1})^{-1} \to (\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}$ as $n \to \infty$. Supposing $\frac{-\mathbf{U^n}}{\sqrt{n}} \to_D \mathbf{U}$ element-wise, then

$$\begin{aligned} & n\sqrt{n}\big((\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1} - (E\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\big) \\ & \to_D (\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}\mathbf{U}(\mathbf{\Sigma}_{11} + \tau^2\mathbf{I}_{q,q})^{-1}. \end{aligned}$$

It follows that

$$\begin{aligned} Var(\gamma_1'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_1) &= \frac{Var(\gamma_1'(\mathbf{X_1}+\mathbf{\Xi_1})'(\mathbf{X_1}+\mathbf{\Xi_1})\gamma_1)}{n} \\ &= \frac{2tr(\mathbf{V_1^2}) + 4\mu_1'\mathbf{V_1}\mu_1}{n} \end{aligned}$$

where $\mu_1 = E(\mathbf{X_1}+\mathbf{\Xi_1})\gamma_1 = \mathbf{X_1}\gamma_1$ and $\mathbf{V_1} = Var(\mathbf{\Xi_1}\gamma_1) = \tau^2\|\gamma_1\|_2^2\mathbf{I_{nn}}$. Noting that $\frac{tr(\mathbf{V_1^2})}{n} = \tau^4\|\gamma_1\|_2^4$ and $\frac{\mu_1'\mathbf{V_1}\mu_1}{n} = \tau^2\|\gamma_1\|_2^2\gamma_1'\mathbf{X_1'X_1}\gamma_1$ where $\gamma_1\|\gamma_1\|_2^2 \le \gamma_1'\mathbf{X_1'X_1}\gamma_1 \le \Lambda_q\|\gamma_1\|_2^2$, so

$$\frac{2q^2}{\tau^4}(1 - \frac{2q}{\tau^2} + \frac{2\Lambda_1}{\tau^2})^2 \le Var(\gamma_1'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_1) \le \frac{2q^2}{\tau^4} + \frac{4q^3}{\tau^6}. \tag{A.6}$$

Similarly, $Var(\gamma_2'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_2) = \frac{2tr(\mathbf{V_2^2})+4\mu_2'\mathbf{V_2}\mu_2}{n}$, $\mathbf{V_2} = Var(\mathbf{\Xi_1}\gamma_2) = \tau^2\|\gamma_2\|_2^2\mathbf{I_{nn}}$ where $\mu_2 = \mathbf{X_1}\gamma_2$. Noting that

$$\begin{aligned} & \frac{tr(\mathbf{V_2^2})}{n} = \tau^4\|\gamma_2\|_2^4 \\ & \mu_2'\mathbf{V_2}\mu_2 = \tau^2\|\gamma_2\|_2^2\gamma_2'\mathbf{X_1'X_1}\gamma_2 \le q\tau^2\|\gamma_2\|_2^4, \end{aligned}$$

so

$$\frac{2}{\tau^4}(1 - \frac{2q}{\tau^2} + \frac{2\Lambda_1}{\tau^2})^2 \le Var(\gamma_2'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_2) \le \frac{2}{\tau^4} + \frac{4q}{\tau^6}. \tag{A.7}$$

Concerning $Var(\gamma_{\mathbf{2}}'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_{\mathbf{1}})$, we have $Var(\gamma_{\mathbf{2}}'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_{\mathbf{1}}) = tr(\mathbf{V_{12}^2}) + tr(\mathbf{V_2V_1}) + \mu_{\mathbf{1}}'\mathbf{V_2}\mu_{\mathbf{1}} + \mu_{\mathbf{2}}'\mathbf{V_1}\mu_{\mathbf{2}} + 2\mu_{\mathbf{1}}'\mathbf{V_{12}}\mu_{\mathbf{2}}$ where $\mathbf{V_{12}} = E(\mathbf{\Xi_1}\gamma_{\mathbf{2}}\gamma_{\mathbf{1}}'\mathbf{\Xi_1'}) = \tau^2\gamma_{\mathbf{1}}'\gamma_2\mathbf{I_{nn}}$.

