In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference

Sen Zhao[*], Ali Shojaie[†], and Daniela Witten[‡]

University of Washington

Abstract

In recent years, a great deal of interest has focused on conducting inference on the parameters in a linear model in the high-dimensional setting. In this paper, we consider a simple and very naïve two-step procedure for this task, in which we (i) fit a lasso model in order to obtain a subset of the variables; and (ii) fit a least squares model on the lasso-selected set. Conventional statistical wisdom tells us that we cannot make use of the standard statistical inference tools for the resulting least squares model (such as confidence intervals and p-values), since we peeked at the data twice: once in running the lasso, and again in fitting the least squares model. However, in this paper, we show that under a certain set of assumptions, with high probability, the set of variables selected by the lasso is deterministic. Consequently, the naïve two-step approach can yield confidence intervals that have asymptotically correct coverage, as well as p-values with proper Type-I error control. Furthermore, this two-step approach unifies two existing camps of work on high-dimensional inference: one camp has focused on inference based on a sub-model selected by the lasso, and the other has focused on inference using a debiased version of the lasso estimator.

Keywords— Confidence interval; Lasso; p-value; Post-selection inference; Significance testing.

1 Introduction

In this paper, we consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{1}$$

[*]sen-zhao@sen-zhao.com.
[†]ashojaie@u.washington.edu.
[‡]dwitten@u.washington.edu.

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ is an $n \times p$ deterministic design matrix, $\boldsymbol{\epsilon}$ is a vector of independent and identically distributed errors with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma_{\boldsymbol{\epsilon}}^2$, and $\boldsymbol{\beta}^*$ is a p-vector of coefficients. Without loss of generality, we assume that the columns of $\boldsymbol{X}$ are centered and standardized, such that $\sum_{i=1}^n X_{(i,k)} = 0$ and $\|\boldsymbol{x}_k\|_2^2 = n$ for $k = 1, \ldots, p$.

When the number of variables p is much smaller than the sample size n, estimation and inference for the vector $\boldsymbol{\beta}^*$ are straightforward. For instance, estimation can be performed using ordinary least squares, and inference can be conducted using classical approaches (see, e.g., Gelman and Hill, 2006; Weisberg, 2013).

As the scope and scale of data collection have increased across virtually all fields, there is an increase in data sets that are *high dimensional*, in the sense that the number of variables, p, is larger than the number of observations, n. In this setting, classical approaches for estimation and inference of $\boldsymbol{\beta}^*$ cannot be directly applied. In the past 20 years, a vast statistical literature has focused on estimating $\boldsymbol{\beta}^*$ in high dimensions. In particular, penalized regression methods, such as the lasso (Tibshirani, 1996),

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\}, \tag{2}$$

can be used to estimate $\boldsymbol{\beta}^*$. However, the topic of inference in the high-dimensional setting remains relatively unexplored, despite promising recent work in this area. Roughly speaking, recent work on inference in the high-dimensional setting falls into two classes: (i) methods that examine the null hypothesis $H_{0,j}^* : \beta_j^* = 0$; and (ii) methods that make inference based on a sub-model. We will review these two classes of methods in turn.

First, we review methods that examine the null hypothesis $H_{0,j}^* : \beta_j^* = 0$, i.e. that the variable $\boldsymbol{x}_j$ is unassociated with the outcome $\boldsymbol{y}$, conditional on *all other variables*. It might be tempting to estimate $\boldsymbol{\beta}^*$ using the lasso (2), and then (for instance) to construct a confidence interval around $\hat{\beta}_{\lambda,j}$. Unfortunately, such an approach is problematic, because $\hat{\boldsymbol{\beta}}_\lambda$ is a biased estimate of $\boldsymbol{\beta}^*$. To remedy this problem, we can apply a one-step adjustment to $\hat{\boldsymbol{\beta}}_\lambda$, such that under appropriate assumptions, the resulting *debiased estimator* is asymptotically unbiased for $\boldsymbol{\beta}^*$. Then, p-values and confidence intervals can be constructed around this debiased estimator. For example, such an approach is taken by the low dimensional projection estimator (LDPE; Zhang and Zhang, 2014; van de Geer et al., 2014), the debiased lasso test with unknown population covariance (SSLasso; Javanmard and Montanari, 2013, 2014a), the debiased lasso test with known population covariance (SDL; Javanmard and Montanari, 2014b), and the decorrelated score test (dScore; Ning and Liu, 2016). See Dezeure et al. (2015) for a review of such procedures. In what follows, we will refer to these and related approaches for testing $H^*_{0,j} : \beta^*_j = 0$ as *debiased lasso tests*.

Now, we review recent work that makes statistical inference based on a sub-model. Recall that the challenge in high dimensions stems from the fact that when $p > n$, classical statistical methods cannot be applied; for instance, we cannot even perform ordinary least squares (OLS). This suggests a simple approach: given an index set $\mathcal{M} \subseteq \{1, \ldots, p\}$, let $\boldsymbol{X}_{\mathcal{M}}$ denote the columns of $\boldsymbol{X}$ indexed by $\mathcal{M}$. Then, we can consider performing inference *based on the sub-model* composed only of the features in the index set $\mathcal{M}$. That is, rather than considering the model (1), we consider the sub-model

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{(\mathcal{M})} + \boldsymbol{\epsilon}^{(\mathcal{M})}. \tag{3}$$

In (3), the notation $\boldsymbol{\beta}^{(\mathcal{M})}$ and $\boldsymbol{\epsilon}^{(\mathcal{M})}$ emphasizes that the true regression coefficients and corresponding noise are functions of the set $\mathcal{M}$.

Now, provided that $|\mathcal{M}| < n$, we can perform estimation and inference on the vector $\boldsymbol{\beta}^{(\mathcal{M})}$ using classical statistical approaches. For instance, we can consider building confidence intervals $\mathrm{CI}_j^{(\mathcal{M})}$ such that for any $j \in \mathcal{M}$,

$$\Pr\left[\beta_j^{(\mathcal{M})} \in \mathrm{CI}_j^{(\mathcal{M})}\right] \geq 1 - \alpha. \tag{4}$$

At first blush, the problems associated with high dimensionality have been solved!

Of course, there are some problems with the aforementioned approach. The first problem is that the coefficients in the sub-model (3) typically are not the same as the coefficients in the original model (1) (Berk et al., 2013). Roughly speaking, the problem is that the coefficients in the model (1) quantify the linear association between a given variable and the response, *conditional on the other* $p-1$ *variables*, whereas the coefficients in the model (3) quantify the linear association between a variable and the response, *conditional on the other* $|\mathcal{M}|-1$ *variables in the sub-model.* The true regression coefficients in the sub-model are of the form

$$\boldsymbol{\beta}^{(\mathcal{M})} \equiv \left(\boldsymbol{X}_{\mathcal{M}}^{\top}\boldsymbol{X}_{\mathcal{M}}\right)^{-1}\boldsymbol{X}_{\mathcal{M}}^{\top}\boldsymbol{X}\boldsymbol{\beta}^{*}. \tag{5}$$

Thus, $\boldsymbol{\beta}^{(\mathcal{M})} \neq \boldsymbol{\beta}^{*}_{\mathcal{M}}$ unless $\boldsymbol{X}_{\mathcal{M}}^{\top}\boldsymbol{X}_{\mathcal{M}^c}\boldsymbol{\beta}^{*}_{\mathcal{M}^c} = \mathbf{0}$. To see this more concretely, consider the following example with $p = 4$ deterministic variables. We let

$$\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0.6 & 0 \\ 0 & 1 & 0.6 & 0 \\ 0.6 & 0.6 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Set $\boldsymbol{\beta}^{*} = (1,1,0,0)^{\top}$. If $\mathcal{M} = \{2,3\}$, then it is easy to verify that

$$\begin{aligned} \boldsymbol{\beta}^{(\mathcal{M})} &= \left[\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right]^{-1}_{(\{2,3\},\{2,3\})} \left[\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right]_{(\{2,3\},\{1,2,3,4\})} \boldsymbol{\beta}^{*} \\ &= \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0.6 & 0 \\ 0.6 & 0.6 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^{\top} \\ &= \begin{bmatrix} 0.4375 \\ 0.9375 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \boldsymbol{\beta}^{*}_{\mathcal{M}}. \end{aligned}$$

The second problem that arises in restricting our attention to the sub-model (3) is that in practice, the index set $\mathcal{M}$ is not pre-specified. Instead, it is typically chosen based on the data. For instance, one might take $\mathcal{M}$ to equal the support of the lasso estimate,

$$\hat{\mathcal{A}}_{\lambda} \equiv \operatorname{supp}(\hat{\boldsymbol{\beta}}_{\lambda}) \equiv \left\{ j : \hat{\beta}_{\lambda,j} \neq 0 \right\}. \tag{6}$$

The problem is that if we construct the index set $\mathcal{M}$ based on the data, and then apply classical inference approaches on the vector $\boldsymbol{\beta}^{(\mathcal{M})}$, the resulting p-values and confidence intervals will not be valid (see, e.g., Pötscher, 1991; Kabaila, 1998; Leeb and Pötscher, 2003, 2005, 2006a,b, 2008; Kabaila, 2009; Berk et al., 2013). This is because we peeked at the data twice: once to determine which variables to include in $\mathcal{M}$, and then again to test hypotheses associated with those variables. Consequently, an extensive recent body of literature has focused on the task of performing inference on $\boldsymbol{\beta}^{(\mathcal{M})}$ in (3) given that $\mathcal{M}$ was chosen based on the data. Cox (1975) proposed the idea of sample-splitting to break up the dependence of variable selection and hypothesis testing, whereas Wasserman and Roeder (2009) studied sample-splitting in application to the lasso, marginal regression and forward step-wise regression. Meinshausen et al. (2009) extended the single-splitting proposal of Wasserman and Roeder (2009) to multi-splitting, which improved statistical power and reduced the number of falsely selected variables. Berk et al. (2013) instead considered simultaneous inference, which is universally valid under all possible model selection procedures without sample-splitting. More recently, Lee et al. (2016); Tibshirani et al. (2016) studied the geometry of the lasso and sequential regression, respectively, and proposed exact post-selection inference methods conditional on the random set of selected variables. See Taylor and Tibshirani (2015) for a review of post-selection inference procedures.

In a recent *Statistical Science* paper, Leeb et al. (2015) performed a simulation study, in which they obtained a set $\mathcal{M}$ using variable selection, and then calculated *"naïve" confidence intervals* for $\boldsymbol{\beta}^{(\mathcal{M})}$ using ordinary least squares, *without accounting for the fact that the set $\mathcal{M}$ was chosen based on the data.* Of course, conventional wisdom dictates that the resulting confidence intervals will be much too narrow. In fact, this is what Leeb et al. (2015) found, when they used best subset selection to construct the set $\mathcal{M}$. However, surprisingly, when the lasso was used to construct the set $\mathcal{M}$, the confidence intervals had approximately correct coverage, in the sense that (4) holds. This is in stark contrast to the existing literature!

In this paper, we present a theoretical justification for the empirical finding in Leeb et al. (2015) that selecting a set $\mathcal{M}$ based on the lasso and then constructing naïve confidence intervals based on the selected set leads to valid inference of the vector

$\boldsymbol{\beta}^{(\mathcal{M})}$. Furthermore, we make use of our theoretical findings in order to develop the *naïve score test*, a simple procedure for testing the null hypothesis $H^*_{0,j} : \beta^*_j = 0$ for $j = 1, \ldots, p$.

The rest of this paper is organized as follows. In Sections 2 and 3, we focus on post-selection inference: we seek to perform inference on $\boldsymbol{\beta}^{(\mathcal{M})}$ in (3), where $\mathcal{M}$ is selected based on the lasso, i.e., $\mathcal{M} = \hat{\mathcal{A}}_\lambda$ (6). In Section 2, we point out a previously overlooked fact: although $\hat{\mathcal{A}}_\lambda$ (6) is random, it converges in probability to a deterministic and non-data-dependent set. This result implies that we can use classical methods for inference on $\boldsymbol{\beta}^{(\mathcal{M})}$, when $\mathcal{M} = \hat{\mathcal{A}}_\lambda$. In Section 3, we provide empirical evidence in support of these theoretical findings. In Sections 4 and Section 5, we instead focus on the task of performing inference on $\boldsymbol{\beta}^*$ in (1). We propose the naïve score test in Section 4, and study its empirical performance in Section 5. We end with a discussion of future research directions in Section 6. Technical proofs are relegated to the online Supplementary Materials.

We now introduce some notation that will be used throughout the paper. We use "$\equiv$" to denote equalities by definition, and "$\asymp$" for the asymptotic order. We use $1\{\cdot\}$ for the indicator function; "$\vee$" and "$\wedge$" denote the maximum and minimum of two real numbers, respectively. For any real number $a \in \mathbb{R}$, $a_+ \equiv a \vee 0$. Given a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality and $-\mathcal{S} \equiv \mathcal{S}^c$ denotes its complement. We use bold upper case fonts to denote matrices, bold lower case fonts for vectors, and normal fonts for scalars. We use symbols with a superscript "$*$", e.g., $\boldsymbol{\beta}^*$ and $\mathcal{A}^* \equiv \mathrm{supp}(\boldsymbol{\beta}^*)$, to denote the true population parameters associated with the full linear model (1); we use symbols superscripted by a set in the parentheses, e.g., $\boldsymbol{\beta}^{(\mathcal{M})}$, to denote quantities related to the sub-model (3). Symbols subscripted by "λ" and with a hat, e.g., $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\mathcal{A}}_\lambda$, denote parameter estimates from the lasso estimator (2) with tuning parameter $\lambda > 0$; symbols subscripted by "λ" and without a hat, e.g., $\boldsymbol{\beta}_\lambda$ and $\mathcal{A}_\lambda \equiv \mathrm{supp}(\boldsymbol{\beta}_\lambda)$, are associated with the noiseless lasso estimator,

$$\boldsymbol{\beta}_\lambda \equiv \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \right] + \lambda \|\boldsymbol{b}\|_1 \right\}. \tag{7}$$

For any vector $\boldsymbol{b}$, matrix $\boldsymbol{\Sigma}$, and index sets $\mathcal{S}_1$ and $\mathcal{S}_2$, we use $\boldsymbol{b}_{\mathcal{S}_1}$ to denote the sub-vector of $\boldsymbol{b}$ comprised of elements of $\mathcal{S}_1$, and $\boldsymbol{\Sigma}_{(\mathcal{S}_1,\mathcal{S}_2)}$ to denote the sub-matrix of $\boldsymbol{\Sigma}$ with rows in $\mathcal{S}_1$ and columns in $\mathcal{S}_2$.

2 Theoretical Justification for Naïve Confidence Intervals

Recall that $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ was defined in (5). The simulation results of Leeb et al. (2015) suggest that if we perform ordinary least squares using the variables contained in the support set of the lasso, $\hat{\mathcal{A}}_\lambda$, then the classical confidence intervals associated with the least squares estimator,

$$\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \equiv \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y}, \tag{8}$$

have approximately correct coverage, where correct coverage means that for all $j \in \hat{\mathcal{A}}_\lambda$,

$$P\left(\beta_j^{(\hat{\mathcal{A}}_\lambda)} \in \mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}\right) \geq 1 - \alpha. \tag{9}$$

We reiterate that in (9), $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}$ is the confidence interval output by standard least squares software applied to the data $(\boldsymbol{y}, \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})$. This goes against our statistical intuition: it seems that by fitting a lasso model and then performing least squares on the selected set, we are peeking at the data twice, and thus we would expect the confidence interval $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}$ to be much too small.

In this section, we present a theoretical result that suggests that, in fact, this "double-peeking" might not be so bad. Our key insight is as follows: under appropriate assumptions, *the set of variables selected by the lasso is deterministic and non-data-dependent with high probability*. Thus, fitting a least squares model on the variables selected by the lasso does not really constitute peeking at the data twice: effectively, with high probability, we are only peeking at it once. That means that the naïve confidence intervals obtained from ordinary least squares will have approximately correct coverage, in the sense of (9).

We first introduce the required conditions for our theoretical result.

(**M1**) The design matrix $\boldsymbol{X}$ is deterministic, with columns in *general position*; see Definition 2.1. Furthermore, the columns of $\boldsymbol{X}$ are centered and standardized, i.e., for any $j = 1, \dots, p$, $\mathbf{1}^\top \boldsymbol{x}_j = 0$, $\boldsymbol{x}_j^\top \boldsymbol{x}_j = n$.

(**M2**) The error $\boldsymbol{\epsilon}$ in (1) has independent entries and sub-Gaussian tails. That is, there exist some constant $h > 0$ such that for all $x > 0$, we have $\Pr[|\epsilon_i| > x] < \exp(1 - hx^2)$, for all $i = 1, \dots, n$.

(**M3**) The sample size n, dimension p and tuning parameter λ satisfy

$$\sqrt{\frac{\log(p)}{n}} \frac{1}{\lambda} \to 0.$$

(**M4**) Recall that $\mathcal{A}^* \equiv \mathrm{supp}(\boldsymbol{\beta}^*)$. Let $\mathcal{S}^* \equiv \{j : |\beta_j^*| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$, where $q^* \equiv |\mathcal{A}^*| \equiv |\mathrm{supp}(\boldsymbol{\beta}^*)|$, and ϕ^* is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda \backslash \mathcal{S}^*}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right),$$

and

$$\left\|\boldsymbol{X}_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)} \boldsymbol{\beta}^*_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)}\right\|_2 = \mathcal{O}\left(\sqrt{\log(p)}\right),$$

where $\mathcal{A}_\lambda \equiv \mathrm{supp}(\boldsymbol{\beta}_\lambda)$, with $\boldsymbol{\beta}_\lambda$ defined in (7).

(**E**) Let $\hat{\boldsymbol{\Sigma}} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n$.

$$\limsup_{n\to\infty} \left\|\hat{\boldsymbol{\Sigma}}\right\|_2^2 \sqrt{\frac{q^*}{\log(p)}} < \infty.$$

In addition, there exists a constant $\phi^* > 0$, such that for any index set $\mathcal{I}$ with $|\mathcal{I}| = \mathcal{O}(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2)$, and all $\boldsymbol{a} \in \mathbb{R}^p$ that satisfy $\|\boldsymbol{a}_{\mathcal{I}^c}\|_1 \leq \|\boldsymbol{a}_{\mathcal{I}}\|_1$,

$$\liminf_{n\to\infty} \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{a}}{\|\boldsymbol{a}_{\mathcal{B}}\|_2^2} \geq \phi^{*2} > 0,$$

for any index set $\mathcal{B}$ such that $\mathcal{B} \supseteq \mathcal{I}$, $|\mathcal{B} \backslash \mathcal{I}| \leq |\mathcal{I}|$ and $\|\boldsymbol{a}_{\mathcal{B}^c}\|_\infty \leq \min_{j \in (\mathcal{B} \backslash \mathcal{I})} |a_j|$.

(**T**) Define $\boldsymbol{\tau}_\lambda$ based on the stationary condition of (7),

$$\boldsymbol{\tau}_\lambda = \frac{1}{\lambda}\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right). \tag{10}$$

Then,

$$\limsup_{n\to\infty} \left\|\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda^c}\right\|_\infty \leq 1-\delta \quad \text{and} \quad \frac{\sqrt{\log(p)/n}/\lambda}{\min_{j\in\mathcal{A}_\lambda\backslash\mathcal{S}^*}\left|\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}\right]^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}\right|_j} \to 0,$$

such that $\sqrt{\log(p)/n}/(\lambda\delta) \to 0$.