Noting that $\frac{tr(\mathbf{V_{12}^2})}{n} = \tau^4(\gamma_{\mathbf{1}}'\gamma_{\mathbf{2}})^2 \le \tau^4\|\gamma_{\mathbf{1}}\|_2^2\|\gamma_{\mathbf{2}}\|_2^2 \le \frac{q}{\tau^4}$, $\frac{q}{\tau^4}(1-\frac{2q}{\tau^2})^2 \le \frac{tr(\mathbf{V_2V_1})}{n} = \tau^4\|\gamma_{\mathbf{1}}\|_2^2\|\gamma_{\mathbf{2}}\|_2^2 \le \frac{q}{\tau^4}$, $\frac{\mu_{\mathbf{1}}'\mathbf{V_2}\mu_{\mathbf{1}}}{n} \le \tau^2\|\gamma_{\mathbf{2}}\|_2^2\|\mu_{\mathbf{1}}\|_2^2 \le q\tau^2\|\gamma_{\mathbf{1}}\|_2^2\|\gamma_{\mathbf{2}}\|_2^2 \le \frac{q^2}{\tau^6}$, $\frac{\mu_{\mathbf{2}}'\mathbf{V_1}\mu_{\mathbf{2}}}{n} \le \frac{q^2}{\tau^6}$ and $\frac{\mu_{\mathbf{1}}\mathbf{V_{12}}\mu_{\mathbf{2}}}{n} = \tau^2\gamma_{\mathbf{1}}'\gamma_{\mathbf{2}}\mu_{\mathbf{1}}'\mu_{\mathbf{2}} \le q\tau^2\|\gamma_{\mathbf{1}}\|_2^2\|\gamma_{\mathbf{2}}\|_2^2 \le \frac{q^2}{\tau^6}$, so

$$\frac{q}{\tau^4}(1-\frac{2q}{\tau^2}+\frac{2\Lambda_1}{\tau^2})^2 \le Var(\gamma_{\mathbf{2}}'\frac{\mathbf{U^n}}{\sqrt{n}}\gamma_{\mathbf{1}}) \le \frac{2q}{\tau^4}+\frac{4q^2}{\tau^6}. \tag{A.8}$$

This completes the proof. □

Proof of Proposition 3.1

By KKT condition [3] we have the following conclusions:

$$(\mathbf{Z}_h'\mathbf{W}_h - \mathbf{Z}_h'\mathbf{Z}_h\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda})_{[j]} = \frac{\lambda}{2}\operatorname{sign}(\widehat{\theta}_{h,j}^{\mathrm{LA},\lambda}) \quad \text{for } j \text{ s.t. } \widehat{\theta}_{h,j}^{\mathrm{LA},\lambda} \neq 0 \tag{A.9}$$

$$|(\mathbf{Z}_h'\mathbf{W}_h - \mathbf{Z}_h'\mathbf{Z}_h\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda})_{[j]}| < \frac{\lambda}{2} \quad \text{for } j \text{ s.t. } \widehat{\theta}_{h,j}^{\mathrm{LA},\lambda} = 0 \tag{A.10}$$

where $[j]$ denotes the jth entry of a vector and $\widehat{\theta}_{h,j}^{\mathrm{LA},\lambda}$ $(1 \le j \le p)$ is the j-th entry of $\widehat{\boldsymbol{\theta}}_h^{\mathrm{LA},\lambda}$.