Definition 2.1 (Definition 1.1 in Dossal (2012))**.** *Let $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ be p points of $\mathbb{R}^n$. These points are said in general position if all affine subspaces of $\mathbb{R}^n$ of dimension $k < n \vee p$ contain at most $k+1$ points in $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$. Columns of matrix $\boldsymbol{X}$ are in general position if for all sign vectors $\boldsymbol{s} \in \{-1,1\}^p$, points $(s_1\boldsymbol{x}_1, \ldots, s_p\boldsymbol{x}_p)$ are in general position.*

Condition (**M1**), presented in Rosset et al. (2004); Dossal (2012); Tibshirani (2013), is a mild assumption that guarantees the uniqueness of $\boldsymbol{\beta}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$. (**M2**) enables the dimension p to grow at an exponential rate relative to the sample size n, i.e., $p = \mathcal{O}(\exp(n^\nu))$ for some $0 \leq \nu < 1$. (**M3**) requires the lasso tuning parameter λ to approach zero at a slightly slower rate than the ℓ_2-estimation and -prediction consistent rate $\lambda \asymp \sqrt{\log(p)/n}$; this helps further control the randomness of the error $\boldsymbol{\epsilon}$. Unfortunately, this condition complicates the task of tuning parameter selection; we further discuss this issue in Sections 3 and 6.

In (**M4**), the requirements that $\mathcal{S}^* \equiv \{j : |\beta_j^*| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$ and $\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\backslash\mathcal{S}^*}\|_\infty = \mathcal{O}(\sqrt{\log(p)/n})$ indicate that the regression coefficients of the variables in $\mathcal{A}_\lambda$ are either asymptotically no smaller than $\lambda\sqrt{q^*}$, or else no larger than $\sqrt{\log(p)/n}$. Given that λ is asymptotically slightly larger than $\sqrt{\log(p)/n}$ by (**M3**), these requirements imply that there needs to be a gap in signal strength of order at least $\sqrt{q^*}$ between the *strong signal variables* (those in $\mathcal{S}^*$) and the *weak signal* variables (those in $\mathcal{A}^*\backslash\mathcal{S}^*$). We note that this is substantially milder than the $\boldsymbol{\beta}$-min condition that is commonly used to establish model selection consistency (i.e., $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}^*] \to 1$) or the variable screening property (i.e., $\Pr[\hat{\mathcal{A}}_\lambda \supseteq \mathcal{A}^*] \to 1$) of the lasso (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; Bühlmann and van de Geer, 2011), which does not allow for the presence of weak signal variables. If we imposed additional stringent conditions on the design matrix (Buena et al., 2007; Zhang, 2009; Candés and Plan, 2009), then we could allow for the signal strength of $\boldsymbol{\beta}^*_{\mathcal{A}_\lambda \setminus \mathcal{S}^*}$ to achieve $\sqrt{\log(p) q^*/n}$ in magnitude, in which case there is no gap in signal strength. $\|\boldsymbol{X}_{\mathcal{A}^* \setminus (\mathcal{A}_\lambda \cup \mathcal{S}^*)} \boldsymbol{\beta}^*_{\mathcal{A}^* \setminus (\mathcal{A}_\lambda \cup \mathcal{S}^*)}\|_2 = \mathcal{O}(\sqrt{\log(p)})$ of (**M4**) implies that the total signal strength of weak signal variables that are not selected by the noiseless lasso cannot be too large.

While our theoretical results rely on (**M4**), an alternative set of assumptions of (**M4**) is presented in Section E of the online Supplementary Materials.

Condition (**E**) manifests the behavior of the eigenvalues of $\hat{\boldsymbol{\Sigma}}$. Specifically, the first part of (**E**) indicates that the maximum eigenvalue of $\hat{\boldsymbol{\Sigma}}$ cannot grow faster than the $\sqrt{\log(p)/q^*}$ rate. The second part of (**E**) is a modification of the $(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2, q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2, 1)$-restricted eigenvalue condition (Bickel et al., 2009; van de Geer and Bühlmann, 2009), which is a standard condition in the literature.

The first part of (**T**) requires that δ converges to zero at a slower rate than $\sqrt{\log(p)/n}/\lambda$, which means that λ does not converge to a transition point too fast, at which some variable enters or leaves $\mathcal{A}_\lambda$. Since $\sqrt{\log(p)/n}/\lambda \to 0$ by (**M3**), the second part of (**T**) requires that $\min_{j \in \mathcal{A}_\lambda \setminus \mathcal{S}^*} |[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda, \mathcal{A}_\lambda)}]^{-1} \boldsymbol{\tau}_{\lambda, \mathcal{A}_\lambda}|_j$ also does not converge to zero too fast. We empirically examine the stringency of Condition (**T**) in Section G of the online Supplementary Materials. We also compare the condition with the irrepresentable condition, which is required for variable selection consistency of lasso (Zhao and Yu, 2006). Given the discussion in Section 1, in particular, (5), when the irrepresentable condition fails to hold, sample splitting and exact post selection procedures may no longer test the population parameters, $\boldsymbol{\beta}^*_{\mathcal{M}}$. The empirical results in Section G of the online Supplementary Materials clearly indicate that Condition (**T**) is much more likely to hold than the irrepresentable condition.

Conditions (**M4**) and (**T**) critically depend on the set of variables selected by the noiseless lasso, $\mathcal{A}_\lambda$, which could be hard to interpret. However, as shown in Remarks 2.2

and 2.3, with some simple designs, we can simplify these two conditions to make them more interpretable. These two remarks are proven in Section F in the online Supplementary Materials.

Remark 2.2. *If the design matrix* $\boldsymbol{X}$ *is orthonormal, i.e.,* $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}$*, then* $\boldsymbol{\beta}_\lambda = \operatorname{sign}(\boldsymbol{\beta}^*)(|\boldsymbol{\beta}^*| - \lambda)_+$ *can be obtained by soft-thresholding* $\boldsymbol{\beta}^*$ *with threshold* λ *and for* n *sufficiently large,* $\mathcal{A}_\lambda = \mathcal{S}^*$*. Furthermore, in this case,*

$$\frac{\sqrt{\log(p)/n}/\lambda}{\min_{j \in \mathcal{A}_\lambda \setminus \mathcal{S}^*} \left| \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda, \mathcal{A}_\lambda)} \right]^{-1} \boldsymbol{\tau}_{\lambda, \mathcal{A}_\lambda} \right|_j} \to 0,$$

Remark 2.3. *If the covariance of design matrix* $\boldsymbol{X}$ *is block-diagonal, such that* $\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*, -\mathcal{A}^*)} = \boldsymbol{0}$*, then* $\mathcal{S}^* \subseteq \mathcal{A}_\lambda \subseteq \mathcal{A}^*$*. In this case, we can rephrase (***M4***) and (***T***) to make them hold for any set* $\mathcal{A}_\lambda = \mathcal{M}$ *such that* $\mathcal{S}^* \subseteq \mathcal{M} \subseteq \mathcal{A}^*$*.*

We now present Proposition 2.4, which is proven in Section A of the online Supplementary Materials.

Proposition 2.4. *Suppose conditions (***M1***)-(***M4***), (***E***) and (***T***) hold. Then, we have* $\lim_{n\to\infty} \Pr\left[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda\right] = 1$*, where* $\mathcal{A}_\lambda \equiv \operatorname{supp}(\boldsymbol{\beta}_\lambda)$*, with* $\boldsymbol{\beta}_\lambda$ *defined in* (7)*.*

It is important to emphasize the difference between the result in Proposition 2.4 and variable selection consistency of the lasso. Variable selection consistency asserts that $\Pr\left[\hat{\mathcal{A}}_\lambda = \mathcal{A}^*\right] \to 1$, and requires the stringent irrepresentable and $\boldsymbol{\beta}$-min conditions (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009). For estimation methods with folded-concave penalties (e.g., Fan and Li, 2001; Zhang, 2010), the irrepresentable condition may be relaxed. However, to achieve variable selection consistency, they still do not allow the existence of weak signal variables. In contrast, Proposition 2.4 asserts that under milder conditions, $\hat{\mathcal{A}}_\lambda$ converges with high probability to a deterministic set $\mathcal{A}_\lambda$ with cardinality smaller than n, *which is likely different from* $\mathcal{A}^*$. Based on Proposition 2.4, we could build asymptotically valid confidence intervals, as shown in Theorem 2.5.

Proposition 2.4 suggests that asymptotically, we "pay no price" for peeking at our data by performing the lasso: we should be able to perform downstream analyses on the subset of variables in $\hat{\mathcal{A}}_\lambda$ *as though we had obtained that subset without looking at the data*. This intuition will be formalized in Theorem 2.5.

Theorem 2.5, which is proven in Section B in the online Supplementary Materials, shows that $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ in (8) is asymptotically normal, with mean and variance suggested by classical least squares theory: that is, *the fact that $\hat{\mathcal{A}}_\lambda$ was selected by peeking at the data has no effect on the asymptotic distribution of $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$*. This result requires that λ be chosen in a non-data-adaptive way. Otherwise, $\mathcal{A}_\lambda$ will be affected by the random error $\boldsymbol{\epsilon}$ through λ, which complicates the distribution of $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$. Theorem 2.5 requires Condition (**W**), which is used to apply the Lindeberg-Feller Central Limit Theorem. This condition can be relaxed if the noise $\boldsymbol{\epsilon}$ is normally distributed.

(**W**) λ, $\boldsymbol{\beta}^*$ and $\boldsymbol{X}$ are such that $\lim_{n\to\infty} \|\mathbf{r}^w\|_\infty / \|\mathbf{r}^w\|_2 \to 0$, where

$$\mathbf{r}^w \equiv \boldsymbol{e}^j (\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top,$$

and $\boldsymbol{e}^j$ is the row vector of length $|\mathcal{A}_\lambda|$ with the entry corresponding to β_j^* equal to one, and zero otherwise.

Theorem 2.5. *Suppose (**M1**)–(**M4**), (**E**), (**T**) and (**W**) hold. Then, for any $j \in \hat{\mathcal{A}}_\lambda$,*

$$\frac{\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} - \beta_j^{(\hat{\mathcal{A}}_\lambda)}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\left[(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})^{-1} \right]_{(j,j)}}} \to_d \mathcal{N}(0,1), \tag{11}$$

where $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ is defined in (8) *and $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ in* (5), *and $\sigma_{\boldsymbol{\epsilon}}$ is the variance of $\boldsymbol{\epsilon}$ in* (1).

The error standard deviation $\sigma_{\boldsymbol{\epsilon}}$ in (11) is usually unknown. It can be estimated using various high-dimensional estimation methods, e.g., the scaled lasso (Sun and Zhang, 2012), cross-validation (CV) based methods (Fan et al., 2012) or method-of-moments based methods (Dicker, 2014); see a comparison study of high dimensional error variance estimation methods in Reid et al. (2016). Alternatively, Theorem 2.6 shows that we could also consistently estimate the error variance using the post-selection OLS residual sum of square (RSS).

Theorem 2.6. *Suppose (**M1**)–(**M4**), (**E**) and (**T**) hold, and* $\log(p)/(n - q_\lambda) \to 0$*, where* $q_\lambda \equiv |\mathcal{A}_\lambda|$*. Then*

$$\frac{1}{n - \hat{q}_\lambda} \left\| \boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \right\|_2^2 \to_p \sigma_{\boldsymbol{\epsilon}}^2, \tag{12}$$

where $\hat{q}_\lambda \equiv |\hat{\mathcal{A}}_\lambda|$.

Theorem 2.6 is proven in Section C in the online Supplementary Materials. In (12), $\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \tilde{\boldsymbol{\beta}}^{\hat{\mathcal{A}}_\lambda}$ is the fitted OLS residual on the sub-model (3). $\log(p)/(n - q_\lambda) \to 0$ is a weak condition. Since $\log(p)/n \to 0$, $\log(p)/(n - q_\lambda) \to 0$ is satisfied if $\lim_{n\to\infty} q_\lambda/n < 1$.

To summarize, in this section, we have provided a theoretical justification for a procedure that seems, intuitively, to be statistically unjustifiable:

1. Perform the lasso in order to obtain the support set $\hat{\mathcal{A}}_\lambda$;
2. Use least squares to fit the sub-model containing just the features in $\hat{\mathcal{A}}_\lambda$;
3. Use the classical confidence intervals from that least squares model, without accounting for the fact that $\hat{\mathcal{A}}_\lambda$ was obtained by peeking at the data.

Theorem 2.5 guarantees that the naïve confidence intervals in Step 3 will indeed have approximately correct coverage, in the sense of (9).

3 Numerical Examination of Naïve Confidence Intervals

In this section, we perform simulation studies to examine the coverage probability (9) of the *naïve* confidence intervals obtained by applying standard least squares software to the data $\boldsymbol{y}, \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}$.

Recall from Section 1 that (9) involves the probability that the confidence interval contains the quantity $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$, which in general does not equal the population regression coefficient vector $\boldsymbol{\beta}^*_{\hat{\mathcal{A}}_\lambda}$. Inference for $\boldsymbol{\beta}^*$ is discussed in Sections 4 and 5.

The results in this section complement simulation findings in Leeb et al. (2015).

3.1 Methods for Comparison

Following Theorem 2.5, for $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ defined in (8), and for each $j \in \hat{\mathcal{A}}_\lambda$, the 95% naïve confidence interval takes the form

$$\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)} \equiv \left(\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} - 1.96 \times \hat{\sigma}_{\boldsymbol{\epsilon}} \sqrt{\left[\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right]_{(j,j)}},\ \tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} + 1.96 \times \hat{\sigma}_{\boldsymbol{\epsilon}} \sqrt{\left[\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right]_{(j,j)}}\right). \quad (13)$$

where $\Phi_{\mathcal{N}}^{-1}[\cdot]$ is the quantile function of the standard normal distribution. In order to obtain the set $\hat{\mathcal{A}}_\lambda$, we must apply the lasso using some value of λ. Note that (**M3**) requires that $\lambda \succ \sqrt{\log(p)/n}$, which is slightly larger than the prediction optimal rate, $\lambda \asymp \sqrt{\log(p)/n}$ (Bickel et al., 2009; van de Geer and Bühlmann, 2009). Thus, we propose to use the tuning parameter value $\lambda_{1\mathrm{SE}}$, which is the largest value of λ for which the 10-fold CV prediction mean squared error (PMSE) is within one standard error of the minimum CV PMSE (see Section 7.10.1 in Hastie et al., 2009). We leave the optimal choice of tuning parameters to future research.

As a comparison, we also report the confidence intervals for $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ output by the R package `selectiveInference`, which implements the exact lasso post-selection inference procedure proposed in Lee et al. (2016). For exact lasso post-selection confidence intervals, we adopt the procedure in Section 7 of Lee et al. (2016) to choose its tuning parameters: we let $\lambda_{\mathrm{sup}} \equiv 2\mathbb{E}[\|\boldsymbol{X}^\top \boldsymbol{e}\|_\infty]/n$, where we simulate $\boldsymbol{e} \sim \mathcal{N}_n(\mathbf{0}, \hat{\sigma}_{\boldsymbol{\epsilon}}^2 \boldsymbol{I})$, and approximate its expectation based on the average of 1000 replicates. For fairer comparisons, we also report the performance of naïve confidence intervals with λ_{sup}. Unlike $\lambda_{1\mathrm{SE}}$, λ_{sup} does not depend on the randomness in $\boldsymbol{y}$.

In both approaches, the standard deviation of errors, $\sigma_{\boldsymbol{\epsilon}}$ in (1), is estimated using the scaled lasso (Sun and Zhang, 2012).

3.2 Simulation Set-Up

For the simulations, we consider two partial correlation settings for $\boldsymbol{X}$, generated based on (i) a scale-free graph and (ii) a stochastic block model (see, e.g., Kolaczyk, 2009), each containing $p = 100$ nodes. These settings are relaxations of the simple orthogonal and block-diagonal settings considered in Section 2, and are displayed in Figure 1.

In the scale-free graph setting, we used the `igraph` package in R to simulate an undirected, scale-free network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with power-law exponent parameter $\gamma = 5$, and edge density 0.05. Here, $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes in the graph, and $\mathcal{E}$ is the set of edges. This resulted in a total of $|\mathcal{E}| = 247$ edges in the graph. We then order the indices of the nodes in the graph so that the first, second, third, fourth, and fifth nodes correspond to the 10th, 20th, 30th, 40th, and 50th least-connected nodes in the graph.

In the stochastic block model setting, we first generate two dense Erdős-Rényi graphs (Erdős and Rényi, 1959; Gilbert, 1959) with five nodes and 95 nodes, respectively. In each graph, the edge density is 0.3. We then added edges randomly between these two graphs to achieve an inter-graph edge density of 0.05. The indices of the nodes are ordered so that the nodes in the five-node graph precede the remaining nodes.

Next, for both graph settings, we define the weighted adjacency matrix, $\boldsymbol{A}$, as follows:

$$A_{(j,k)} = \begin{cases} 1 & \text{for } j = k \\ \rho & \text{for } (j,k) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

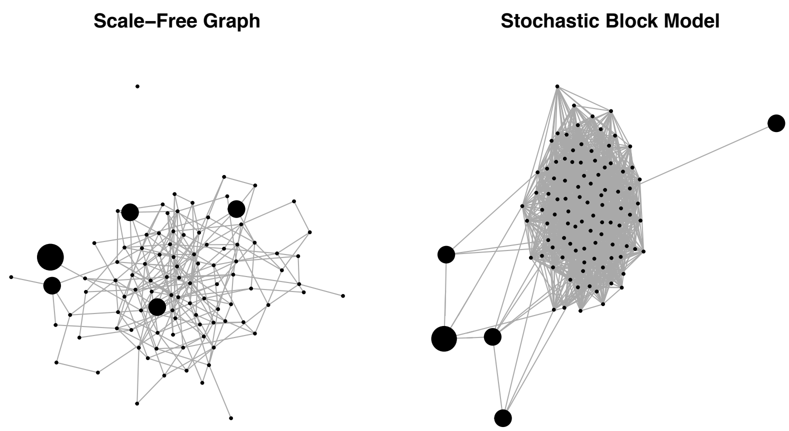
where $\rho \in \{0.2, 0.6\}$. We then set $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}$, and standardize $\boldsymbol{\Sigma}$ so that $\Sigma_{(j,j)} = 1$, for all $j = 1, \ldots, p$. We simulate observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and generate the outcome $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I}_n)$, $n \in \{300, 400, 500\}$, where

$$\beta_j^* = \begin{cases} 1 & \text{for } j = 1 \\ 0.1 & \text{for } 2 \leq j \leq 5 \\ 0 & \text{otherwise} \end{cases} \cdot$$

A range of error variances $\sigma_{\boldsymbol{\epsilon}}^2$ are used to produce signal-to-noise ratios, $\text{SNR} \equiv (\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}\boldsymbol{\beta}^*)/\sigma_{\boldsymbol{\epsilon}}^2 \in \{0.1, 0.3, 0.5\}$.

Throughout the simulations, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}^*$ are held fixed over $B = 1000$ repetitions of

Figure 1: The scale-free graph and stochastic block model settings. The size of a given node indicates the magnitude of the corresponding element of $\boldsymbol{\beta}^*$.



the simulation study, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

3.3 Simulation Results

We calculate the average length and coverage proportion of the 95% naïve confidence intervals, where the coverage proportion is defined as

$$\text{Coverage Proportion} \equiv \sum_{b=1}^{B} \sum_{j \in \hat{\mathcal{A}}_\lambda^b} 1\left\{\beta_j^{(\hat{\mathcal{A}}_\lambda^b)} \in \mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda^b),b}\right\} \Big/ \left|\hat{\mathcal{A}}_\lambda^b\right|, \tag{15}$$

where $\hat{\mathcal{A}}_\lambda^b$ and $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda^b),b}$ are the set of variables selected by the lasso in the bth repetition, and the 95% naïve confidence interval (13) for the jth variable in the bth repetition, respectively. Recall that $\beta_j^{(\hat{\mathcal{A}}_\lambda^b)}$ was defined in (5). In order to calculate the average length and coverage proportion associated with the exact lasso post selection procedure of Lee et al. (2016), we replace $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda^b),b}$ in (15) with the confidence interval output by the `selectiveinference` R package.

Tables 1 and 2 show the coverage proportion and average length of 95% naïve confidence intervals and 95% exact lasso post-selection confidence intervals under the scale-free graph and stochastic block model settings, respectively. The result shows that the coverage probability of the naïve and exact post-selection confidence intervals with tuning parameter λ_{sup} is approximately correct. This corroborates the findings in Leeb et al. (2015), in which the authors consider settings with $n = 30$ and $p = 10$. The coverage probability of the naïve confidence intervals with tuning parameter λ_{1SE} is slightly too small. Tables 1 and 2 also show that naïve confidence intervals are narrower than exact lasso post-selection confidence intervals, especially when the signal is weak.

Table 1: Coverage proportions and average lengths of 95% naïve confidence intervals with tuning parameters λ_{sup} and λ_{1SE}, and 95% exact post-selection confidence intervals under the scale-free graph partial correlation setting with $\rho \in \{0.2, 0.6\}$, sample size $n \in \{300, 400, 500\}$, dimension $p = 100$ and signal-to-noise ratio SNR $\in \{0.1, 0.3, 0.5\}$.

	ρ	0.2								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{sup}	0.949	0.949	0.949	0.950	0.959	0.959	0.956	0.964	0.965
	naïve λ_{sup}	0.951	0.948	0.948	0.977	0.958	0.958	0.969	0.964	0.964
	naïve λ_{1SE}	0.922	0.936	0.928	0.944	0.937	0.930	0.950	0.938	0.930
Length	exact λ_{sup}	1.902	0.427	0.327	1.148	0.368	0.284	0.815	0.327	0.254
	naïve λ_{sup}	0.714	0.418	0.325	0.623	0.363	0.282	0.561	0.325	0.252
	naïve λ_{1SE}	0.719	0.418	0.324	0.625	0.363	0.282	0.559	0.325	0.252
	ρ	0.6								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{sup}	0.960	0.954	0.954	0.956	0.955	0.956	0.953	0.945	0.945
	naïve λ_{sup}	0.962	0.950	0.950	0.972	0.954	0.954	0.966	0.944	0.942
	naïve λ_{1SE}	0.918	0.939	0.932	0.974	0.937	0.932	0.968	0.936	0.928
Length	exact λ_{sup}	1.669	0.426	0.326	1.081	0.365	0.283	0.781	0.326	0.255
	naïve λ_{sup}	0.711	0.418	0.324	0.624	0.363	0.281	0.559	0.324	0.251
	naïve λ_{1SE}	0.716	0.416	0.323	0.623	0.361	0.280	0.556	0.324	0.251

In addition, to evaluate whether $\hat{\mathcal{A}}_\lambda$ is deterministic, among repetitions that $\hat{\mathcal{A}}_\lambda^b \neq \emptyset$ (there is no confidence interval if $\hat{\mathcal{A}}_\lambda^b = \emptyset$), we also calculate the proportion of $\hat{\mathcal{A}}_\lambda^b = \mathcal{D}$, where $\mathcal{D}$ is the most common $\hat{\mathcal{A}}_\lambda^b$, $b = 1, \ldots, 1000$. The result is summarized in Table 3, which shows that $\hat{\mathcal{A}}_\lambda$ is almost deterministic with tuning parameter λ_{sup}. With λ_{1SE}, due to the randomness in the tuning parameter, $\hat{\mathcal{A}}_\lambda$ is less deterministic, which may explain the result that the coverage probability is slightly smaller than the desired level in this case.

Table 2: Coverage proportions and average lengths of 95% naïve confidence intervals with tuning parameters λ_{sup} and λ_{1SE}, and 95% exact post-selection confidence intervals under the stochastic block model setting. Details are as in Table 1.

	ρ	0.2								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{sup}	0.957	0.948	0.948	0.948	0.952	0.952	0.949	0.958	0.958
	naïve λ_{sup}	0.958	0.946	0.946	0.969	0.952	0.952	0.972	0.956	0.956
	naïve λ_{1SE}	0.907	0.938	0.935	0.938	0.939	0.933	0.947	0.928	0.919
Length	exact λ_{sup}	1.843	0.427	0.325	1.177	0.363	0.285	0.812	0.323	0.251
	naïve λ_{sup}	0.704	0.415	0.322	0.615	0.360	0.279	0.554	0.322	0.250
	naïve λ_{1SE}	0.710	0.414	0.321	0.617	0.359	0.279	0.554	0.322	0.250
	ρ	0.6								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{sup}	0.961	0.959	0.959	0.946	0.957	0.958	0.947	0.946	0.945
	naïve λ_{sup}	0.969	0.958	0.958	0.961	0.956	0.956	0.968	0.945	0.945
	naïve λ_{1SE}	0.925	0.941	0.934	0.919	0.927	0.921	0.959	0.944	0.940
Length	exact λ_{sup}	1.759	0.420	0.323	1.133	0.361	0.280	0.778	0.323	0.250
	naïve λ_{sup}	0.703	0.412	0.320	0.614	0.358	0.278	0.552	0.321	0.249
	naïve λ_{1SE}	0.705	0.411	0.319	0.612	0.358	0.278	0.553	0.321	0.249

4 Inference for β^* With the Naïve Score Test

Sections 2 and 3 focused on the task of developing confidence intervals for $\boldsymbol{\beta}^{(\mathcal{M})}$ in (3), where $\mathcal{M} = \hat{\mathcal{A}}_\lambda$, the set of variables selected by the lasso. However, recall from (5) that typically $\boldsymbol{\beta}^{(\mathcal{M})} \neq \boldsymbol{\beta}^*_{\mathcal{M}}$, where $\boldsymbol{\beta}^*$ was introduced in (1).

In this section, we shift our focus to performing inference on $\boldsymbol{\beta}^*$. We will exploit Proposition 2.4 to develop a simple approach for testing $H^*_{0,j} : \beta^*_j = 0$, for $j = 1, \ldots, p$.

Recall that in the low-dimensional setting, the classical score statistic for the hypothesis $H^*_{0,j} : \beta^*_j = 0$ is proportional to $\boldsymbol{x}_j^T(\boldsymbol{y} - \hat{\boldsymbol{y}}^0)$, where $\hat{\boldsymbol{y}}^0$ is the vector of fitted values that results from least squares regression of $\boldsymbol{y}$ onto the $p-1$ features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p$. In order to adapt the classical score test statistic to the high-dimensional setting, we define the *naïve score test statistic* for testing $H^*_{0,j} : \beta^*_j = 0$ as

$$S^j \equiv \boldsymbol{x}_j^\top \left(\boldsymbol{y} - \tilde{\boldsymbol{y}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}\right) \equiv \boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}\right) \boldsymbol{y}, \tag{16}$$

where

$$\tilde{\boldsymbol{y}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})} \equiv \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}} \tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})},$$

Table 3: Among repetitions that $\hat{\mathcal{A}}^b_\lambda \neq \emptyset$, the proportion of $\hat{\mathcal{A}}^b_\lambda$ that equals the most common $\hat{\mathcal{A}}^b_\lambda$, $b = 1, \ldots, 1000$, under the scale-free graph and stochastic block model settings with tuning parameters λ_{sup} and λ_{1SE}. In the simulation, $\rho \in \{0.2, 0.6\}$, sample size $n \in \{300, 400, 500\}$, dimension $p = 100$ and signal-to-noise ratio SNR $\in \{0.1, 0.3, 0.5\}$.

ρ	0.2								
n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Scale-free λ_{sup}	1.000	0.999	0.999	1.000	0.999	0.999	0.999	1.000	1.000
Scale-free λ_{1SE}	0.976	0.960	0.943	0.978	0.981	0.967	0.990	0.978	0.958
Stochastic block λ_{sup}	0.999	0.997	0.997	1.000	0.999	0.999	1.000	1.000	0.999
Stochastic block λ_{1SE}	0.966	0.969	0.963	0.984	0.968	0.959	0.990	0.980	0.962
ρ	0.6								
n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Scale-free λ_{sup}	0.999	0.996	0.996	0.999	1.000	1.000	1.000	0.999	0.996
Scale-free λ_{1SE}	0.969	0.970	0.958	0.986	0.978	0.965	0.992	0.986	0.967
Stochastic block λ_{sup}	1.000	0.999	0.999	1.000	0.999	0.999	1.000	1.000	1.000
Stochastic block λ_{1SE}	0.972	0.963	0.953	0.987	0.969	0.955	0.997	0.979	0.967

and

$$\mathbf{P}^{(\hat{\mathcal{A}}_\lambda \backslash \{j\})} \equiv \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \backslash \{j\}} \left(\boldsymbol{X}^\top_{\hat{\mathcal{A}}_\lambda \backslash \{j\}} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \backslash \{j\}} \right)^{-1} \boldsymbol{X}^\top_{\hat{\mathcal{A}}_\lambda \backslash \{j\}}$$

is the orthogonal projection matrix onto the set of variables in $\hat{\mathcal{A}}_\lambda \backslash \{j\}$. $\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda \backslash \{j\})}$ is defined in (8). In (16), the notation $\hat{\mathcal{A}}_\lambda \backslash \{j\}$ represents the set $\hat{\mathcal{A}}_\lambda$ in (6) with j removed, if $j \in \hat{\mathcal{A}}_\lambda$. If $j \notin \hat{\mathcal{A}}_\lambda$, then $\hat{\mathcal{A}}_\lambda \backslash \{j\} = \hat{\mathcal{A}}_\lambda$.

In Theorem 4.1, we will derive the asymptotic distribution for S^j under $H^*_{0,j} : \beta^*_j = 0$. We first introduce two new conditions.

First, we require that the total signal strength of variables not selected by the noiseless lasso, (7), is small.

(**M4***) Recall that $\mathcal{A}^* \equiv \text{supp}(\boldsymbol{\beta}^*)$. Let $\mathcal{S}^* \equiv \{j : |\beta^*_j| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$, where $q^* \equiv |\mathcal{A}^*| \equiv |\text{supp}(\boldsymbol{\beta}^*)|$, and ϕ^* is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda \backslash \mathcal{S}^*}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right),$$

and

$$\left\|\boldsymbol{X}_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)} \boldsymbol{\beta}^*_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)}\right\|_2 = \mathcal{O}(1),$$

where $\mathcal{A}_\lambda \equiv \text{supp}(\boldsymbol{\beta}_\lambda)$, with $\boldsymbol{\beta}_\lambda$ defined in (7).

Condition (**M4***) closely resembles (**M4**), which was required for Theorem 2.5 in Section 2. The only difference between the two is that (**M4***) requires $\|\boldsymbol{X}_{\mathcal{A}^*\backslash(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\backslash(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\|_2 = \mathcal{o}(1)$, whereas (**M4**) requires only that $\|\boldsymbol{X}_{\mathcal{A}^*\backslash(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\backslash(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\|_2 = \mathcal{O}(\sqrt{\log(p)})$. In other words, testing the population regression parameter $\boldsymbol{\beta}^*$ in (1) requires more stringent assumptions than constructing confidence intervals for the parameters in the sub-model (3).

The following condition, required to apply the Lindeberg-Feller Central Limit Theorem, can be relaxed if the noise $\boldsymbol{\epsilon}$ in (1) is normally distributed.

(**S**) λ, $\boldsymbol{\beta}^*$ and $\boldsymbol{X}$ satisfy $\lim_{n\to\infty}\|\mathbf{r}^s\|_\infty/\|\mathbf{r}^s\|_2 = 0$, where $\mathbf{r}^s \equiv \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\backslash\{j\})}\right)\boldsymbol{x}_j$.

We now present Theorem 4.1, which is proven in Section D of the online Supplementary Materials.

Theorem 4.1. *Suppose (**M1**)–(**M3**), (**M4***), (**E**), (**T**) and (**S**) hold. For any* $j = 1,\ldots,p$, *under the null hypothesis* $H^*_{0,j} : \beta^*_j = 0$,

$$T \equiv \frac{S^j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{x}_j^\top\left(\mathbf{I}_n - \mathbf{P}^{(\hat{\mathcal{A}}_\lambda\backslash\{j\})}\right)\boldsymbol{x}_j}} \to_d \mathcal{N}(0,1), \tag{17}$$

where S^j *was defined in* (16), *and where* $\sigma_{\boldsymbol{\epsilon}}$ *is the variance of* $\boldsymbol{\epsilon}$ *in* (1).

Theorem 4.1 states that the distribution of the naïve score test statistic S^j is asymptotically the same as if $\hat{\mathcal{A}}_\lambda$ were a fixed set, as opposed to being selected by fitting a lasso model on the data. Based on (17), we reject the null hypothesis $H^*_{0,j} : \beta^*_j = 0$ at level $\alpha > 0$ if $|T| > \Phi^{-1}_{\mathcal{N}}(1-\alpha/2)$, where $\Phi^{-1}_{\mathcal{N}}(\cdot)$ is the quantile function of the standard normal distribution function.

We emphasize that Theorem 4.1 holds for any variable $j = 1,\ldots,p$, and thus can be used to test $H^*_{0,j} : \beta^*_j = 0$, for all $j = 1,\ldots,p$. (This is in contrast to Theorem 2.5, which concerns confidence intervals for the parameters in the sub-model (3) consisting of the variables in $\hat{\mathcal{A}}_\lambda$, and hence holds only for $j \in \hat{\mathcal{A}}_\lambda$.)

5 Numerical Examination of the Naïve Score Test

In this section, we compare the performance of the naïve score test (16) to three recent proposals from the literature for testing $H^*_{0,j} : \beta^*_j = 0$: namely, LDPE (Zhang and Zhang, 2014; van de Geer et al., 2014), SSLasso (Javanmard and Montanari, 2014a), and the decorrelated score test (dScore; Ning and Liu, 2016). R code for SSLasso, and dScore was provided by the authors; LDPE is implemented in the R package `hdi`. For the naïve score test, we estimate $\sigma_{\boldsymbol{\epsilon}}$, the standard deviation of the errors in (1), using the scaled lasso (Sun and Zhang, 2012).

All four of these methods require us to select the value of the lasso tuning parameter. For LDPE, SSLasso, and dScore, we use 10-fold cross-validation to select the tuning parameter value that produces the smallest cross-validated mean square error, λ_{min}. As in the numerical study of the naïve confidence intervals in Section 3, we implement the naïve score test using the tuning parameter value λ_{1SE} and λ_{sup}.

Unless otherwise noted, all tests are performed at a significance level of 0.05.

In Section 5.1, we investigate the powers and type-I errors of the above tests in simulation experiments. Section 5.2 contains an analysis of a glioblastoma gene expression dataset.

5.1 Power and Type-I Error

5.1.1 Simulation Set-Up

In this section, we adapt the scale-free graph and the stochastic block model presented in Section 3.2 to have $p = 500$.

In the scale-free graph setting, we generate a scale-free graph with $\gamma = 5$, edge density 0.05, and $p = 500$ nodes. The resulting graph has $|\mathcal{E}| = 6237$ edges. We order the nodes in the graph so that jth node is the $(30 \times j)$th least-connected node in the graph, for $1 \leq j \leq 10$. For example, the 4th node is the 120th least-connected node in the graph.

In the stochastic block model setting, we generate two dense Erdős-Rényi graphs with ten nodes and 490 nodes, respectively; each has an intra-graph edge density of

0.3. The node indices are ordered so that the nodes in the smaller graph precede those in the larger graph. We then randomly connect nodes between the two graphs in order to obtain an inter-graph edge density of 0.05.

Next, for both graph settings, we generate $\boldsymbol{A}$ as in (14), where $\rho \in \{0.2, 0.6\}$. We then set $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}$, and standardize $\boldsymbol{\Sigma}$ so that $\Sigma_{(j,j)} = 1$, for all $j = 1, \ldots, p$. We simulate observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim_{i.i.d.} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, and generate the outcome $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I}_n)$, $n \in \{100, 200, 400\}$, where

$$\beta_j^* = \begin{cases} 1 & \text{for } 1 \leq j \leq 3 \\ 0.1 & \text{for } 4 \leq j \leq 10 \\ 0 & \text{otherwise} \end{cases} \cdot$$

A range of error variances $\sigma_{\boldsymbol{\epsilon}}^2$ are used to produce signal-to-noise ratios, SNR $\equiv (\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}\boldsymbol{\beta}^*)/\sigma_{\boldsymbol{\epsilon}}^2 \in \{0.1, 0.3, 0.5\}$.

We hold $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}^*$ fixed over $B = 100$ repetitions of the simulation, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

5.1.2 Simulation Results

For each test, the average power on the strong signal variables, the average power on the weak signal variables, and the average type-I error rate are defined as

$$\text{Power}_{\text{strong}} \equiv \frac{1}{B}\frac{1}{3}\sum_{b=1}^{B}\sum_{j:\beta_j^*=1} 1\{p_{jb} < 0.05\}, \tag{18}$$

$$\text{Power}_{\text{weak}} \equiv \frac{1}{B}\frac{1}{7}\sum_{b=1}^{B}\sum_{j:\beta_j^*=0.1} 1\{p_{jb} < 0.05\}, \tag{19}$$

$$\text{Type-1 Error} \equiv \frac{1}{B}\frac{1}{490}\sum_{b=1}^{B}\sum_{j:\beta_j^*=0} 1\{p_{jb} < 0.05\}, \tag{20}$$

respectively. In (18)–(20), p_{jb} is the p-value associated with null hypothesis $H_{0,j}^* : \beta_j^* = 0$ in the bth simulated data set. In the simulations, the graphs and $\boldsymbol{\beta}^*$ are held fixed over $B = 100$ repetitions of the simulation study, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

Tables 4 and 5 summarize the results in the two simulation settings. Naïve score test with λ_{sup} has slightly worse control of type-1 error rate and better power than the other four methods, which have approximate control over the type-I error rate and comparable power.