Let $\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} = (\widehat{\theta}_{h,1}^{\mathrm{LA},\lambda},\cdots,\widehat{\theta}_{h,q}^{\mathrm{LA},\lambda})'$ and $\widehat{\boldsymbol{\theta}}_{h,2}^{\mathrm{LA},\lambda} = (\widehat{\theta}_{h,q+1}^{\mathrm{LA},\lambda},\cdots,\widehat{\theta}_{h,p}^{\mathrm{LA},\lambda})'$. The following condition

$$\operatorname{sign}(\boldsymbol{\beta}_1^0) * \widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} >_c 0 \tag{A.11}$$

is equivalent to $\operatorname{sign}(\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}) = \operatorname{sign}(\boldsymbol{\beta}_1^0)$, which indicates

$$\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} = (\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}_{h,1}'\mathbf{W}_h - \frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)\big). \tag{A.12}$$

Furthermore, as $\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}$ satisfies (A.9), (A.11) is equivalent to

$$\operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}_{h,1}'\mathbf{W}_h - \frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)\big) >_c 0, \tag{A.13}$$

where

$$\begin{aligned}
& \operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}(\mathbf{Z}_{h,1}'\mathbf{W}_h - \frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)) \\
= \; & \operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}_{h,1}'(\mathbf{X_1}\beta_{\mathbf{1}}^{\mathbf{0}} + \boldsymbol{\varepsilon} + \tau\mathbf{\Xi}_h\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0}) \\
& -\frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)\big) \\
= \; & \operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1}\boldsymbol{\beta}_1^0 - \tau\mathbf{Z}_{h,1}'\mathbf{\Xi}_h\boldsymbol{\beta}^0 + \boldsymbol{Z}_{h,1}'\boldsymbol{\varepsilon}
\end{aligned}$$

$$
\begin{aligned}
&+\tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} - \frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)\big)\\
= \;& |\boldsymbol{\beta}_1^0| + \operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}'_{h,1}\boldsymbol{\varepsilon} - \frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0)\\
&+\tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}^0)\big).
\end{aligned}
$$

Therefore, $\operatorname{sign}(\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}) = \operatorname{sign}(\boldsymbol{\beta}_1^0)$ is equivalent to

$$
\begin{aligned}
&|\boldsymbol{\beta}_1^0| + \operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\big(\mathbf{Z}'_{h,1}\boldsymbol{\varepsilon}\\
&-\frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0) + \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}^0)\big) >_c 0. \qquad \text{(A.14)}
\end{aligned}
$$

Hence, for a scalar η the following two inequalities

$$
\begin{aligned}
&|\boldsymbol{\beta}_1^0| + \operatorname{sign}(\boldsymbol{\beta}_1^0) * \tau(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} >_c \eta\\
&\operatorname{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\big(\frac{\lambda}{2}\operatorname{sign}(\boldsymbol{\beta}_1^0) - \mathbf{X}'_1\boldsymbol{\varepsilon} - \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}^0)\big) <_c \eta
\end{aligned}
$$

indicate $\operatorname{sign}(\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}) = \operatorname{sign}(\boldsymbol{\beta}_1^0)$.

Similarly, as (A.9) holds

$$
\begin{aligned}
&\mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda} = \mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1}\mathbf{W}_h\\
&+\frac{\lambda}{2}\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&|\mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}\widehat{\boldsymbol{\theta}}_{h,1}^{\mathrm{LA},\lambda}| \le |\mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1}\mathbf{W}_h|\\
&+\frac{\lambda}{2}|\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0)|.
\end{aligned}
$$

Supposing $1 > \delta > 0$, $\widehat{\boldsymbol{\theta}}_{h,2}^{\mathrm{LA},\lambda} =_c 0$ is implied by the following two inequalities:

$$
\begin{aligned}
|\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0)| &\le_c (1-\delta)\\
|\mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1}\mathbf{W}_h| &\le_c 2^{-1}\delta\lambda
\end{aligned}
$$

. The proof is complete.