5.2 Application to Glioblastoma Data

We investigate a glioblastoma gene expression data set previously studied in Horvath et al. (2006). For each of 130 patients, a survival outcome is available; we removed the twenty patients who were still alive at the end of the study. This resulted in a data set with $n = 110$ observations. The gene expression measurements were normalized using the method of Gautier et al. (2004). We limited our analysis to $p = 3600$ highly-connected genes (Zhang and Horvath, 2005; Horvath and Dong, 2008). The normalized data can be found at the website of Dr. Steve Horvath of UCLA Biostatistics: `http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/ASPMgene/`. We log-transformed the survival response and centered it to have mean zero. Furthermore, we log-transformed the expression data, and then standardized each gene to have mean zero and standard deviation one across the $n = 110$ observations.

Our goal is to identify individual genes whose expression levels are associated with survival time, after adjusting for the other 3599 genes in the data set. With family-wise error rate (FWER) controlled at level 0.1 using the Holm procedure (Holm, 1979), the naïve score test identifies three such genes: CKS2, H2AFZ, and RPA3. You et al. (2015) observed that CKS2 is highly expressed in glioma. Vardabasso et al. (2014) found that histone genes, of which H2AFZ is one, are related to cancer progression. Jin et al. (2015) found that RPA3 is associated with glioma development. As a comparison, SSLasso finds two genes associated with patient survival: PPAP2C and RGS3. LDPE and dScore identify no genes at FWER of 0.1.

6 Discussion

In this paper, we examined a very naïve two-step approach to high-dimensional inference:

1. Perform the lasso in order to select a small set of variables, $\hat{\mathcal{A}}_\lambda$.
2. Fit a least squares regression model using just the variables in $\hat{\mathcal{A}}_\lambda$, and make use of standard regression inference tools. Make no adjustment for the fact that $\hat{\mathcal{A}}_\lambda$ was selected based on the data.

It seems clear that this naïve approach is problematic, since we have peeked at the data twice, but are not accounting for this double-peeking in our analysis.

In this paper, we have shown that under appropriate assumptions, $\hat{\mathcal{A}}_\lambda$ converges with high probability to a deterministic set, $\mathcal{A}_\lambda$. This key insight allows us to establish that the confidence intervals resulting from the aforementioned naïve two-step approach have asymptotically correct coverage, in the sense of (4). This constitutes a theoretical justification for the recent simulation findings of Leeb et al. (2015). Furthermore, we used this key insight in order to establish that the score test that results from the naïve two-step approach has asymptotically the same distribution as though the selected set of variables had been fixed in advance; thus, it can be used to test the null hypothesis $H^*_{0,j} : \beta^*_j = 0$, $j = 1, \ldots, p$.

Our simulation results corroborate our theoretical findings. In fact, we find essentially no difference between the empirical performance of these naïve proposals, and a host of other recent proposals in the literature for high-dimensional inference (Javanmard and Montanari, 2014a; Zhang and Zhang, 2014; van de Geer et al., 2014; Lee et al., 2016; Ning and Liu, 2016).

From a bird's-eye view, the recent literature on high-dimensional inference falls into two camps. The work of Wasserman and Roeder (2009); Meinshausen et al. (2009); Berk et al. (2013); Lee et al. (2016); Tibshirani et al. (2016) focuses on performing inference on the sub-model (3), whereas the work of Javanmard and Montanari (2013, 2014a,b); Zhang and Zhang (2014); van de Geer et al. (2014); Ning and Liu (2016) focuses on testing hypotheses associated with (1). In this paper, we have shown that the confidence intervals that result from the naïve approach can be used to perform inference on the sub-model (3), whereas the score test that results from the naïve approach can be used to test hypotheses associated with (1). Thus, the naïve approach to inference considered in this paper serves to unify these two camps of high-dimensional inference.

In the era of big data, simple analyses that are easy to apply and easy to understand are especially attractive to scientific investigators. Therefore, a careful investigation of such simple approaches is worthwhile, in order to determine which have the potential to yield accurate results, and which do not. We do not advocate applying the naïve two-step approach described above in most practical data analysis settings: we are confident that in practice, our intuition is correct, and this approach will perform poorly when the sample size is small or moderate, and/or the assumptions are not met. However, in very large data settings, our results suggest that this naïve approach may indeed be viable for high-dimensional inference, and warrants further investigation.

When choosing among existing inference procedures based on lasso, the *target of inference* should be taken into consideration. The target of inference can either be the population parameters, $\boldsymbol{\beta}^*$ in (1), or the parameters induced by the sub-model chosen by lasso, $\boldsymbol{\beta}^{(\mathcal{M})}$ in (5). Sample-splitting (Wasserman and Roeder, 2009; Meinshausen et al., 2009) and exact post selection (Lee et al., 2016; Tibshirani et al., 2016) methods provide valid inferences for $\boldsymbol{\beta}^{(\mathcal{M})}$. The latter is a particularly appealing choice for inference on $\boldsymbol{\beta}^{(\mathcal{M})}$, as it provides non-asymptotic confidence intervals under minimal assumptions. However, as we discussed in Section 1, $\boldsymbol{\beta}^{(\mathcal{M})}$ is, in general, different from $\boldsymbol{\beta}^*_{\mathcal{M}}$. A set of sufficient conditions for $\boldsymbol{\beta}^{(\mathcal{M})} = \boldsymbol{\beta}^*_{\mathcal{M}}$ is the irrepresentable condition together with a beta-min condition. Unfortunately, these assumptions are unverifiable and may not hold in practice. In contrast, debiased lasso tests (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2013, 2014a; Ning and Liu, 2016) provide asymptotically valid inference for entries of $\boldsymbol{\beta}^*$, without requiring a beta-min condition. However, these method require sparsity of the inverse covariance matrix of covariates, $\boldsymbol{\Sigma}^{-1}$, which is also unverifiable. While our theoretical analysis and empirical studies suggests that the naïve two-step approach described above requires less stringent assumptions than beta-min and irrepresentability, this method is also asymptotic and relies on unverifiable assumptions.

We close with some suggestions for future research. One reviewer brought up an interesting comment: Methods with folded-concave penalties (e.g., Fan and Li, 2001; Zhang, 2010) require milder conditions to achieve variable selection consistency than the lasso, i.e., $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}^*] \to 1$. Inspired by this observation, we wonder whether Fan and Li (2001); Zhang (2010) also require milder conditions to achieve $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda] \to 1$. If so, then we could replace lasso with Fan and Li (2001); Zhang (2010) in the variable selection step, and improve the robustness of the naïve approaches. We believe this could be a fruitful area of future research. In addition, extending the proposed theory and methods to generalized linear models and M-estimators may also be fruitful areas for future research.

A Proof of Proposition 2.4

We first state and prove Lemmas A.1–A.5, which are required to prove Proposition 2.4.

Lemma A.1. *Suppose (*$\mathbf{M1}$*) holds. Then,* $\boldsymbol{\beta}_\lambda$ *and* $\hat{\boldsymbol{\beta}}_\lambda$ *as defined in* (7) *and* (2)*, respectively, are unique.*

Proof. First, by, e.g., Lemma 3 of Tibshirani (2013), (**M1**) implies that $\hat{\boldsymbol{\beta}}_\lambda$ is unique. We now prove $\boldsymbol{\beta}_\lambda$ is also unique. The proof is similar to the proof of Lemmas 1 and 3 in Tibshirani (2013).

To show $\boldsymbol{\beta}_\lambda$ is unique, we first show that the fitted value of $\boldsymbol{X}\boldsymbol{\beta}_\lambda$ is unique. This is because, suppose, to the contrary, that we have two solutions to the problem (7), $\boldsymbol{\beta}_\lambda^{\mathrm{I}}$ and $\boldsymbol{\beta}_\lambda^{\mathrm{II}}$, which give different fitted values, $\boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{I}} \neq \boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{II}}$, but achieve the same minimum value of the objective function, c, i.e.,

$$\frac{1}{2n}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{I}}\right\|_2^2\right] + \lambda\left\|\boldsymbol{\beta}_\lambda^{\mathrm{I}}\right\|_1 = \frac{1}{2n}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right\|_2^2\right] + \lambda\left\|\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right\|_1 = c. \tag{21}$$

Then the value of the objective function of $\boldsymbol{\beta}_\lambda^{\mathrm{I}}/2 + \boldsymbol{\beta}_\lambda^{\mathrm{II}}/2$ is

$$
\begin{aligned}
&\frac{1}{2n}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}\left(\frac{1}{2}\boldsymbol{\beta}_\lambda^{\mathrm{I}} + \frac{1}{2}\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right)\right\|_2^2\right] + \lambda\left\|\frac{1}{2}\boldsymbol{\beta}_\lambda^{\mathrm{I}} + \frac{1}{2}\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right\|_1 \\
&< \frac{1}{4n}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{I}}\right\|_2^2\right] + \frac{\lambda}{2}\left\|\boldsymbol{\beta}_\lambda^{\mathrm{I}}\right\|_1 + \frac{1}{4n}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right\|_2^2\right] + \frac{\lambda}{2}\left\|\boldsymbol{\beta}_\lambda^{\mathrm{II}}\right\|_1 = c. \qquad (22)
\end{aligned}
$$

The inequality is due to the strict convexity of the squared ℓ_2 norm function and the convexity of the ℓ_1 norm function. Thus, $\boldsymbol{\beta}_\lambda^{\mathrm{I}}/2 + \boldsymbol{\beta}_\lambda^{\mathrm{II}}/2$ achieves a smaller value of the objective function than either $\boldsymbol{\beta}_\lambda^{\mathrm{I}}$ or $\boldsymbol{\beta}_\lambda^{\mathrm{II}}$, which is a contradiction. Hence, all solutions to the problem (7) have the same fitted value.

Therefore, based on the stationary condition in (10),

$$\lambda n \boldsymbol{\tau}_\lambda = \boldsymbol{X}^\top \boldsymbol{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda\right), \tag{23}$$

$\boldsymbol{\tau}_\lambda$ is unique. Define $\mathcal{T}_\lambda \equiv \{j : |\tau_{\lambda,j}| = 1\}$. $\mathcal{T}_\lambda$ is also unique. Furthermore, because $|\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}| = \mathbf{1}$, $\mathcal{T}_\lambda \supseteq \mathcal{A}_\lambda$, and

$$\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda^c} = \mathbf{0}. \tag{24}$$

Also according to the stationary condition in (10), and using the fact that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{y} - \boldsymbol{\epsilon}$, we have

$$\lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda - \boldsymbol{\epsilon}\right) = \boldsymbol{X}_{\mathcal{T}_\lambda}^\top\left(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}_\lambda}\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} - \boldsymbol{\epsilon}\right). \tag{25}$$

The last equality holds because as shown in (24), $\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda^c} = \mathbf{0}$. Equation (25) indicates that $\lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda}$ is in the row space of $\boldsymbol{X}_{\mathcal{T}_\lambda}$, or $\lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}^\top(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda}$, where $(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top)^+$ is the Moorse-Penrose pseudoinverse of $\boldsymbol{X}_{\mathcal{T}_\lambda}^\top$. Properties of the Moorse-Penrose pseudoinverse include $(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top)^+ = (\boldsymbol{X}_{\mathcal{T}_\lambda}^+)^\top$, $\boldsymbol{X}_{\mathcal{T}_\lambda}^+ = (\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \boldsymbol{X}_{\mathcal{T}_\lambda})^+ \boldsymbol{X}_{\mathcal{T}_\lambda}^\top$ and $\boldsymbol{X}_{\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}\boldsymbol{X}_{\mathcal{T}_\lambda}^+\boldsymbol{X}_{\mathcal{T}_\lambda}$.

Rearranging terms in (25), and pluging in $\lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}^\top (\boldsymbol{X}_{\mathcal{T}_\lambda}^\top)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda}$, we get

$$\begin{aligned}\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} &= \boldsymbol{X}_{\mathcal{T}_\lambda}^\top (\boldsymbol{y} - \boldsymbol{\epsilon}) - \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \\ &= \boldsymbol{X}_{\mathcal{T}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \right)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \right). \end{aligned} \tag{26}$$

Hence, according to (26) and the properties of the Moorse-Penrose pseudoinverse,

$$\begin{aligned}\boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} &= \boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{X}_{\mathcal{T}_\lambda}^+ \boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} \\ &= \boldsymbol{X}_{\mathcal{T}_\lambda} \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \boldsymbol{X}_{\mathcal{T}_\lambda} \right)^+ \boldsymbol{X}_{\mathcal{T}_\lambda}^\top \boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} \\ &= \boldsymbol{X}_{\mathcal{T}_\lambda} \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \boldsymbol{X}_{\mathcal{T}_\lambda} \right)^+ \boldsymbol{X}_{\mathcal{T}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \right)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \right) \\ &= \boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{X}_{\mathcal{T}_\lambda}^+ \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \right)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \right). \end{aligned} \tag{27}$$

Thus,

$$\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}^+ \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \right)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \right) + \boldsymbol{d}, \tag{28}$$

where $\boldsymbol{X}_{\mathcal{T}_\lambda} \boldsymbol{d} = \boldsymbol{0}$. Therefore, if $\text{null}(\boldsymbol{X}_{\mathcal{T}_\lambda}) = \{\boldsymbol{0}\}$, $\boldsymbol{d} = \boldsymbol{0}$. Because $\boldsymbol{\tau}_\lambda$ and $\mathcal{T}_\lambda$ are unique based on (23), $\text{null}(\boldsymbol{X}_{\mathcal{T}_\lambda}) = \{\boldsymbol{0}\}$ also implies $\boldsymbol{\beta}_\lambda$ is unique, i.e.,

$$\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda} = \boldsymbol{X}_{\mathcal{T}_\lambda}^+ \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\mathcal{T}_\lambda}^\top \right)^+ \lambda n \boldsymbol{\tau}_{\lambda,\mathcal{T}_\lambda} \right), \tag{29}$$

$$\boldsymbol{\beta}_{\lambda,\mathcal{T}_\lambda^c} = \boldsymbol{0}. \tag{30}$$

To see that (**M1**) implies $\text{null}(\boldsymbol{X}_{\mathcal{T}_\lambda}) = \{\boldsymbol{0}\}$, we use a similar argument as Tibshirani (2013). Specifically, we assume to the contrary, $\text{null}(\boldsymbol{X}_{\mathcal{T}_\lambda}) \neq \{\boldsymbol{0}\}$. Then, for any $j \in \mathcal{T}_\lambda$, we can write $\boldsymbol{x}_j = \sum_{k \in \mathcal{T}_\lambda \setminus \{j\}} a_k \boldsymbol{x}_k$, and multiplying both sides by $\tau_{\lambda,j}$,

$$\tau_{\lambda,j} \boldsymbol{x}_j = \sum_{k \in \mathcal{T}_\lambda \setminus \{j\}} \tau_{\lambda,j} a_k \boldsymbol{x}_k. \tag{31}$$

Multiplying both sides of (31) by $\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda)/n$,

$$\tau_{\lambda,j}\frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right) = \sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} \tau_{\lambda,j}a_k\frac{1}{n}\boldsymbol{x}_k^\top \boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right). \tag{32}$$

Based on the stationary condition in (10),

$$\lambda\tau_{\lambda,j} = \frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right), \tag{33}$$

(32) implies that $\tau_{\lambda,j}^2 = \sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} \tau_{\lambda,j}a_k\tau_{\lambda,k}$. Since $\tau_{\lambda,j}^2 = 1$ for any $j \in \mathcal{T}_\lambda$,

$$\sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} \tau_{\lambda,j}a_k\tau_{\lambda,k} = 1. \tag{34}$$

Therefore,

$$\begin{aligned}\tau_{\lambda,j}\boldsymbol{x}_j &= \sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} \tau_{\lambda,j}a_k\boldsymbol{x}_k \\ &= \sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} \tau_{\lambda,j}a_k\tau_{\lambda,k}^2\boldsymbol{x}_k \\ &\equiv \sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} c_k\tau_{\lambda,k}\boldsymbol{x}_k, \end{aligned} \tag{35}$$

where $c_k \equiv \tau_{\lambda,j}a_k\tau_{\lambda,k}$. By (34), we have $\sum_{k\in\mathcal{T}_\lambda\backslash\{j\}} c_k = 1$. This shows that $\tau_{\lambda,j}\boldsymbol{x}_j$, $j \in \mathcal{T}_\lambda$, is a weighted average of $\tau_{\lambda,k}\boldsymbol{x}_k$, $k \in \mathcal{T}_\lambda\backslash\{j\}$, which contradicts (**M1**). Thus, (**M1**) implies that $\boldsymbol{\beta}_\lambda$ is unique. □

Lemma A.2. *Suppose (***M1***) and (***M2***) hold. Then,*

$$\frac{1}{n}\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty = \mathcal{O}_p\left(\sqrt{\frac{\log(p)}{n}}\right).$$

Proof. There is an equivalence between the tail probability of a sub-Gaussian random variable and its moment generating function. For example, according to Lemma 5.5 in Vershynin (2012), since $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, for the constant $h > 0$ stated in (**M2**), there exists some $k > 0$ such that, $M_{\epsilon_i}(t) \equiv \mathbb{E}[\exp(t\epsilon_i)] \leq \exp(kt^2)$ for all $t \in \mathbb{R}$ if $\Pr[|\epsilon_i| \geq x] \leq$

$\exp(1 - hx^2)$ for any $x > 0$, where $M_{\epsilon_i}(t)$ is the moment generating function of ϵ_i. Thus, for any $j = 1, \ldots, p$, denoting $T_j \equiv \sum_{i=1}^n X_{ij}\epsilon_i/\sqrt{n} = (\boldsymbol{X}^\top \boldsymbol{\epsilon})_j/\sqrt{n}$, we have

$$
\begin{aligned}
M_{T_j}(t) = \mathbb{E}\left[\exp\left(t \sum_{i=1}^n \frac{1}{\sqrt{n}} X_{ij}\epsilon_i\right)\right] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{n}} X_{(i,j)}^2 t\epsilon_i\right)\right] \\
&\leq \prod_{i=1}^n \exp\left(\frac{X_{(i,j)}^2}{n} kt^2\right) \\
&= \exp\left(\frac{\|\boldsymbol{x}_j\|_2^2}{n} kt^2\right) = \exp\left(kt^2\right). \qquad (36)
\end{aligned}
$$

The last equality holds because columns of $\boldsymbol{X}$ are standardized such that $\|\boldsymbol{x}_j\|_2^2 = n$ for $j = 1, \ldots, p$ by (**M1**). Using Chebyshev's inequality, (36) shows that for any $j = 1, \ldots, p$, we have $\Pr[|T_j| \geq x] \leq \exp(1 - h'x^2)$ for some $h' > 0$. Applying Boole's inequality,

$$
\begin{aligned}
\Pr\left[\frac{1}{\sqrt{n}} \left\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\right\|_\infty \equiv \max_{j=1,\ldots,p} |T_j| > t\sqrt{\log(p)}\right] &= \Pr\left[\bigcup_{j=1,\ldots,p} \left\{|T_j| > t\sqrt{\log(p)}\right\}\right] \\
&\leq \sum_{j=1}^p \Pr\left[|T_j| > t\sqrt{\log(p)}\right] \\
&\leq p \exp\left(1 - h't^2 \log(p)\right) \\
&= \exp\left(\log(p)\left(1 - h't^2 \log(p)\right)\right) \\
&\leq \exp\left(1 - h't^2\right).
\end{aligned}
$$

Sicne $\exp(\log(p)(1 - h't^2 \log(p)))$ is a decreasing function of p with $h't^2 > 1$ and $p \geq e$, the last inequality holds with $h't^2 > 1$ and $p \geq 3$. Thus, for any $\xi > 0$, we can choose a large value of t, such that $\Pr[\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty/\sqrt{n} > t\sqrt{\log(p)}] < \xi$. This shows that $\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty/\sqrt{n} = \mathcal{O}_p(\sqrt{\log(p)})$. Dividing both sides by $\sqrt{n}$ completes the proof. □

Lemma A.3. *Suppose* (**E**) *holds. Then,* $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$*, where* $\mathcal{A}_\lambda \equiv \operatorname{supp}(\boldsymbol{\beta}_\lambda)$ *and* $\mathcal{S}^* \equiv \{j : |\beta_j^*| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$.

Proof. First, if $q^* \equiv |\mathcal{A}^*| = 0$, we trivially have $\mathcal{A}_\lambda \supseteq \mathcal{S}^* = \emptyset$.

If $q^* \geq 1$, by Corollary 2.1 in van de Geer and Bühlmann (2009), (**E**) guarantees that

$$\|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*\|_\infty \le \|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*\|_2 \le \frac{2\lambda\sqrt{2q^*}}{\phi^{*2}}. \tag{37}$$

But, for any $j = 1, \dots, p$ such that $j \in \mathcal{S}^*$, we have $|\beta_j^*| > 3\lambda\sqrt{q^*}/\phi^{*2}$. Thus, by (37), $|\beta_{\lambda,j}| > (3 - 2\sqrt{2})\lambda\sqrt{q^*}/\phi^{*2} > 0$ i.e., $j \in \mathcal{A}_\lambda$, or, $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$. □

Lemma A.4. *Suppose (**M2**), (**M3**) and (**E**) hold. Then, the estimator $\hat{\boldsymbol{\beta}}_\lambda$ defined in* (2) *and it population version, $\boldsymbol{\beta}_\lambda$, defined in* (7) *satisfy*

$$\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_2 = \mathcal{O}_p\left(\sqrt{\sqrt{q^* \log(p)}/n}\right).$$

Proof. To prove the result, we first show that $\|\hat{\boldsymbol{\beta}}_\lambda\|_0 \equiv |\hat{\mathcal{A}}_\lambda| = \mathcal{O}_p(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$ and $\|\boldsymbol{\beta}_\lambda\|_0 \equiv |\mathcal{A}_\lambda| = \mathcal{O}(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$; these imply that $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_0 = \mathcal{O}_p(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$.

To show that $\|\hat{\boldsymbol{\beta}}_\lambda\|_0 \equiv |\hat{\mathcal{A}}_\lambda| = \mathcal{O}_p(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$, we observe that based on Lemma 2 in Belloni and Chernozhukov (2013), and, specifically (3.3), (**M3**) and (**E**) imply that $\|\hat{\boldsymbol{\beta}}_\lambda\|_0 \equiv |\hat{\mathcal{A}}_\lambda| = \mathcal{O}_p(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$.

To show that $\|\boldsymbol{\beta}_\lambda\|_0 \equiv |\mathcal{A}_\lambda| = \mathcal{O}(q^*\|\hat{\boldsymbol{\Sigma}}\|_2^2)$, we observe that from the stationary condition of (7),

$$\lambda\boldsymbol{\tau}_\lambda = \hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right), \tag{38}$$

where $\hat{\boldsymbol{\Sigma}} \equiv \boldsymbol{X}^\top\boldsymbol{X}/n$. Thus,

$$\begin{aligned} \|\lambda\boldsymbol{\tau}_\lambda\|_2 = \left\|\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right)\right\|_2 &\le \left\|\hat{\boldsymbol{\Sigma}}\right\|_2 \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\|_2 \\ &= \mathcal{O}\left(\lambda\sqrt{q^*}\|\hat{\boldsymbol{\Sigma}}\|_2\right). \end{aligned} \tag{39}$$

The last equality is based on Corollary 2.1 in van de Geer and Bühlmann (2009) that (**M3**) implies that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\|_2 = \mathcal{O}(\lambda\sqrt{q^*})$.

On the other hand, because for any $j \in \mathcal{A}_\lambda$, $\tau_j^2 = 1$,

$$\|\lambda \boldsymbol{\tau}_\lambda\|_2 = \lambda \sqrt{\sum_{j \in \mathcal{A}_\lambda} \tau_j^2 + \sum_{k \notin \mathcal{A}_\lambda} \tau_k^2} \geq \lambda \sqrt{\sum_{j \in \mathcal{A}_\lambda} \tau_j^2} = \lambda \sqrt{|\mathcal{A}_\lambda|}. \tag{40}$$

Thus, combining with (39), we get that $|\mathcal{A}_\lambda| = \mathcal{O}(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2)$. Hence, $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_0 = \mathcal{O}_p(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2)$.

Now we proceed to show that $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_2 = \mathcal{O}_p(\sqrt{\sqrt{q^* \log(p)}/n})$. Theorem 2.1 in van de Geer (2017) shows that under (**M2**) and (**M3**),

$$\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right)\right\|_2 \leq \frac{1}{\lambda\sqrt{n}} \left\|\hat{\boldsymbol{\Sigma}}\right\|_2 \|\boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right)\|_2 \mathcal{O}_p(1) + \mathcal{O}_p(1). \tag{41}$$

Note that aalthough the above theorem assumes Gaussian random errors $\boldsymbol{\epsilon}$, it continues to hold for sub-Gaussian errors. This is because Lemma 15.5 in van de Geer (2017) can be proven with sub-Gaussian data as shown in Hsu et al. (2012).

With (**E**), Lemma 2.1 in van de Geer and Bühlmann (2009) shows that

$$\|\boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right)\|_2 = \mathcal{O}\left(\lambda\sqrt{nq^*}\right). \tag{42}$$

Given that $\|\hat{\boldsymbol{\Sigma}}\|_2 = \mathcal{O}((\log(p)/q^*)^{1/4})$ by (**E**), we have $\|\hat{\boldsymbol{\Sigma}}\|_2 \|\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda)\|_2/(\lambda\sqrt{n}) = \mathcal{O}((q^* \log(p))^{1/4})$ and

$$\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right)\right\|_2 = \mathcal{O}\left((q^* \log(p))^{1/4}\right). \tag{43}$$

Let $\mathcal{I} = \mathrm{supp}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda)$. Then, $|\mathcal{I}| = \|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_0 = \mathcal{O}_p(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2)$. Moreover, $\|\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{I}^c} - \boldsymbol{\beta}_{\lambda,\mathcal{I}^c}\|_1 = 0 \leq \|\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{I}} - \boldsymbol{\beta}_{\lambda,\mathcal{I}}\|_1$. Thus, by (**E**),

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right\|_2^2 = \mathcal{O}_p\left(\frac{1}{n}\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right)\right\|_2^2\right) = \mathcal{O}_p\left(\frac{\sqrt{q^* \log(p)}}{n}\right), \tag{44}$$

which completes the proof. $\square$

Lemma A.5. *Suppose* (**M1**), (**M3**), (**M4**), (**E**) *and* (**T**) *hold. For* $\mathcal{A}_\lambda \neq \emptyset$,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{b_{\lambda\min}} \to 0, \tag{45}$$

where $b_{\lambda\min} \equiv \min_{j\in\mathcal{A}_\lambda} |\beta_{\lambda,j}|$, *and* $\boldsymbol{\beta}_\lambda$ *is defined in* (7).

Proof. To prove the result, we show that $\sqrt{\log(p)/n}/|\beta_{\lambda,j}| \to 0$ for any $j \in \mathcal{A}_\lambda$. This is proved separately for entries in $\mathcal{S}^* \cap \mathcal{A}_\lambda$ and in $\mathcal{A}_\lambda \backslash \mathcal{S}^*$, where $\mathcal{S}^* \equiv \{j : |\beta^*_j| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$. By Lemma A.3, $\mathcal{S}^* \cap \mathcal{A}_\lambda = \mathcal{S}^*$.

We first show that for any $j \in \mathcal{S}^* \subseteq \mathcal{A}_\lambda$, $\sqrt{\log(p)/n}/|\beta_{\lambda,j}| \to 0$. For n sufficiently large, by (**E**), Lemma A.3 gives us

$$\|\boldsymbol{\beta}_{\lambda,\mathcal{S}^*} - \boldsymbol{\beta}^*_{\mathcal{S}^*}\|_\infty \leq \|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*\|_2 \leq \frac{2\lambda\sqrt{2q^*}}{\phi^{*2}}.$$

Thus, for any $j \in \mathcal{S}^*$, i.e., $|\beta^*_j| > 3\lambda\sqrt{q^*}/\phi^{*2}$, $|\beta_{\lambda,j}| > (3-2\sqrt{2})\lambda\sqrt{q^*}/\phi^{*2}$. Therefore,

$$0 < \sqrt{\frac{\log(p)}{n}}\frac{1}{|\beta_{\lambda,j}|} < \sqrt{\frac{\log(p)}{n}}\frac{1}{\lambda} \cdot \frac{\phi^{*2}}{(3-2\sqrt{2})\sqrt{q^*}} \to 0,$$

by (**M3**).

If $\mathcal{A}_\lambda = \mathcal{S}^*$, then our proof is complete. If $\mathcal{A}_\lambda \neq \mathcal{S}^*$, by Lemma A.3, (**E**) implies that $\mathcal{A}_\lambda \supset \mathcal{S}^*$. We now proceed to show that in the case that $\mathcal{A}_\lambda \supset \mathcal{S}^*$, $\sqrt{\log(p)/n}/|\beta_{\lambda,j}| \to 0$ for $j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*$. Consider the stationary condition of (7),

$$\begin{aligned} n\lambda\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda} &= \boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda\right) \\ &= \boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}_{\mathcal{A}_\lambda}\left(\boldsymbol{\beta}^*_{\mathcal{A}_\lambda} - \boldsymbol{\beta}_{\lambda,\mathcal{A}_\lambda}\right) + \boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}_{\mathcal{A}^c_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^c_\lambda}, \end{aligned} \tag{46}$$

where the second equality holds because $\boldsymbol{\beta}_{\lambda,\mathcal{A}^c_\lambda} = \mathbf{0}$. Rearranging terms,

$$\boldsymbol{\beta}_{\lambda,\mathcal{A}_\lambda} = \boldsymbol{\beta}^*_{\mathcal{A}_\lambda} + \left(\boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}_{\mathcal{A}^c_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^c_\lambda} - n\lambda\left(\boldsymbol{X}^\top_{\mathcal{A}_\lambda}\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}. \tag{47}$$

Thus, for any $j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*$,

$$
\begin{aligned}
|\boldsymbol{\beta}_{\lambda,j}| &= \left| \left[n\lambda \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda} \right]_j - \beta_j^* - \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right]_j \right| \\
&\geq \left| \left| \lambda \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right]^{-1} \boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda} \right|_j - \left| \beta_j^* + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right]_j \right| \right|. \qquad (48)
\end{aligned}
$$

We now bound the term $(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^*$. By definition,

$$
\phi_{\min}^2 \left(\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right) \equiv \min_{\boldsymbol{a} \in \mathbb{R}^p : \boldsymbol{a}_{\mathcal{A}_\lambda^c} = \boldsymbol{0}} \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{a}}{\|\boldsymbol{a}\|_2^2} = \min_{\boldsymbol{a} \in \mathbb{R}^p : \boldsymbol{a}_{\mathcal{A}_\lambda^c} = \boldsymbol{0}} \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{a}}{\|\boldsymbol{a}_{\mathcal{A}_\lambda}\|_2^2}.
$$

For any $\boldsymbol{a} \in \mathbb{R}^p$ such that $\boldsymbol{a}_{\mathcal{A}_\lambda^c} = \boldsymbol{0}$, we have $|\mathcal{A}_\lambda| = \mathcal{O}(q^* \|\hat{\boldsymbol{\Sigma}}\|_2^2)$ by Lemma A.4 and $\|\boldsymbol{a}_{\mathcal{A}_\lambda^c}\|_1 \leq \|\boldsymbol{a}_{\mathcal{A}_\lambda}\|_1$, and by (**E**),

$$
\liminf_{n \to \infty} \phi_{\min}^2 \left(\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right) = \liminf_{n \to \infty} \min_{\boldsymbol{a} \in \mathbb{R}^p : \boldsymbol{a}_{\mathcal{A}_\lambda^c} = \boldsymbol{0}} \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{a}}{\|\boldsymbol{a}_{\mathcal{A}_\lambda}\|_2^2} \geq \phi^{*2} > 0.
$$

Therefore,

$$
\begin{aligned}
\left\| \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \right\|_2 &= \sqrt{\phi_{\max}^2 \left(\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \right)} \\
&= \sqrt{\phi_{\max}^2 \left(\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \right)} = \sqrt{\frac{1}{n} \phi_{\max}^2 \left(\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right]^{-1} \right)} \\
&= \sqrt{\frac{1}{n} \phi_{\min}^{-2} \left(\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right)} \leq \sqrt{\frac{1}{n\phi^{*2}}} = \mathcal{O} \left(\frac{1}{\sqrt{n}} \right). \qquad (49)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\left\| \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right\|_\infty &\leq \left\| \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right\|_2 \\
&\leq \left\| \left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda} \right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \right\|_2 \left\| \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right\|_2 \\
&= \mathcal{O} \left(\frac{1}{\sqrt{n}} \left\| \boldsymbol{X}_{\mathcal{A}_\lambda^c} \boldsymbol{\beta}_{\mathcal{A}_\lambda^c}^* \right\|_2 \right). \qquad (50)
\end{aligned}
$$

If $\mathcal{A}_\lambda \supseteq \mathcal{A}^*$, we have $\|\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\|_2 = 0$. Otherwise, if $\mathcal{S}^* \subset \mathcal{A}_\lambda \not\supseteq \mathcal{A}^*$, based on (**M4**),

$$\left\|\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right\|_2 = \left\|\boldsymbol{X}_{\mathcal{A}^*\backslash\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^*\backslash\mathcal{A}_\lambda}\right\|_2 = \mathcal{O}\left(\sqrt{\log(p)}\right), \tag{51}$$

and

$$\left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right). \tag{52}$$

Thus, for any $j \in \mathcal{A}_\lambda\backslash\mathcal{S}^*$, based on (**M4**) that $\|\boldsymbol{X}_{\mathcal{A}_\lambda\backslash\mathcal{S}^*}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\backslash\mathcal{S}^*}\|_2 = \mathcal{O}(\sqrt{\log(p)/n})$,

$$\begin{aligned}\left|\boldsymbol{\beta}^*_j + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right]_j\right| &\le \left\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\backslash\mathcal{S}^*}\right\|_\infty + \left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right\|_\infty \\ &= \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right). \end{aligned} \tag{53}$$

Now, by (48), for any $j \in \mathcal{A}_\lambda\backslash\mathcal{S}^*$,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{|\beta_{\lambda,j}|} \le \frac{\sqrt{\log(p)/n}}{\left|\left|n\lambda\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}\right|_j - \left|\boldsymbol{\beta}^*_j + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right]_j\right|\right|}.$$

But, by (53) and (**T**),

$$\frac{\left|\boldsymbol{\beta}^*_j + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right]_j\right|}{\min_{j\in\mathcal{A}_\lambda\backslash\mathcal{S}^*}\left|\lambda\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}\right]^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}\right|_j} \to 0$$

Therefore,

$$\begin{aligned}&\frac{\sqrt{\log(p)/n}}{\left|\min_{j\in\mathcal{A}_\lambda\backslash\mathcal{S}^*}\left|\lambda\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}\right]^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}\right|_j - \left|\boldsymbol{\beta}^*_j + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right]_j\right|\right|} \\ &\to \frac{\sqrt{\log(p)/n}}{\min_{j\in\mathcal{A}_\lambda\backslash\mathcal{S}^*}\left|\lambda\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}\right]^{-1}\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda}\right|_j} \to 0.\end{aligned}$$

Thus,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{|\beta_{\lambda,j}|} \to 0,$$

as desired. $\square$

Proof of Proposition 2.4. According to the stationary conditions of (7) and (2), respectively,

$$\lambda n \boldsymbol{\tau}_\lambda = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda) - \boldsymbol{X}^\top \boldsymbol{\epsilon}, \tag{54}$$

$$\lambda n \hat{\boldsymbol{\tau}}_\lambda = \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right). \tag{55}$$

This implies that

$$\hat{\boldsymbol{\tau}}_\lambda - \boldsymbol{\tau}_\lambda = \frac{1}{n\lambda}\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{\beta}_\lambda - \hat{\boldsymbol{\beta}}_\lambda\right) + \frac{1}{n\lambda}\boldsymbol{X}^\top\boldsymbol{\epsilon}. \tag{56}$$

We now bound both terms on the right hand side of (56). By Lemma A.2,

$$\frac{\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty}{n\lambda} = \mathcal{O}_p\left(\frac{1}{\lambda}\sqrt{\frac{\log(p)}{n}}\right).$$

In addition, Lemma A.4 shows that

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right\|_2 = \mathcal{O}_p\left(\sqrt{\frac{\sqrt{q^*\log(p)}}{n}}\right).$$

Therefore,

$$\begin{aligned}
\frac{1}{n\lambda}\left\|\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{\beta}_\lambda - \hat{\boldsymbol{\beta}}_\lambda\right)\right\|_\infty &\leq \frac{1}{n}\left\|\boldsymbol{X}^\top\boldsymbol{X}\right\|_2 \frac{\left\|\boldsymbol{\beta}_\lambda - \hat{\boldsymbol{\beta}}_\lambda\right\|_2}{\lambda} \\
&= \left\|\hat{\boldsymbol{\Sigma}}\right\|_2 \frac{\left\|\boldsymbol{\beta}_\lambda - \hat{\boldsymbol{\beta}}_\lambda\right\|_2}{\lambda} \\
&= \mathcal{O}_p\left(\frac{1}{\lambda}\sqrt{\frac{\log(p)}{n}}\right),
\end{aligned}$$

where the last equality is based on the ℓ_2 norm of $\hat{\boldsymbol{\Sigma}}$ in (**E**).