Proof of Theorem 1

First of all we derive the probability lower bounds of the Irrepresentable Condition, Exclusion, Beta-min and Inclusion conditions. Suppose $\tau \ge \sqrt{8}q^{3/4}$ and let

$$
\delta = \frac{1}{2}; \qquad \lambda = 2\sigma\tau^2\sqrt{\frac{2nq\log p}{\kappa_0^2}}; \qquad \eta = \frac{2\lambda}{n\tau^2} = 4\sigma\sqrt{\frac{2q\log p}{n\kappa_0^2}}.
$$

Step 1: Perturbed Irrepresentable Condition

Let A_h denote the Perturbed Irrepresentable Condition holding at the h-th perturbation and $Pr(A_h) \geq Pr(A_{1h} \cap A_{2h})$ where

$$Pr(A_{1h}) = Pr\big(|\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)| \leq_c \frac{1-\delta}{2}\big)$$
$$Pr(A_{2h}) = Pr\big(|\mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}((\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1} - (\mathbf{R}^n_{11})^{-1})\operatorname{sign}(\boldsymbol{\beta}^0_1)| \leq_c \frac{1-\delta}{2}\big).$$

Next, $Pr(A_{1h}) \geq Pr\big(|\mathbf{X}'_2\mathbf{X}_1(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)|+\tau|\boldsymbol{\Xi}'_{h,2}\mathbf{X}_1(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)|+\tau|\mathbf{X}'_2\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)| + \tau^2|\boldsymbol{\Xi}'_{h,2}\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)| \leq_c (1-\delta)/2\big)$. Let $\mathbf{X}'_{[,j]}$ be j-th column of $\mathbf{X}$ $(j = q+1, \cdots, p)$. Then,

$$|\mathbf{X}'_{[,j]}\mathbf{X}_1(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)| \leq \frac{q}{\tau^2}.$$

Let $\boldsymbol{\xi}_{h,j}$ be the j-th column of $\boldsymbol{\Xi}_h$, then $E\big(\tau^2\boldsymbol{\xi}'_{h,j}\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)\big) = 0$ and $Var\big(\tau^2\boldsymbol{\xi}'_{h,j}\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}$
$\operatorname{sign}(\boldsymbol{\beta}^0_1)\big) = n\tau^4\operatorname{sign}((\boldsymbol{\beta}^0_1)')(\mathbf{R}^n_{11})^{-1}\mathbf{I}_{q,q}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1) \leq \frac{q}{n}$.

Similarly,

$$E(\tau\boldsymbol{\xi}'_{h,j}\mathbf{X}_1(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)) = 0;$$
$$E(\tau\mathbf{X}'_{[,j]}\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1)) = 0$$
$$Var(\tau\boldsymbol{\xi}'_{h,j}\mathbf{X}_1(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1))$$
$$= n\tau^2\operatorname{sign}((\boldsymbol{\beta}^0_1)')(\mathbf{R}^n_{11})^{-1}\boldsymbol{\Sigma}^n_{11}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1) \leq \frac{q^2}{n\tau^2};$$
$$Var(\tau\mathbf{X}'_{[,j]}\boldsymbol{\Xi}_{h,1}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1))$$
$$= n\tau^2\operatorname{sign}((\boldsymbol{\beta}^0_1)')(\mathbf{R}^n_{11})^{-1}\mathbf{I}_{q,q}(\mathbf{R}^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^0_1) \leq \frac{q}{n\tau^2}.$$

Therefore, $Pr(A^c_{2h}) = o(Pr(A^c_{1h}))$ and

$$Pr(A^c_{1h}) \leq (p-q)\Phi^c\Big(\frac{1/2 - \delta/2 - q/\tau^2}{\sqrt{q}/\sqrt{n}}\Big) = (p-q)\Phi^c\big(\frac{\sqrt{n}}{8\sqrt{q}}\big(\frac{1}{4} - \frac{q}{\tau^2}\big)\big). \quad \text{(A.15)}$$