Therefore, it follows from (56) that $\|\boldsymbol{\tau}_\lambda - \hat{\boldsymbol{\tau}}_\lambda\|_\infty = \mathcal{O}_p(\sqrt{\log(p)/n}/\lambda)$. By (**T**), $\limsup_{n\to\infty} \|\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda^c}\|_\infty \leq 1-\delta$ for some δ such that $\sqrt{\log(p)/n}/(\lambda\delta) \to 0$. Hence, $\lim_{n\to\infty} \Pr[\|\hat{\boldsymbol{\tau}}_{\lambda,\mathcal{A}_\lambda^c}\|_\infty < 1] = 1$, and

$$\lim_{n\to\infty} \Pr\left[\mathcal{A}_\lambda \supseteq \hat{\mathcal{A}}_\lambda\right] = 1. \tag{57}$$

To prove the other direction, if $\mathcal{A}_\lambda = \emptyset$, then, $\mathcal{A}_\lambda \subseteq \hat{\mathcal{A}}_\lambda$, Otherwise, if $\mathcal{A}_\lambda \neq \emptyset$, by Lemma A.5,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{b_{\lambda\min}} \to 0. \tag{58}$$

Based on Lemma A.4, $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_2 = \mathcal{O}_p(\sqrt{\sqrt{\log(p)q^*}/n})$. Thus, for any $\xi > 0$, there exists a constant $C > 0$, not depending on n, such that for n sufficiently large,

$$\Pr\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\right\|_\infty > C\sqrt{\frac{\sqrt{\log(p)q^*}}{n}}\right] < \xi. \tag{59}$$

Based on (58), for n sufficiently large, $b_{\lambda\min} > C\sqrt{\log(p)/n} > C\sqrt{\sqrt{\log(p)q^*}/n}$, where $b_{\lambda\min} \equiv \min_{j\in\mathcal{A}_\lambda} |\beta_{\lambda,j}|$. Thus, combining (58) and (59), for n sufficiently large, whenever $|\beta_{\lambda,j}| > 0$, $|\beta_{\lambda,j}| > C\sqrt{\log(p)/n}$ and hence $\Pr[|\hat{\beta}_{\lambda,j}| > 0] > 1-\xi$. Therefore

$$\lim_{n\to\infty} \Pr\left[\mathcal{A}_\lambda \subseteq \hat{\mathcal{A}}_\lambda\right] = 1, \tag{60}$$

which completes the proof. $\square$

B Proof of Theorem 2.5

Proof of Theorem 2.5. By Proposition 2.4, $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda] \to 1$. Therefore, with probability tending to one,

$$\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} \equiv \left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y}\right]_j = \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{y}\right]_j. \tag{61}$$

Thus,

$$
\begin{aligned}
\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} &= \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \left(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}\right)\right]_j \\
&= \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{\epsilon}\right]_j + \left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}\boldsymbol{\beta}^*\right]_j. \qquad (62)
\end{aligned}
$$

We proceed to prove the asymptotic distribution of $[(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{\epsilon}]_j$. Dividing it by its standard deviation, $\sigma_{\boldsymbol{\epsilon}}\sqrt{[(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1}]_{(j,j)}}$, where $\sigma_{\boldsymbol{\epsilon}}$ is the error standard deviation, we get

$$
\frac{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1} \boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{\epsilon}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\right]_{(j,j)}}} = \frac{\mathbf{r}^w \boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}} \left\|\mathbf{r}^w\right\|_2}, \qquad (63)
$$

where $\boldsymbol{r}^w \equiv \boldsymbol{e}^j(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \in \mathbb{R}^n$, and $\boldsymbol{e}^j$ is the row vector of length $|\mathcal{A}_\lambda|$ with the entry that corresponds to β_j^* equal to one, and zero otherwise. In order to use the Lindeberg-Feller Central Limit Theorem to prove the asymptotic normality of (63), we need to show that the Lindeberg's condition holds, i.e.,

$$
\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\left[\frac{(r_i^w \epsilon_i)^2}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\mathbf{r}^w\right\|_2^2}\mathbf{1}\left\{\frac{|r_i^w \epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^w\right\|_2} > \eta\right\}\right] = 0, \quad \forall \eta > 0.
$$

Given that $|r_i^w| \le \|\mathbf{r}^w\|_\infty$, and that the ϵ_i's are identically distributed,

$$
\begin{aligned}
0 \le \sum_{i=1}^n \mathbb{E}\left[\frac{(r_i^w \epsilon_i)^2}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\mathbf{r}^w\right\|_2^2}\mathbf{1}\left\{\frac{|r_i^w \epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^w\right\|_2} > \eta\right\}\right] &\le \sum_{i=1}^n \mathbb{E}\left[\frac{(r_i^w \epsilon_i)^2}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\mathbf{r}^w\right\|_2^2}1\left\{\frac{|\epsilon_i|\left\|\mathbf{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^w\right\|_2} > \eta\right\}\right] \\
&= \sum_{i=1}^n \frac{r_i^{w2}}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\mathbf{r}^w\right\|_2^2}\mathbb{E}\left[\epsilon_i^2 1\left\{\frac{|\epsilon_i|\left\|\mathbf{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^w\right\|_2} > \eta\right\}\right] \\
&= \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{|\epsilon_1|\left\|\mathbf{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^w\right\|_2} > \eta\right\}\right].
\end{aligned}
$$

Since $\|\mathbf{r}^w\|_\infty/\|\mathbf{r}^w\|_2 \to 0$ by Condition (**W**), $\epsilon_1^2 1\{|\epsilon_1|\|\mathbf{r}^w\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}^w\|_2) > \eta\} \to_p 0$. Thus, because $\epsilon_1^2 \ge \epsilon_1^2 1\{|\epsilon_1|\|\mathbf{r}^w\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}^w\|_2) > \eta\}$ with probability one and $\mathbb{E}[\epsilon_1^2] = \sigma_\epsilon^2 < \infty$, we use ϵ_1^2 as the dominant random variable, and apply the Dominated Convergence Theorem,

$$\lim_{n\to\infty}\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{|\epsilon_1|\,\|\mathbf{r}^w\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\,\|\mathbf{r}^w\|_2}>\eta\right\}\right]=\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\lim_{n\to\infty}\epsilon_1^2 1\left\{\frac{|\epsilon_1|\,\|\mathbf{r}^w\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\,\|\mathbf{r}^w\|_2}>\eta\right\}\right]=0,$$

which gives the Lindeberg's condition.

Thus,

$$\frac{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{\epsilon}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\right]_{(j,j)}}}\to_d\mathcal{N}(0,1). \tag{64}$$

Using, again, the fact that by Proposition 2.4, $\lim_{n\to\infty}\Pr\left[\mathcal{A}_\lambda=\hat{\mathcal{A}}_\lambda\right]=1$, we can write

$$\begin{aligned}\frac{\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)}-\beta_j^{(\hat{\mathcal{A}}_\lambda)}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}&\equiv\frac{\tilde{\beta}_j^{(\hat{\mathcal{A}}_\lambda)}-\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}\boldsymbol{\beta}^*\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}\\&\to_p\frac{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{\epsilon}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\right]_{(j,j)}}}\to_d\mathcal{N}\left(0,1\right).\end{aligned}$$

□

C Proof of Theorem 2.6

Proof. Based on Proposition 2.4 that $\Pr[\hat{\mathcal{A}}_\lambda=\mathcal{A}_\lambda]\to 1$, we have

$$\frac{1}{n-\hat{q}_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}\right\|_2^2\to_p\frac{1}{n-q_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2, \tag{65}$$

where $\hat{q}_\lambda \equiv |\hat{\mathcal{A}}_\lambda|$, $q_\lambda \equiv |\mathcal{A}_\lambda|$. Denoting $\boldsymbol{P}^{(\mathcal{A}_\lambda)} \equiv \boldsymbol{X}_{\mathcal{A}_\lambda}(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda})^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top$,

$$
\begin{aligned}
&\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2 \\
&\equiv \left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top \boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}\boldsymbol{y}\right\|_2^2 \\
&= \boldsymbol{y}^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)^2\boldsymbol{y} \\
&= \boldsymbol{y}^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)\boldsymbol{y} \\
&= \left(\boldsymbol{X}_{\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda} + \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c} + \boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda} + \boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c} + \boldsymbol{\epsilon}\right) \\
&= \left(\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c} + \boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c} + \boldsymbol{\epsilon}\right) \\
&= \left(\boldsymbol{X}_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus\mathcal{A}_\lambda} + \boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus\mathcal{A}_\lambda} + \boldsymbol{\epsilon}\right) \qquad (66) \\
&= \left(\boldsymbol{X}_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)} + \boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)} + \boldsymbol{\epsilon}\right) \qquad (67)
\end{aligned}
$$

where (66) is based on the fact that $\boldsymbol{\beta}^*_{-\mathcal{A}^*} \equiv \boldsymbol{0}$ and (67) holds because based on Lemma A.3, $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$. To simplify notations, denote $\boldsymbol{\theta} \equiv \boldsymbol{X}_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}$ and $\boldsymbol{Q} \equiv \boldsymbol{I} - \boldsymbol{P}^{(\mathcal{A}_\lambda)}$. Expanding (67),

$$
(\boldsymbol{\theta} + \boldsymbol{\epsilon})^\top \boldsymbol{Q}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) = \boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta} + 2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}.
$$

Because $\boldsymbol{Q}$ is an idempotent matrix, $\boldsymbol{Q}$ is positive semidefinite, whose eigenvalues are all zeros and ones. Thus,

$$
0 \leq \boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta} \leq \phi^2_{\max}[\boldsymbol{Q}]\,\|\boldsymbol{\theta}\|_2^2 = \mathcal{O}(\log(p)),
$$

where the last equality is based on (**M4**). Since $\log(p)/(n - q_\lambda) \to 0$,

$$
0 \leq \frac{1}{n - q_\lambda}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta} = \mathcal{O}\left(\frac{\log(p)}{n - q_\lambda}\right) = o(1),
$$

which means that

$$
\frac{1}{n - q_\lambda}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta} = o(1). \qquad (68)
$$

Therefore,

$$\frac{1}{n-q_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2=\frac{2}{n-q_\lambda}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}+\frac{1}{n-q_\lambda}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}+\mathcal{O}(1). \tag{69}$$

We now derive the expected value and variance of $\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\|_2^2/(n-q_\lambda)$. First, because $\mathbb{E}[\boldsymbol{\epsilon}]=\mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}]=\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I}$, according to the formula of the expectation of quadratic forms,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n-q_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right]&=\mathbb{E}\left[\frac{2}{n-q_\lambda}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]+\mathbb{E}\left[\frac{1}{n-q_\lambda}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]+\mathcal{O}(1)\\&=\mathbb{E}\left[\frac{1}{n-q_\lambda}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]+\mathcal{O}(1)\\&=\frac{1}{n-q_\lambda}\sigma_{\boldsymbol{\epsilon}}^2\mathrm{tr}\left[\boldsymbol{Q}\right]+\mathcal{O}(1).\end{aligned} \tag{70}$$

Because $\boldsymbol{Q}$ is an idempotent matrix, we have $\mathrm{tr}[\boldsymbol{Q}]=n-q_\lambda$. Thus

$$\mathbb{E}\left[\frac{1}{n-q_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right]\to\sigma_{\boldsymbol{\epsilon}}^2. \tag{71}$$

We now calculate the variance of $\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\|_2^2/(n-q_\lambda)$. Since

$$\begin{aligned}&\mathrm{Var}\left[\frac{1}{n-q_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right]\\&=\frac{1}{(n-q_\lambda)^2}\left(\mathbb{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^4\right]-\mathbb{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right]^2\right)\\&\to\frac{1}{(n-q_\lambda)^2}\mathbb{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^4\right]-\sigma_{\boldsymbol{\epsilon}}^4,\end{aligned} \tag{72}$$

where the second term is derived in (71), we now derive the expected value of $\|\boldsymbol{y}-$

$\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\|_2^4/(n-q_\lambda)^2$. Based on (69),

$$
\begin{aligned}
\frac{1}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^4\right] = \mathcal{O}(1)\cdot\frac{1}{n-q_\lambda}&\mathbb{E}\left[2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\
&+\frac{4}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] \\
&+\frac{2}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\
&+\frac{1}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]+\mathcal{O}(1). \qquad (73)
\end{aligned}
$$

We now consider each term in the above formulation. First,

$$
\frac{1}{n-q_\lambda}\mathbb{E}\left[2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]=\frac{1}{n-q_\lambda}\mathbb{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]=\sigma_{\boldsymbol{\epsilon}}^2, \qquad (74)
$$

where the last equality is based on (70) and (71). For the second term, based on (68),

$$
\begin{aligned}
\frac{4}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] &= \frac{4}{\left(n-q_\lambda\right)^2}\boldsymbol{\theta}^\top\boldsymbol{Q}\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right]\boldsymbol{Q}\boldsymbol{\theta} \\
&= \frac{4\sigma_{\boldsymbol{\epsilon}}^2}{\left(n-q_\lambda\right)^2}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{Q}\boldsymbol{\theta}=\frac{4\sigma_{\boldsymbol{\epsilon}}^2}{\left(n-q_\lambda\right)^2}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta}=\mathcal{O}\left(\frac{1}{n-q_\lambda}\right).
\end{aligned} \qquad (75)
$$

For the third term,

$$
\begin{aligned}
&\frac{2}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\
&=\frac{2}{\left(n-q_\lambda\right)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\mathbb{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]+\mathrm{Cor}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon},\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\sqrt{\frac{2\mathrm{Var}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{\left(n-q_\lambda\right)^2}\frac{2\mathrm{Var}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{\left(n-q_\lambda\right)^2}} \\
&=\mathrm{Cor}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon},\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\sqrt{\frac{2\mathrm{Var}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{\left(n-q_\lambda\right)^2}\frac{2\mathrm{Var}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{\left(n-q_\lambda\right)^2}},
\end{aligned}
$$

where, based on (68),

$$
\begin{aligned}
\frac{2}{(n-q_\lambda)^2}\mathrm{Var}\left[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \frac{2}{(n-q_\lambda)^2}\left(\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\theta}\right] - \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right]^2\right) \\
&= \frac{2\sigma_{\boldsymbol{\epsilon}}^2}{(n-q_\lambda)^2}\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\theta} = \mathcal{O}\left(\frac{1}{n-q_\lambda}\right), \qquad (76)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{2}{(n-q_\lambda)^2}\mathrm{Var}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \frac{2}{(n-q_\lambda)^2}\left(\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right]^2\right) \\
&= \frac{2}{(n-q_\lambda)^2}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - 2\sigma_{\boldsymbol{\epsilon}}^4. \qquad (77)
\end{aligned}
$$

The last equality is based on (70) and (71). Since $-1 \leq \mathrm{Cor}[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}] \leq 1$, based on (76) and (77), we have

$$
\begin{aligned}
-\sqrt{\mathcal{O}\left(\frac{1}{(n-q_\lambda)^3}\right)\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathcal{O}\left(\frac{1}{n-q_\lambda}\right)} &\leq \frac{2}{(n-q_\lambda)^2}\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] \\
&\leq \sqrt{\mathcal{O}\left(\frac{1}{(n-q_\lambda)^3}\right)\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathcal{O}\left(\frac{1}{n-q_\lambda}\right)}.
\end{aligned}
$$

Therefore, collecting (74), (75), (76) and (77), we have

$$
\begin{aligned}
0 &\leq \frac{1}{(n-q_\lambda)^2}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^4\right] \\
&\leq \mathcal{O}\left(\frac{1}{(n-q_\lambda)^{3/2}}\right)\sqrt{\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right]} + \frac{1}{(n-q_\lambda)^2}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] + \mathcal{O}(1). \qquad (78)
\end{aligned}
$$

We now calculate $\mathbb{E}[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}]$.

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathbb{E}\left[\sum_{i,j,l,k}^{n} \epsilon_i\epsilon_j\epsilon_l\epsilon_k Q_{(i,j)}Q_{(l,k)}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \epsilon_i^4 Q_{(i,i)}^2\right] + \mathbb{E}\left[\sum_{i=l\neq j=k}^{n} \epsilon_i^2\epsilon_j^2 Q_{(i,j)}Q_{(l,k)}\right] \\ &+ \mathbb{E}\left[\sum_{i=k\neq j=l}^{n} \epsilon_i^2\epsilon_j^2 Q_{(i,j)}Q_{(l,k)}\right] + \mathbb{E}\left[\sum_{i=j\neq l=k}^{n} \epsilon_i^2\epsilon_l^2 Q_{(i,j)}Q_{(l,k)}\right]. \end{aligned} \tag{79}$$

Because $\boldsymbol{\epsilon}$ is independent and identically distributed with $\mathbb{E}[\epsilon_1^2] = \sigma_{\boldsymbol{\epsilon}}^2$,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i=1}^{n} Q_{(i,i)}^2 + \mathbb{E}\left[\epsilon_1^2\right]^2 \sum_{i=l\neq j=k}^{n} Q_{(i,j)}Q_{(l,k)} \\ &+ \mathbb{E}\left[\epsilon_1^2\right]^2 \sum_{i=k\neq j=l}^{n} Q_{(i,j)}Q_{(l,k)} + \mathbb{E}\left[\epsilon_1^2\right]^2 \sum_{i=j\neq l=k}^{n} Q_{(i,j)}Q_{(l,k)} \\ &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i=1}^{n} Q_{(i,i)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,i)}Q_{(j,j)} \\ &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i=1}^{n} Q_{(i,i)}^2 + 2\sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,i)}Q_{(j,j)}. \end{aligned} \tag{80}$$

Because $(\sum_{i=1}^{n} Q_{(i,i)})^2 = \sum_{i\neq j}^{n} Q_{(i,i)}Q_{(j,j)} + \sum_{i=1}^{n} Q_{(i,i)}^2$,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i=1}^{n} Q_{(i,i)}^2 + 2\sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \left(\sum_{i=1}^{n} Q_{(i,i)}\right)^2 - \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i=1}^{n} Q_{(i,i)}^2 \\ &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i=1}^{n} Q_{(i,i)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \left(\sum_{i=1}^{n} Q_{(i,i)}\right)^2 \\ &\leq \mathbb{E}\left[\epsilon_1^4\right]\left(\sum_{i=1}^{n} Q_{(i,i)}^2 + \sum_{i\neq j}^{n} Q_{(i,j)}^2\right) + \sigma_{\boldsymbol{\epsilon}}^4 \left(\sum_{i=1}^{n} Q_{(i,i)}\right)^2 \\ &= \mathbb{E}\left[\epsilon_1^4\right]\sum_{i,j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \mathrm{tr}\left[\boldsymbol{Q}\right]^2, \end{aligned} \tag{81}$$

where the inequality is based on Jensen's inequality that $\mathbb{E}[\epsilon_1^2]^2 \leq \mathbb{E}[\epsilon_1^4]$. Because

$\sum_{i,j}^{n} Q^2_{(i,j)} = \mathrm{tr}[\boldsymbol{Q}^2] = \mathrm{tr}[\boldsymbol{Q}] = n - q_\lambda$, we have

$$\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] = (n - q_\lambda)\,\mathbb{E}\left[\epsilon_1^4\right] + (n - q_\lambda)^2\,\sigma^4_{\boldsymbol{\epsilon}}. \tag{82}$$

Since ϵ_1 has sub-Gaussian tails, we have $\mathbb{E}[\epsilon_1^4] = \mathcal{O}(1)$ (see, e.g., Lemma 5.5 in Vershynin, 2012). Thus, based on (78),

$$0 \le \frac{1}{(n - q_\lambda)^2}\mathbb{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^4\right] \le \sigma^4_{\boldsymbol{\epsilon}} + \scriptstyle\mathcal{O}(1), \tag{83}$$

and hence, based on (72),

$$\mathrm{Var}\left[\frac{1}{n - q_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right] = \scriptstyle\mathcal{O}(1). \tag{84}$$

Finally, applying Chebyshev's inequality, we obtain

$$\begin{aligned}\frac{1}{n - \hat{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\tilde{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}\right\|_2^2 &\to_p \frac{1}{n - q_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2 \\ &\to_p \mathbb{E}\left[\frac{1}{n - q_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_\lambda}\tilde{\boldsymbol{\beta}}^{(\mathcal{A}_\lambda)}\right\|_2^2\right] = \sigma^2_{\boldsymbol{\epsilon}}.\end{aligned} \tag{85}$$

□

D Proof of Theorem 4.1

Proof of Theorem 4.1. By Proposition 2.4, $\Pr\left[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda\right] \to 1$. Therefore, we also have $\Pr\left[\left(\hat{\mathcal{A}}_\lambda \backslash \{j\}\right) = \left(\mathcal{A}_\lambda \backslash \{j\}\right)\right] \to 1$, and with probability tending to one,

$$\begin{aligned}S^j \equiv \boldsymbol{x}_j^\top\left(\mathbf{I}_n - \mathbf{P}^{(\hat{\mathcal{A}}_\lambda\backslash\{j\})}\right)\boldsymbol{y} &= \boldsymbol{x}_j^\top\left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\backslash\{j\})}\right)\boldsymbol{y} \\ &= \boldsymbol{x}_j^\top\left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\backslash\{j\})}\right)\left(\boldsymbol{X}_{\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda} + \boldsymbol{X}_{\mathcal{A}^c_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^c_\lambda} + \boldsymbol{\epsilon}\right),\end{aligned} \tag{86}$$

where $\mathbf{P}^{(\mathcal{A}_\lambda\backslash\{j\})} \equiv \boldsymbol{X}_{\mathcal{A}_\lambda\backslash\{j\}}(\boldsymbol{X}^\top_{\mathcal{A}_\lambda\backslash\{j\}}\boldsymbol{X}_{\mathcal{A}_\lambda\backslash\{j\}})^{-1}\boldsymbol{X}^\top_{\mathcal{A}_\lambda\backslash\{j\}}$.