Step 2: Perturbed Exclusion Condition

Let B_h denote the Exclusion Condition holding at the h-perturbation. First,

$$\begin{aligned} & \mathbf{Z}'_{h,2}\mathbf{W}_h - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1}\mathbf{W}_h \\ = \; & (\mathbf{Z}'_{h,2} - \mathbf{Z}'_{h,2}\mathbf{Z}_{h,1}(\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\mathbf{Z}'_{h,1})(\mathbf{Z}_{h,1}\boldsymbol{\beta}^0_1 + \boldsymbol{\varepsilon} + \tau\boldsymbol{\Xi}_{h,1}(\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0}_1 - \boldsymbol{\beta}^0_1) \\ & + \tau\boldsymbol{\Xi}_{h,2}\widehat{\boldsymbol{\beta}}^{\mathrm{LA},\lambda_0}_2) \end{aligned}$$

$$
\begin{aligned}
&= (\mathbf{X}_2' + \tau\boldsymbol{\Xi}_{h,2}' - \mathbf{Z}_{h,2}'\mathbf{Z}_{h,1}(\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\mathbf{Z}_{h,1}')(\boldsymbol{\varepsilon} + \tau\boldsymbol{\Xi}_{h,1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0) \\
&\quad +\tau\boldsymbol{\Xi}_{h,2}\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0}) \\
&= \tau^2\boldsymbol{\Xi}_{h,2}'\boldsymbol{\Xi}_{h,2}\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0} + \big(\mathbf{X}_2' - \mathbf{Z}_{h,2}'\mathbf{Z}_{h,1}(\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\mathbf{Z}_{h,1}'\big)(\boldsymbol{\varepsilon} \\
&\quad +\tau\boldsymbol{\Xi}_{h,1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0) + \tau\boldsymbol{\Xi}_{h,2}\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0}) + \tau\boldsymbol{\Xi}_{h,2}'\big(\boldsymbol{\varepsilon} + \tau\boldsymbol{\Xi}_{h,1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0)\big).
\end{aligned}
$$

Then $Pr(B_h) \geq Pr(B_{1h} \cap B_{2h})$ where

$$
\begin{aligned}
Pr(B_{1h}) &= Pr\big(\tau^2\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0} - \tau^2\mathbf{X}_2'\mathbf{X}_1(\mathbf{R}_{11}^n)^{-1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0) \leq_c \lambda/8\big) \\
Pr(B_{2h}) &= Pr\big(\mathbf{Z}_{h,2}'\mathbf{W}_h - \mathbf{Z}_{h,2}'\mathbf{Z}_{h,1}(\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\mathbf{Z}_{h,1}'\mathbf{W}_h - \tau^2\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0} \\
&\quad +\tau^2\mathbf{X}_2'\mathbf{X}_1(\mathbf{R}_{11}^n)^{-1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0) \leq_c \lambda/8\big)
\end{aligned}
$$

In $Pr(B_{1h})$

$$
\begin{aligned}
&Pr(\tau^2\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0} \leq_c \lambda/16) \geq Pr(\tau^2\|\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0}\|_2 \leq \lambda/16); \\
&Pr(-\tau^2\mathbf{X}_2'\mathbf{X}_1(\mathbf{R}_{11}^n)^{-1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0) \leq_c \lambda/16) \\
&\qquad \geq Pr(\tau^2\|\mathbf{X}_2'\mathbf{X}_1(\mathbf{R}_{11}^n)^{-1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0)\|_2 \leq \lambda/16)
\end{aligned}
$$

According to Theorem 2 of [2],

$$
\begin{aligned}
&Pr\big(\|\widehat{\boldsymbol{\beta}}_2^{\mathrm{LA},\lambda_0}\|_2 \geq \sqrt{\frac{q\sigma^2\log p}{n\kappa_0^2}}\big) \leq \Phi^c(\sqrt{2\log p}); \\
&Pr\big(\tau^2\|\mathbf{X}_2'\mathbf{X}_1(\mathbf{R}_{11}^n)^{-1}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{LA},\lambda_0} - \boldsymbol{\beta}_1^0)\|_2 \geq \sqrt{\frac{q\sigma^2\log p}{n\kappa_0^2}}\big) \leq \Phi^c(\sqrt{2\log p}).
\end{aligned}
$$

Finally, $Pr(B_{2h}^c) = o(Pr(B_{1h}^c))$.

Step 3: Perturbed Beta-min Condition

Let C_h denote the Perturbed Beta-min Condition holding at the h-th perturbation. Consider $|\boldsymbol{\beta}_1^0| + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \tau(\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1}\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} = |\boldsymbol{\beta}_1^0| + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \tau(\mathbf{R}_{11}^n)^{-1}\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \tau\big((\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1} - (\mathbf{R}_{11}^n)^{-1}\big)\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} = |\boldsymbol{\beta}_1^0| + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \frac{1}{n\tau}(\mathbf{I}_{q,q} - \mathbf{D}_{11}^n)\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \tau\big((\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1} - (\mathbf{R}_{11}^n)^{-1}\big)\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon}$. Thus,

$$
Pr(C_h) \geq Pr(C_{1h} \cap C_{2h}) \tag{A.16}
$$

where

$$
\begin{aligned}
Pr(C_{1h}) &= Pr\big(|\boldsymbol{\beta}_1^0| + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \frac{1}{n\tau}\mathbf{I}_{q,q}\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} >_c \eta\big) \\
Pr(C_{2h}) &= Pr\big(\mathrm{sign}(\boldsymbol{\beta}_1^0) * \frac{1}{n\tau}(-\mathbf{D}_{11}^n)\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} \\
&\quad + \mathrm{sign}(\boldsymbol{\beta}_1^0) * \tau\big((\mathbf{Z}_{h,1}'\mathbf{Z}_{h,1})^{-1} - (\mathbf{R}_{11}^n)^{-1}\big)\boldsymbol{\Xi}_{h,1}'\boldsymbol{\varepsilon} >_c -\frac{\eta}{2}\big).
\end{aligned}
$$

Let $\boldsymbol{\xi}'_{h,j}$ be the j-th column of $\boldsymbol{\Xi}_h$, then $Pr(C_{1h}) = \prod_{j=1}^{q} p_j^0$ where $p_j^0 = E\big(P(\text{sign}(\beta_j^0)*\boldsymbol{\xi}'_{h,j}\boldsymbol{\varepsilon}+n\tau|\beta_j^0| >_c n\tau\eta|\boldsymbol{\epsilon})\big) = E\Phi^c\big(\frac{\tau\sqrt{n}}{\|n^{-1}\boldsymbol{\epsilon}\|_2}(\eta-|\beta_j^0|)\big)$. As $\eta > |\beta_j^0|$, $p_j^0 = E\Phi^c\big(\frac{\tau\sqrt{n}}{\|n^{-1}\boldsymbol{\epsilon}\|_2}(\eta - |\beta_j^0|)\big) \geq \Phi^c\big(E[\frac{\tau\sqrt{n}}{\|n^{-1}\boldsymbol{\epsilon}\|_2}](\eta - |\beta_j^0|)\big) = \Phi^c\big(\frac{\tau\sqrt{n}}{\sigma}(\eta - |\beta_j^0|)\big)$ by Jensen's Inequality. Furthermore, $Pr(C_{2h}) \geq Pr(C_{21h} \cap C_{22h})$ where

$$Pr(C_{21h}) \geq Pr\big(\text{sign}(\boldsymbol{\beta}_1^0) * \frac{1}{n\tau}(-\mathbf{D}_{11}^n)\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} \leq_c \frac{\eta}{4}\big)$$
$$Pr(C_{22h}) \geq Pr\big(\text{sign}(\boldsymbol{\beta}_1^0) * \tau\big((\mathbf{R}_{11}^n)^{-1} - (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}\big)\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} \leq_c \frac{\eta}{4}\big).$$