Thus, under the null hypothesis $H^*_{0,j} : \beta^*_j = 0$, (86) is equal to

$$\begin{aligned} &\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})}\right) \left(\boldsymbol{X}_{\mathcal{A}_\lambda\setminus\{j\}}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\setminus\{j\}} + \boldsymbol{X}_{\mathcal{A}^c_\lambda\setminus\{j\}}\boldsymbol{\beta}^*_{\mathcal{A}^c_\lambda\setminus\{j\}} + \boldsymbol{\epsilon}\right) \\ &= \boldsymbol{x}_j^\top \left(\boldsymbol{I}_n - \boldsymbol{P}^{(\mathcal{A}_\lambda\setminus\{j\})}\right)\boldsymbol{\epsilon} + \boldsymbol{x}_j^\top \left(\boldsymbol{I}_n - \boldsymbol{P}^{(\mathcal{A}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_{\mathcal{A}^c_\lambda\setminus\{j\}}\boldsymbol{\beta}^*_{\mathcal{A}^c_\lambda\setminus\{j\}}. \end{aligned} \tag{87}$$

The equality holds because $(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})})\boldsymbol{X}_{\mathcal{A}_\lambda\setminus\{j\}}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\setminus\{j\}} = \mathbf{0}$.

We first show the asymptotic distribution of $\boldsymbol{x}_j^\top(\boldsymbol{I}_n - \boldsymbol{P}^{(\mathcal{A}_\lambda\setminus\{j\})})\boldsymbol{\epsilon}$. Dividing it by its standard deviation $\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{x}_j^\top(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})})\boldsymbol{x}_j}$, where $\sigma_{\boldsymbol{\epsilon}}$ is the error standard deviation,

$$\frac{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})}\right)\boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})}\right)\boldsymbol{x}_j}} = \frac{\mathbf{r}^{s\top}\boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^s\right\|_2}, \tag{88}$$

where $\mathbf{r}^{s\top} \equiv \boldsymbol{x}_j^\top(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda\setminus\{j\})})$. Now, we use the Lindeberg-Feller Central Limit Theorem to prove the asymptotic normality of (88). Similar to the proof of Theorem 2.5, we need to prove that the Lindeberg's condition holds, i.e.,

$$\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\left[\frac{\left(r_i^s\epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\mathbf{r}^s\right\|_2^2}1\left\{\frac{\left|r_i^s\epsilon_i\right|}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^s\right\|_2} > \eta\right\}\right] = 0, \quad \forall\eta > 0.$$

Given that $|r_i^s| \leq \|\mathbf{r}^s\|_\infty$, and that the ϵ_i's are identically distributed,

$$0 \leq \sum_{i=1}^n \mathbb{E}\left[\frac{\left(r_i^s\epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\mathbf{r}^s\right\|_2^2}1\left\{\frac{\left|r_i^s\epsilon_i\right|}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^s\right\|_2} > \eta\right\}\right] \leq \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{\left|\epsilon_1\right|\left\|\mathbf{r}^s\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^s\right\|_2} > \eta\right\}\right].$$

Since $\|\mathbf{r}^s\|_\infty/\|\mathbf{r}^s\|_2 \to 0$ by Condition (**S**), $\epsilon_1^2 1\left\{|\epsilon_1|\|\mathbf{r}^s\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}^s\|_2) > \eta\right\} \to_p 0$. Thus, because $\epsilon_1^2 \geq \epsilon_1^2 1\left\{|\epsilon_1|\|\mathbf{r}^s\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}^s\|_2) > \eta\right\}$ with probability one and $\mathbb{E}[\epsilon_1^2] = \sigma_\epsilon^2 < \infty$, we use ϵ_1^2 as the dominant random variable, and apply the Dominated Convergence Theorem,

$$\lim_{n\to\infty}\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{\left|\epsilon_1\right|\left\|\mathbf{r}^s\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\mathbf{r}^s\right\|_2} > \eta\right\}\right] = 0,$$

which in turn gives the Lindeberg's condition.

Thus,

$$\frac{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda \setminus \{j\})}\right) \boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda \setminus \{j\})}\right) \boldsymbol{x}_j}} \to_d \mathcal{N}(0,1), \tag{89}$$

We now prove the asymptotic unbiasedness of the naïve score test on $\boldsymbol{\beta}^*$. Dividing the second term in (87) by $\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{x}_j^\top (\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda \setminus \{j\})})\boldsymbol{x}_j}$, we get

$$\begin{aligned}
\left| \frac{\boldsymbol{x}_j^\top \left(\boldsymbol{I}_n - \boldsymbol{P}^{(\mathcal{A}_\lambda \setminus \{j\})}\right) \boldsymbol{X}_{\mathcal{A}_\lambda^c \setminus \{j\}} \boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c \setminus \{j\}}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\mathcal{A}_\lambda \setminus \{j\})}\right) \boldsymbol{x}_j}} \right| &= \left| \frac{\mathbf{r}^s \boldsymbol{X}_{\mathcal{A}_\lambda^c \setminus \{j\}} \boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c \setminus \{j\}}}{\sigma_{\boldsymbol{\epsilon}} \left\|\mathbf{r}^s\right\|_2} \right| \\
&\leq \frac{\left\|\mathbf{r}^s\right\|_2}{\left\|\mathbf{r}^s\right\|_2} \frac{\left\|\boldsymbol{X}_{\mathcal{A}_\lambda^c \setminus \{j\}} \boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c \setminus \{j\}}\right\|_2}{\sigma_{\boldsymbol{\epsilon}}} \\
&= \frac{\left\|\boldsymbol{X}_{\mathcal{A}_\lambda^c \setminus \{j\}} \boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c \setminus \{j\}}\right\|_2}{\sigma_{\boldsymbol{\epsilon}}}.
\end{aligned} \tag{90}$$

Based on Lemma A.3, $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$, and by (**M4***),

$$\left\|\boldsymbol{X}_{\mathcal{A}_\lambda^c \setminus \{j\}} \boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c \setminus \{j\}}\right\|_2 = \left\|\boldsymbol{X}_{(\mathcal{A}^* \setminus \mathcal{A}_\lambda) \setminus \{j\}} \boldsymbol{\beta}^*_{(\mathcal{A}^* \setminus \mathcal{A}_\lambda) \setminus \{j\}}\right\|_2 = \left\|\boldsymbol{X}_{\mathcal{A}^* \setminus \mathcal{A}_\lambda} \boldsymbol{\beta}^*_{\mathcal{A}^* \setminus \mathcal{A}_\lambda}\right\|_2 = \scriptstyle{\mathcal{O}}\displaystyle(1),$$

where the second equality holds under $H_{0,j} : \beta_j^* = 0$.

Using, again, the fact that by Proposition 2.4, $\lim_{n\to\infty} \Pr\left[\mathcal{A}_\lambda = \hat{\mathcal{A}}_\lambda\right] = 1$, we get

$$\frac{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}\right) \boldsymbol{y}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{x}_j^\top \left(\mathbf{I}_n - \mathbf{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}\right) \boldsymbol{x}_j}} \to_d \mathcal{N}(0,1). \tag{91}$$

□

E Additional Comments on Condition (M4)

In this section, we further comment on Condition (**M4**), which we re-state here for convenience:

(**M4**) Recall that $\mathcal{A}^* \equiv \text{supp}(\boldsymbol{\beta}^*)$. Let $\mathcal{S}^* \equiv \{j : |\beta_j^*| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$, where $q^* \equiv$

$|\mathcal{A}^*| \equiv |\text{supp}(\boldsymbol{\beta}^*)|$, and ϕ^* is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda \backslash \mathcal{S}^*}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right),$$

and

$$\left\|\boldsymbol{X}_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)} \boldsymbol{\beta}^*_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)}\right\|_2 = \mathcal{O}\left(\sqrt{\log(p)}\right),$$

where $\mathcal{A}_\lambda \equiv \text{supp}(\boldsymbol{\beta}_\lambda)$, with $\boldsymbol{\beta}_\lambda$ defined in (7).

A conceptual illustration of (**M4**) is presented in Figure 2, in which lines show the required magnitude of β^*_j. Figure 2 shows that (**M4**) allows for the presence of both strong signal and weak signal variables, with a gap in signal strength. Specifically, let "$\gtrsim$" and "$\lesssim$" denote asymptotic inequalities.

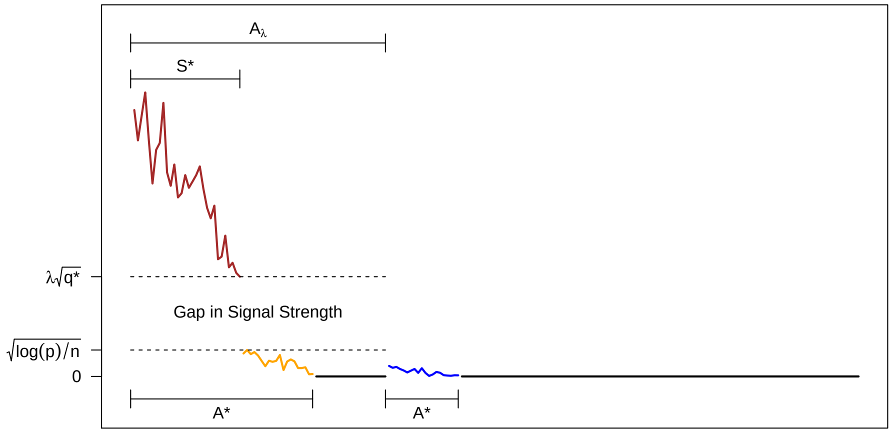
- Brown line represents the signal strength of strong signal variables, $|\beta^*_j| \gtrsim \lambda\sqrt{q^*}$, $j \in \mathcal{S}^*$.
- Orange line represents the signal strength of weak signal variables that are selected by the noiseless lasso, $|\beta^*_j| \lesssim \sqrt{\log(p)/n}$, $j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*$.
- Blue line represents the signal strength of weak signal variables that are not selected by the noiseless lasso, $\|\boldsymbol{X}_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)} \boldsymbol{\beta}^*_{\mathcal{A}^* \backslash (\mathcal{A}_\lambda \cup \mathcal{S}^*)}\|_2 \lesssim \sqrt{\log(p)}$.

The condition in (**M4**) for weak signal variables differs based on whether the variable is in $\mathcal{A}_\lambda$, which may seem unintuitive. We show that the results in the paper can be obtained with (**M4**) replaced with (**M4a**) and (**M4b**). In the new conditions, and in particular, in (**M4a**), the condition on signal strength no longer depends on $\mathcal{A}_\lambda$.

(**M4a**) Recall that $\mathcal{A}^* \equiv \text{supp}(\boldsymbol{\beta}^*)$. Let $\mathcal{S}^* \equiv \{j : |\beta^*_j| > 3\lambda\sqrt{q^*}/\phi^{*2}\}$, where $q^* \equiv |\mathcal{A}^*| \equiv |\text{supp}(\boldsymbol{\beta}^*)|$, and ϕ^* is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}^*_{\mathcal{A}^* \backslash \mathcal{S}^*}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right).$$

Figure 2: Conceptual illustration of the required magnitude of $|\beta_j^*|$ in (**M4**). Brown, orange and blue lines represent the signal strength of strong signal variables, weak signal variables in $\mathcal{A}_\lambda$ and weak signal variables in $\mathcal{A}_\lambda^c$, respectively. Black line represents noise variables.



(**M4b**) $\boldsymbol{X}$, $\mathcal{A}^*$ and $\mathcal{A}_\lambda$ satisfy

$$\left\| \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right]^{-1} \hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}^*\setminus\mathcal{A}_\lambda)} \right\|_\infty = \mathcal{O}\left(1\right). \tag{92}$$

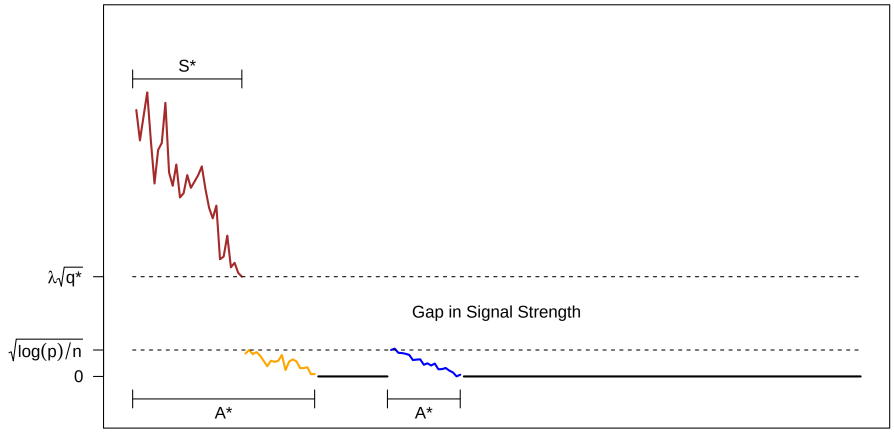
Condition (**M4a**) unifies the condition on $\|\boldsymbol{\beta}^*_{\mathcal{A}_\lambda\setminus\mathcal{S}^*}\|_\infty$ and on $\|\boldsymbol{X}_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\|_2$, and states that the orange and blue lines in Figure 2 can be of the same magnitude; see Figure 3 for a conceptual illustration of (**M4a**). In other words, compared to (**M4**), (**M4a**) no longer depends on $\mathcal{A}_\lambda$.

Condition (**M4b**) is similar to the mutual incoherence condition (see, e.g., Fuchs, 2005; Tropp, 2006; Wainwright, 2009), which requires that

$$\limsup_{n\to\infty} \left\| \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^*)} \right]^{-1} \hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^{*c})} \right\|_\infty < 1.$$

However, (**M4b**) is considerably milder: first, requiring a value to be bounded above is much milder than requiring it to be smaller than one. Second, when the model is sparse, i.e., $q^*/p \to 0$, and when n is large, $|\mathcal{A}^{*c}| \approx p$. In this case, $[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}]^{-1}\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}^*\setminus\mathcal{A}_\lambda)}$ is of dimension no larger than $n \times q^*$, which is substantially smaller than the dimension of $[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^*)}]^{-1}\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^{*c})}$, which is approximately $q^* \times p$. Thus, $\|[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}]^{-1}\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}^*\setminus\mathcal{A}_\lambda)}\|_\infty$

Figure 3: Conceptual illustration of the required magnitude of $|\beta_j^*|$ in (**M4a**). Brown line represents the signal strength of strong signal variables, and orange and blue lines represent the signal strength of weak signal variables. Black line represents noise variables.



is expected to be much smaller than $\|[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^*)}]^{-1}\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^{*c})}\|_\infty$. In other words, the left-hand side of (**M4b**) is smaller than the left-hand side of the mutual incoherence condition, and the right-hand side of (**M4b**) is larger than the right-hand side of the mutual incoherence condition.

To verify that we can replace (**M4**) with (**M4a**) and (**M4b**), recall that the condition $\|\boldsymbol{X}_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus(\mathcal{A}_\lambda\cup\mathcal{S}^*)}\|_2 = \mathcal{O}(\sqrt{\log(p)})$ in (**M4**) is required only in the proof of Lemma A.5 to show (52),

$$\left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right).$$

With (**M4a**) and (**M4b**),

$$\begin{aligned}
\left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda^c}\boldsymbol{\beta}^*_{\mathcal{A}_\lambda^c}\right\|_\infty &= \left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\right\|_\infty \\
&\le \left\|\left(\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}_\lambda}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\right\|_\infty \left\|\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus\mathcal{A}_\lambda}\right\|_\infty \\
&\le \left\|\left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)}\right]^{-1}\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}^*\setminus\mathcal{A}_\lambda)}\right\|_\infty \left\|\boldsymbol{\beta}^*_{\mathcal{A}^*\setminus\mathcal{S}^*}\right\|_\infty \\
&= \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right),
\end{aligned}$$

where the first equality holds because $\boldsymbol{\beta}^*_{\mathcal{A}^{*c}} \equiv \mathbf{0}$ and the second inequality holds because by Lemma A.3, $\mathcal{S}^* \subseteq \mathcal{A}_\lambda$. Hence, (**M4a**) and (**M4b**) also imply (52).

F Proof of Remarks 2.2 and 2.3

Proof of Remark 2.2. When $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}$, based on the the stationary condition of the noiseless lasso,

$$\lambda \boldsymbol{\tau}_\lambda = \hat{\boldsymbol{\Sigma}} \left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda \right) = \boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda. \tag{93}$$

Rearranging terms,

$$\boldsymbol{\beta}_\lambda = \boldsymbol{\beta}^* - \lambda \boldsymbol{\tau}_\lambda. \tag{94}$$

Thus, $\boldsymbol{\beta}_\lambda = \mathrm{sign}(\boldsymbol{\beta}^*)(|\boldsymbol{\beta}^*| - \lambda)_+$. Since for $j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*$, $\beta^*_j \preceq \sqrt{\log(p)/n} \prec \lambda$, for n sufficiently large, we have $\boldsymbol{\beta}_{\lambda, \mathcal{A}_\lambda \backslash \mathcal{S}^*} = \mathbf{0}$, or $\mathcal{A}_\lambda \subseteq \mathcal{S}^*$. Also based on Lemma A.3 that $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$, we have $\mathcal{A}_\lambda = \mathcal{S}^*$.