As for $Pr(C_{21h}^c)$, we establish

$$Pr(C_{21h}^c) \leq \sum_{j=1}^{q} Pr(\mathbf{0}'_{(\mathbf{j})}\mathbf{D}_{11}^n\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} \geq \frac{n\tau\eta}{2})$$
$$\leq \sum_{j=1}^{q} E\big(Pr(\mathbf{0}'_{(\mathbf{j})}\mathbf{D}_{11}^n\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} \geq \frac{n\tau\eta}{2})|\boldsymbol{\epsilon}\big).$$

Obviously, $\mathbf{0}'_{(\mathbf{j})}\mathbf{D}_{11}^n\boldsymbol{\Xi}'_{h,1}\boldsymbol{\varepsilon} \sim N(0, \|\boldsymbol{\epsilon}\|_2^2\mathbf{0}'_{(\mathbf{j})}(\mathbf{D}_{11}^n)^2\mathbf{0}_{(j)})$ conditioning on $\boldsymbol{\epsilon}$ where $\mathbf{0}'_{(\mathbf{j})}(\mathbf{D}_{11}^n)^2\mathbf{0}_{(j)} \leq q^2/(\tau^2+q)^2 \leq 1$. In summary,

$$Pr(C_{21h}^c) = E\big(Pr(C_{21h}^c|\boldsymbol{\epsilon})\big) \leq qE\big(\Phi^c(\frac{n\sqrt{q}\eta}{\|\boldsymbol{\epsilon}\|_2})\big) \leq q\Phi^c(q\sqrt{2\log p})$$

and $Pr(C_{22h}^c) = o\big(Pr(C_{21h}^c)\big)$.

Step 4: Perturbed Inclusion Condition

Let D_h denote the Perturbed Inclusion Condition holding at the h-th perturbation, then

$$Pr(D_h) = Pr(\text{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1}(\frac{\lambda}{2}\text{sign}(\boldsymbol{\beta}_1^0)$$
$$-\mathbf{X}'_1\boldsymbol{\varepsilon} - \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\text{LA},\lambda_0} - \boldsymbol{\beta}^0)) <_c \eta).$$

Moreover, $Pr(D_h) \geq Pr(D_{1h} \cap D_{2h} \cap D_{3h})$ where

$$Pr(D_{1h}) = Pr(\text{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{R}_{11}^n)^{-1}\big(\frac{\lambda}{2}\text{sign}(\boldsymbol{\beta}_1^0) - n\tau^2(\widehat{\boldsymbol{\beta}}_1^{\text{LA},\lambda_0} - \boldsymbol{\beta}_1^0)\big) <_c \frac{\eta}{2})$$
$$Pr(D_{2h}) = Pr(\text{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{R}_{11}^n)^{-1}\big(- \mathbf{X}'_1\boldsymbol{\varepsilon} - \tau(\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_{h,1} - n\tau\mathbf{I}_{q,q})(\widehat{\boldsymbol{\beta}}_1^{\text{LA},\lambda_0}$$
$$-\boldsymbol{\beta}_1^0) - \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_{h,2}\widehat{\boldsymbol{\beta}}_2^{\text{LA},\lambda_0}\big) <_c \frac{\eta}{4})$$
$$Pr(D_{3h}) = Pr(\text{sign}(\boldsymbol{\beta}_1^0) * \big((\mathbf{Z}'_{h,1}\mathbf{Z}_{h,1})^{-1} - (\mathbf{R}_{11}^n)^{-1}\big)(\frac{\lambda}{2}\text{sign}(\boldsymbol{\beta}_1^0)$$
$$-\mathbf{X}'_1\boldsymbol{\varepsilon} - \tau\mathbf{Z}'_{h,1}\boldsymbol{\Xi}_h(\widehat{\boldsymbol{\beta}}^{\text{LA},\lambda_0} - \boldsymbol{\beta}^0)) <_c \frac{\eta}{4}).$$