In addition, since $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}$, we have

$$\min_{j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*} \left| \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda, \mathcal{A}_\lambda)} \right]^{-1} \boldsymbol{\tau}_{\lambda, \mathcal{A}_\lambda} \right|_j = \min_{j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*} |\boldsymbol{\tau}_{\lambda, j}| = 1.$$

Thus based on (**M3**),

$$\frac{\sqrt{\log(p)/n}/\lambda}{\min_{j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*} \left| \left[\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda, \mathcal{A}_\lambda)} \right]^{-1} \boldsymbol{\tau}_{\lambda, \mathcal{A}_\lambda} \right|_j} \to 0,$$

□

Proof of Remark 2.3. Based on the stationary condition of the noiseless lasso,

$$\lambda \boldsymbol{\tau}_\lambda = \hat{\boldsymbol{\Sigma}} \left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda \right). \tag{95}$$

Since $\hat{\mathbf{\Sigma}}_{(\mathcal{A}^*,-\mathcal{A}^*)} = \mathbf{0}$, we get

$$\lambda \boldsymbol{\tau}_{\lambda,\mathcal{A}^*} = \hat{\mathbf{\Sigma}}_{(\mathcal{A}^*,\mathcal{A}^*)} \left(\boldsymbol{\beta}^*_{\mathcal{A}^*} - \boldsymbol{\beta}_{\lambda,\mathcal{A}^*}\right), \tag{96}$$

$$\lambda \boldsymbol{\tau}_{\lambda,-\mathcal{A}^*} = \hat{\mathbf{\Sigma}}_{(-\mathcal{A}^*,-\mathcal{A}^*)} \left(\boldsymbol{\beta}^*_{-\mathcal{A}^*} - \boldsymbol{\beta}_{\lambda,-\mathcal{A}^*}\right) = -\hat{\mathbf{\Sigma}}_{(-\mathcal{A}^*,-\mathcal{A}^*)}\boldsymbol{\beta}_{\lambda,-\mathcal{A}^*}. \tag{97}$$

Observe that (96) is the stationary condition of the noiseless lasso applied on the data $(\boldsymbol{X}_{\mathcal{A}^*}, \boldsymbol{y})$ with tuning parameter λ. Thus,

$$\boldsymbol{\beta}_{\lambda,\mathcal{A}^*} = \underset{\boldsymbol{b}\in\mathbb{R}^{q^*}}{\arg\min} \left\{\mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}^*}\boldsymbol{b}\|_2^2\right] + \lambda \|\boldsymbol{b}\|\right\}. \tag{98}$$

On the other hand, $\boldsymbol{\beta}_{\lambda,-\mathcal{A}^*} = \mathbf{0}$ is the solution to (97). Hence, when $\hat{\mathbf{\Sigma}}_{(\mathcal{A}^*,-\mathcal{A}^*)} = \mathbf{0}$, $\mathcal{A}_\lambda \subseteq \mathcal{A}^*$. Also based on Lemma A.3, $\mathcal{A}_\lambda \supseteq \mathcal{S}^*$. Thus, $\mathcal{S}^* \subseteq \mathcal{A}_\lambda \subseteq \mathcal{A}^*$.

□

G Empirical Examination of Condition (T)

In this section, we examine the stringency of Condition **(T)**. To do so, we adopted a similar approach as in Section 3.3 of Zhao and Yu (2006). Specifically, for $p \in \{8, 16, 32, 64, 128, 256\}$, we first generated the population covariance matrix of $\boldsymbol{X}$, $\mathbf{\Sigma} \sim Wishart(p, \mathcal{I}_p)$. $\mathbf{\Sigma}$ was then standardized, before obtaining $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $n = 1000$. We then set

$$\boldsymbol{\beta}^* = (\underbrace{1, ..., 1}_{q^*}, \underbrace{0, ..., 0}_{p-q^*})^\top,$$

where $q^*/p \in \{1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8\}$. Thus, $\mathcal{A}^* = \mathcal{S}^* = \{1, \ldots, q^*\}$. To get $\boldsymbol{\beta}_\lambda$, we set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^*$ and used the `glmnet` R package on the noiseless data $(\boldsymbol{X}, \boldsymbol{y})$.

Condition **(T)** has two parts. To examine part one, for each p and q^*, we estimated the probability that

$$\left\|\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda^c}\right\|_\infty < 0.999. \tag{99}$$

To examine part two, for each p and q^*, we estimated the probability that

$$\mathcal{A}_\lambda \backslash \mathcal{S}^* = \emptyset \quad \text{or} \quad \frac{\sqrt{\log(p)/n}/\lambda}{\min_{j \in \mathcal{A}_\lambda \backslash \mathcal{S}^*} \left| \left[\hat{\mathbf{\Sigma}}_{(\mathcal{A}_\lambda, \mathcal{A}_\lambda)} \right]^{-1} \boldsymbol{\tau}_{\lambda, \mathcal{A}_\lambda} \right|_j} < 1. \tag{100}$$

As a comparison, we also estimated the probability that the irrepresentable condition (Zhao and Yu, 2006) holds, i.e.,

$$\left\| \hat{\mathbf{\Sigma}}_{(\mathcal{A}^{*c}, \mathcal{A}^*)} \left(\hat{\mathbf{\Sigma}}_{(\mathcal{A}^*, \mathcal{A}^*)} \right)^{-1} \operatorname{sign}\left(\boldsymbol{\beta}^*_{\mathcal{A}^*} \right) \right\|_\infty < 1. \tag{101}$$

The probabilities were estimated empirically from 1000 repetitions.

Tables 6, 7 and 8 show the estimated probability that part 1 of Condition **(T)**, part 2 of Condition **(T)** and the irrepresentable condition holds. They imply that Condition **(T)** is more likely to be satisfied than the irrepresentable condition when p and q^* are large.

Acknowledgements

We thank the Editor, Associate Editor, and two anonymous reviewers for their incredibly insightful comments, which lead to substantial improvements of the manuscript. We thank the authors of Javanmard and Montanari (2014a); Ning and Liu (2016) for providing code for their proposals. We are grateful to Joshua Loftus, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani for helpful responses to our inquiries.

Bibliography

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2).

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Buena, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169 – 194.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer Series in Statistics. Springer Berlin Heidelberg, Berlin and Heidelberg, Germany.

Candés, E. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145 – 2177.

Cox, D. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statistical Science*, 30(4):533–558.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Dossal, C. (2012). A necessary and sufficient condition for exact sparse recovery by $\ell 1$ minimization. *Comptes Rendus Mathematique*, 350(1-2):117–120.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6:290–297.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B*, 74(1):37–65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fuchs, J.-J. (2005). Recovery of exact sparse representations in the presence of noise. *IEEE Transaction on Information Theory*, 51(10):3601–3608.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, New York, NY, USA.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag New York, New York, NY, USA.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horvath, S. and Dong, J. (2008). Geometric interpretation of gene co-expression network analysis. *PLOS Computational Biology*, 4(8):e1000117.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., M., F. R., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., and Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46):17402–17407.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:52.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1187–1195. Curran Associates, Inc.

Javanmard, A. and Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(Oct):2869–2909.

Javanmard, A. and Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transaction on Information Theory*, 60(10):6522 – 6554.

Jin, T., Wang, Y., Li, G., Du, S., Yang, H., Geng, T., Hou, P., and Gong, Y. (2015).

Analysis of difference of association between polymorphisms in the XRCC5, RPA3 and RTEL1 genes and glioma, astrocytoma and glioblastoma. *American Journal of Cancer Research*, 5(7):2294–2300.

Kabaila, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory*, 14(4):463–482.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. *International Statistical Review*, 77(3):405–414.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer Science+Business Media, New York, NY, USA.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Leeb, H. and Pötscher, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22(1):69–97.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Leeb, H., Pötscher, B. M., and Ewald, K. (2015). On various confidence intervals post-model-selection. *Statistical Science*, 30(2):216–227.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Ning, Y. and Liu, H. (2016). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, to appear.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statistica Sinica*, 26(1):35–67.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Tropp, J. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transaction on Information Theory*, 52(3):1030–1051.

van de Geer, S. (2017). Some exercises with the lasso and its compatibility constant. *arXiv preprint arXiv:1701.03326*.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journals of Statistics*, 3:1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vardabasso, C., Hasson, D., Ratnakumar, K., Chung, C.-Y., Duarte, L. F., and Bernstein, E. (2014). Histone variants: emerging players in cancer biology. *Cellular and Molecular Life Sciences*, 71(3):379–404.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing: Theory and Applications*, chapter 5. Cambridge University Press, Cambridge, UK.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (lasso). *IEEE Transaction on Information Theory*, 55(5):2183 – 2202.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201.

Weisberg, S. (2013). *Applied Lienar Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, USA.

You, H., Lin, H., and Zhang, Z. (2015). CKS2 in human cancers: Clinical roles and current perspectives (review). *Molecular and Clinical Oncology*, 3(3):459–463.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(Article 17).

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with $l1$ regularization. *The Annals of Statistics*, 37(5A):2109 – 2144.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563.

Table 4: Average power and type-I error rate for the hypotheses $H^*_{0,j} : \beta^*_j = 0$ for $j = 1, \ldots, p$, as defined in (18)–(20), under the scale-free graph setting with $p = 500$. Results are shown for various values of ρ, n, SNR. Methods for comparison include LDPE, SSLasso, dScore, and the naïve score test with tuning parameter $\lambda_{1\mathrm{SE}}$ and $\lambda_{\sup}$.

	ρ	0.2								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\mathrm{Power}_{\mathrm{strong}}$	LDPE $\lambda_{\min}$	0.400	0.773	0.910	0.627	0.973	1.000	0.923	1.000	1.000
	SSLasso $\lambda_{\min}$	0.410	0.770	0.950	0.650	0.970	1.000	0.910	1.000	1.000
	dScore $\lambda_{\min}$	0.330	0.643	0.857	0.547	0.957	1.000	0.887	1.000	1.000
	nScore $\lambda_{1\mathrm{SE}}$	0.427	0.763	0.893	0.677	0.977	1.000	0.957	1.000	1.000
	nScore $\lambda_{\sup}$	0.403	0.847	0.960	0.727	0.990	0.997	0.940	1.000	1.000
$\mathrm{Power}_{\mathrm{weak}}$	LDPE $\lambda_{\min}$	0.064	0.083	0.056	0.054	0.059	0.079	0.070	0.079	0.113
	SSLasso $\lambda_{\min}$	0.081	0.087	0.060	0.066	0.061	0.086	0.069	0.086	0.113
	dScore $\lambda_{\min}$	0.044	0.056	0.036	0.039	0.039	0.060	0.046	0.056	0.093
	nScore $\lambda_{1\mathrm{SE}}$	0.080	0.077	0.059	0.060	0.061	0.061	0.083	0.076	0.101
	nScore $\lambda_{\sup}$	0.061	0.091	0.109	0.070	0.109	0.107	0.097	0.103	0.114
T1 Error	LDPE $\lambda_{\min}$	0.051	0.052	0.051	0.049	0.051	0.047	0.050	0.051	0.049
	SSLasso $\lambda_{\min}$	0.056	0.056	0.056	0.054	0.055	0.053	0.053	0.054	0.054
	dScore $\lambda_{\min}$	0.035	0.040	0.040	0.033	0.036	0.034	0.035	0.037	0.034
	nScore $\lambda_{1\mathrm{SE}}$	0.061	0.057	0.048	0.056	0.055	0.040	0.060	0.046	0.046
	nScore $\lambda_{\sup}$	0.069	0.082	0.095	0.064	0.083	0.079	0.065	0.068	0.050
	ρ	0.6								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\mathrm{Power}_{\mathrm{strong}}$	LDPE $\lambda_{\min}$	0.330	0.783	0.947	0.627	0.980	1.000	0.887	1.000	1.000
	SSLasso $\lambda_{\min}$	0.347	0.790	0.957	0.623	0.987	1.000	0.867	1.000	1.000
	dScore $\lambda_{\min}$	0.270	0.673	0.883	0.533	0.960	0.993	0.863	1.000	1.000
	nScore $\lambda_{1\mathrm{SE}}$	0.357	0.767	0.887	0.677	0.980	0.997	0.937	1.000	1.000
	nScore $\lambda_{\sup}$	0.430	0.790	0.933	0.707	0.977	1.000	0.923	1.000	1.000
$\mathrm{Power}_{\mathrm{weak}}$	LDPE $\lambda_{\min}$	0.031	0.046	0.063	0.064	0.074	0.076	0.054	0.077	0.119
	SSLasso $\lambda_{\min}$	0.047	0.063	0.076	0.063	0.090	0.099	0.053	0.076	0.121
	dScore $\lambda_{\min}$	0.021	0.037	0.047	0.036	0.060	0.044	0.034	0.050	0.083
	nScore $\lambda_{1\mathrm{SE}}$	0.039	0.060	0.050	0.076	0.074	0.066	0.070	0.067	0.104
	nScore $\lambda_{\sup}$	0.071	0.089	0.136	0.081	0.121	0.104	0.114	0.113	0.123
T1 Error	LDPE $\lambda_{\min}$	0.050	0.051	0.051	0.050	0.049	0.051	0.051	0.050	0.047
	SSLasso $\lambda_{\min}$	0.056	0.056	0.056	0.054	0.055	0.053	0.053	0.054	0.054
	dScore $\lambda_{\min}$	0.033	0.036	0.034	0.031	0.031	0.035	0.036	0.035	0.033
	nScore $\lambda_{1\mathrm{SE}}$	0.056	0.060	0.045	0.061	0.051	0.040	0.058	0.048	0.047
	nScore $\lambda_{\sup}$	0.065	0.080	0.093	0.064	0.084	0.088	0.070	0.071	0.054

Table 5: Average power and type-I error rate for the hypotheses $H^*_{0,j} : \beta^*_j = 0$ for $j = 1, \dots, p$, as defined in (18)–(20), under the stochastic block model setting with $p = 500$. Details are as in Table 4.

	ρ	0.2								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{min}	0.370	0.793	0.937	0.687	0.990	1.000	0.914	1.000	1.000
	SSLasso λ_{min}	0.393	0.803	0.933	0.687	0.990	1.000	0.892	1.000	1.000
	dScore λ_{min}	0.333	0.783	0.917	0.693	0.993	1.000	0.905	1.000	1.000
	nScore λ_{1SE}	0.400	0.797	0.903	0.713	0.997	1.000	0.910	1.000	1.000
	nScore λ_{sup}	0.473	0.857	0.953	0.697	0.993	0.993	0.943	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{min}	0.041	0.044	0.051	0.057	0.050	0.071	0.050	0.093	0.071
	SSLasso λ_{min}	0.054	0.056	0.074	0.071	0.056	0.089	0.071	0.101	0.101
	dScore λ_{min}	0.037	0.044	0.057	0.060	0.046	0.077	0.056	0.101	0.094
	nScore λ_{1SE}	0.047	0.059	0.060	0.059	0.047	0.059	0.062	0.106	0.105
	nScore λ_{sup}	0.059	0.071	0.107	0.043	0.083	0.070	0.069	0.094	0.106
T1ER	LDPE λ_{min}	0.051	0.049	0.048	0.050	0.050	0.050	0.051	0.050	0.049
	SSLasso λ_{min}	0.057	0.056	0.058	0.054	0.054	0.054	0.054	0.053	0.054
	dScore λ_{min}	0.043	0.040	0.041	0.041	0.044	0.042	0.042	0.042	0.041
	nScore λ_{1SE}	0.062	0.058	0.048	0.056	0.052	0.040	0.054	0.047	0.046
	nScore λ_{sup}	0.064	0.074	0.090	0.059	0.076	0.076	0.060	0.060	0.49
	ρ	0.6								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{min}	0.327	0.827	0.960	0.700	0.983	0.997	0.968	1.000	1.000
	SSLasso λ_{min}	0.350	0.853	0.957	0.687	0.990	0.997	0.945	0.996	1.000
	dScore λ_{min}	0.297	0.787	0.937	0.697	0.987	0.993	0.968	0.996	1.000
	nScore λ_{1SE}	0.350	0.800	0.927	0.717	0.980	1.000	0.968	1.000	1.000
	nScore λ_{sup}	0.420	0.870	0.957	0.720	0.987	1.000	0.947	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{min}	0.043	0.049	0.046	0.041	0.077	0.063	0.053	0.066	0.083
	SSLasso λ_{min}	0.061	0.054	0.070	0.053	0.086	0.083	0.067	0.099	0.105
	dScore λ_{min}	0.044	0.047	0.046	0.040	0.081	0.069	0.063	0.077	0.098
	nScore λ_{1SE}	0.059	0.056	0.044	0.054	0.087	0.074	0.067	0.086	0.103
	nScore λ_{sup}	0.054	0.073	0.093	0.063	0.093	0.094	0.061	0.094	0.096
T1 Error	LDPE λ_{min}	0.049	0.049	0.049	0.049	0.050	0.049	0.049	0.047	0.048
	SSLasso λ_{min}	0.057	0.056	0.056	0.053	0.054	0.054	0.053	0.053	0.053
	dScore λ_{min}	0.033	0.039	0.036	0.031	0.033	0.033	0.032	0.030	0.031
	nScore λ_{1SE}	0.057	0.051	0.047	0.056	0.049	0.039	0.055	0.045	0.046
	nScore λ_{sup}	0.063	0.079	0.089	0.062	0.077	0.075	0.060	0.062	0.048

Table 6: Estimated probability that part 1 of Condition **(T)** holds.

	$p = 8$	$p = 16$	$p = 32$	$p = 64$	$p = 128$	$p = 256$
$q^* = p/8$	100%	99.8%	99.5%	99.0%	98.4%	95.9%
$q^* = 2p/8$	99.9%	99.6%	99.0%	98.4%	96.7%	92.1%
$q^* = 3p/8$	99.9%	99.3%	99.3%	96.5%	94.5%	89.8%
$q^* = 4p/8$	99.6%	99.7%	98.8%	96.7%	94.1%	88.2%
$q^* = 5p/8$	99.3%	99.7%	98.0%	96.4%	92.5%	89.4%
$q^* = 6p/8$	99.7%	99.4%	99.2%	97.6%	94.6%	90.1%
$q^* = 7p/8$	99.7%	99.1%	98.7%	97.2%	94.6%	88.2%

Table 7: Estimated probability that part 2 of Condition **(T)** holds.

	$p=8$	$p=16$	$p=32$	$p=64$	$p=128$	$p=256$
$q^*=p/8$	100%	97.9%	96.9%	93.0%	87.5%	81.3%
$q^*=2p/8$	95.6%	91.4%	86.1%	76.4%	65.3%	45.2%
$q^*=3p/8$	96.3%	90.4%	82.8%	74.0%	58.2%	31.1%
$q^*=4p/8$	96.6%	91.4%	84.4%	71.3%	55.6%	30.0%
$q^*=5p/8$	97.1%	93.9%	87.3%	75.2%	60.4%	34.1%
$q^*=6p/8$	97.0%	94.3%	89.0%	79.7%	69.5%	45.9%
$q^*=7p/8$	99.2%	96.9%	95.9%	90.1%	81.0%	68.1%

Table 8: Estimated probability that the irrepresentable condition holds.

	$p=8$	$p=16$	$p=32$	$p=64$	$p=128$	$p=256$
$q^*=p/8$	100%	93.0%	83.6%	64.2%	30.3%	0%
$q^*=2p/8$	76.2%	48.6%	19.9%	3.5%	0%	0%
$q^*=3p/8$	49.6%	21.0%	3.3%	0%	0%	0%
$q^*=4p/8$	33.1%	10.7%	0.7%	0%	0%	0%
$q^*=5p/8$	27.9%	6.3%	0.3%	0%	0%	0%
$q^*=6p/8$	27.5%	6.9%	0.3%	0%	0%	0%
$q^*=7p/8$	36.7%	14.4%	0.8%	0%	0%	0%