Considering the j-th entry of $\text{sign}(\boldsymbol{\beta}_1^0) * (\mathbf{R}_{11}^n)^{-1}\frac{\lambda}{2}\text{sign}(\boldsymbol{\beta}_1^0)$, we establish for $\tau \geq q^{3/4}$

$$|\mathbf{0}'_{(\mathbf{j})}\mathbf{D}_{11}^n \frac{\lambda}{2}\text{sign}(\boldsymbol{\beta}_1^0)| \leq \frac{\lambda}{2}\sum_{j=1}^{q}\frac{\sqrt{q}\Lambda_j^n}{\tau^2+\Lambda_j^n} \leq \frac{\lambda}{2}\frac{q^{3/2}}{\tau^2} \leq \frac{\lambda}{2}. \tag{A.17}$$

Furthermore,

$$\begin{aligned} Pr(D_{1h}^c) &\leq Pr\big(\|\widehat{\boldsymbol{\beta}}_1^{\text{LA},\lambda_0} - \boldsymbol{\beta}_1^0\|_2 + \frac{\lambda}{2\tau^2 n} > \frac{\eta}{2}\big) \\ &\leq Pr\big(\|\widehat{\boldsymbol{\beta}}_1^{\text{LA},\lambda_0} - \boldsymbol{\beta}_1^0\|_2 > \frac{\eta}{4}\big) \leq \Phi^c(\sqrt{2\log p}). \end{aligned}$$

Moreover, $Pr(D_{2h}^c) = o(Pr(D_{1h}^c))$ and $Pr(D_{3h}^c) = o(Pr(D_{1h}^c))$.

We have established the probability lower bounds of $Pr(A_h)$, $Pr(B_h)$, $Pr(C_h)$ and $Pr(D_h)$ and need to derive the upper bound of $Pr(\cap_{h=1}^H(A_h^c \cup B_h^c \cup C_h^c \cup D_h^c))$. The following inequality holds,

$$\begin{aligned} &Pr(\cap_{h=1}^H(A_h^c \cup B_h^c \cup C_h^c \cup D_h^c)) \leq \\ &\quad Pr(A_h^c) + Pr(B_h^c) + Pr(\cap_{h=1}^H C_{1h}^c) + Pr(C_{2h}^c) + Pr(D_h^c) \end{aligned}$$

where

$$\begin{aligned} &Pr(A_h^c) + Pr(B_h^c) + Pr(C_{2h}^c) + Pr(D_h^c) \leq C_0\Phi^c(\sqrt{2\log p}) \\ &Pr(\cap_{h=1}^H C_{1h}^c) \leq \big(1 - \prod_{j=1}^{q} p_j\big)^H. \end{aligned}$$

This completes the proof. □

Acknowledgements

I thank two anonymous referees, the Associate Editor and the Editor, Dr. Domenico Marinucci, for providing me with very insightful comments and valuable suggestions to improve the paper.

References

[1] Bach, F. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *Procedings 25th International Conference Machine Learning* 33–40. Association for Computing Machinery, New York.

[2] Bellec, P. and Tsybakov, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern Problems of Stochastic Analysis and Statistics: Selected Contributions In Honor of Valentin Konakov*, to appear. Springer International Publishing AG, Switzerland. MR2195633

[3] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575
[4] Bühlmann, P. (2011). Comments on 'Regression shrinkage and selection via the lasso: A retrospective'. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 277–279.
[5] Bühlmann, P., Kalisch, M. and Meier, L. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications* **1** 255–278. MR2572443
[6] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York. MR1467237
[7] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523
[8] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242
[9] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701
[10] Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293. MR3160554
[11] Zhang, Y. and Shen, X. (2010). Model selection procedure for high-dimensional data. *Stat. Anal. Data Min.* **3** 350–358. MR2726244
[12] Zhang, Y. and Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics* **187** 95–112. MR3347297
[13] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449
[14] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327