Electronic Journal of Statistics
Vol. 13 (2019) 5306–5333
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1649

Variable selection via adaptive false negative control in linear regression

X. Jessie Jeng* **and Xiongzhi Chen**

Department of Statistics, North Carolina State University, Raleigh, NC 27695
e-mail: xjjeng@ncsu.edu

Department of Mathematics and Statistics
Washington State University, Pullman, WA 99164
e-mail: xiongzhi.chen@wsu.edu

Abstract: Variable selection methods have been developed in linear regression to provide sparse solutions. Recent studies have focused on further interpretations on the sparse solutions in terms of false positive control. In this paper, we consider false negative control for variable selection with the goal to efficiently select a high proportion of relevant predictors. Different from existing studies in power analysis and sure screening, we propose to directly estimate the false negative proportion (FNP) of a decision rule and select the smallest subset of predictors that has the estimated FNP less than a user-specified control level. The proposed method is adaptive to the user-specified control level on FNP by selecting less candidates if a higher level is implemented. On the other hand, when data has stronger effect size or larger sample size, the proposed method controls FNP more efficiently with less false positives. New analytic techniques are developed to cope with the major challenge of FNP control when relevant predictors cannot be consistently separated from irrelevant ones. Our numerical results are in line with the theoretical findings.

MSC 2010 subject classifications: Primary 62J07; secondary 62F03.
Keywords and phrases: debiased Lasso, FNC-Reg, post-selection inference, variable screening.

Received August 2019.

Contents

*The research of X. J. Jeng was supported in part by NSF Grant DMS-1811360.

1. Introduction

We consider a sparse linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the vector of n observations of response, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of unknown coefficients, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \sigma^2 \mathbf{I}\right)$ is the vector of random errors. We assume $\sigma^2 = O(1)$. Let $I_1 = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the set of indices for non-zero coefficients with cardinality $s = |I_1|$ and $I_0 = \{1 \leq j \leq p : \beta_j = 0\}$ with cardinality $p_0 = |I_0|$.

Variable selection methods often provide sparse solutions for the estimation of $\boldsymbol{\beta}$. The non-zero elements of an estimate correspond to variables selected as candidates for relevant predictors. A great amount of literature with many fruitful ideas has contributed to the development of sparse solutions to accommodate the underlying features of the data. We refer to [5] and the references therein for a nice introduction.

Given a selection result, a false positive (FP) occurs when an irrelevant predictor is selected, and a false negative (FN) occurs when a relevant predictor is not selected. It is natural to interpret a selection result in terms of false positive or false negative control, and exciting progress has emerged for false positive control, e.g. [3], [4], [9], [17], [24], [32], [38]. However, the study for efficient false negative control remains relatively underdeveloped.

False negative control is important in many real applications and sometimes a more serious concern than false positive control. For example, in pre-surgical brain mapping with functional MRI, the primary goal is to reduce false negatives where genuine functional areas are not identified. This is because neurosurgical patients are more likely to experience significant harm from mistakenly deeming

a region to be functionally uninvolved and subsequently resecting critical tissue than from incorrectly assigning function to an uninvolved region [25, 26, 31]. Another example where false negative control is of main concern is in the exploratory stage of high-dimensional data analysis, where pre-screening is often conducted to reduce data dimension while keeping a high proportion of true signal variables for follow-up studies.

The problem of false negative control is conceptually related but methodologically very different from Sure Screening in, e.g., [14, 15]. Sure Screening aims to reduce the data dimension by removing only irrelevant predictors. For instance, the Sure Independence Screening procedure in [14] ranks variables by estimated marginal regression coefficients and selects the top d variables where d is fixed at $n-1$ or $n/\log n$. It has been proved that under certain conditions, the screening procedure has eliminated only irrelevant predictors with high probability. The false negative control problem considered here focuses on selecting a high proportion of relevant predictors without including many unnecessary irrelevant predictors. It may be regarded as a more refined screening procedure with a data-adaptive selection rule instead of a fixed d.

We use false negative proportion (FNP) as a measure for false negative control. For a given selection rule, FNP is defined as the ratio of the number of false negatives to the total number of relevant predictors. FNP takes values in $[0, 1]$ and is equivalent to $1 - \mathsf{Sensitivity}$ in binary classification framework. Our work starts with consistently estimating FNP for a given selection rule. To achieve this, we develop novel analyses on the tail behavior of the empirical processes associated with FNP. Based on the estimation of FNP, we develop a new variable selection procedure to control FNP at a user-specified level. If users can tolerate more false negatives, they may implement lower control levels on FNP in the procedure and select less candidates for relevant predictors. On the other hand, if the effect of relevant predictors gets stronger or sample size increases, the procedure controls FNP more efficiently with less false positives.

An important component of the proposed FNP control method is an estimator for the number of relevant predictors. We provide a consistent estimator for dependent test statistics, for which we adopt the recently developed debiased Lasso estimates [20, 34, 37].

Although FNP, by definition, is equivalent to the power in (single) hypothesis testing, our proposed study on FNP control is very different from the existing power analysis in hypothesis testing. In the latter, a decision rule is built upon Type I error control and followed by power calculation with assumptions on the effect size. For such methods to control FNP in addition to controlling family-wise Type I error when multiple hypotheses are considered, the effect sizes of relevant variables need to be larger enough to ensure essentially perfect separation of relevant and irrelevant variables. The proposed method, on the other hand, directly bound the estimated FNP at a user-specified level, which allows a more effective control on FNP. Our condition on effect size for FNP control is shown to be weaker than the existing beta-min conditions that are required for perfect separation of relevant and irrelevant variables.

The rest of the paper is organized as follows. Section 2 presents FNP estimation in two steps: (1) constructing test statistics for regression coefficients and (2) approximating FNP based on the test statistics. Section 3 develops a variable selection method to control FNP at a user-specified level and a computational algorithm to implement the method. Section 4 presents the finite-sample performance of the proposed method in simulation. Conclusion and further discussion are provided in Section 5. Proofs for the main theoretical results are presented in Section 6. Extra technical details are provided in Appendix.

2. False negative proportion estimation

Recall that for a selection rule, FNP is the ratio of the number of false negatives to the total number of relevant predictors. In this section, we rank the predictors based on the debiased Lasso estimates and approximate FNP at a given cut-off point on the list of ranked predictors.

2.1. Test statistics based on debiased Lasso estimates

Recall model (1.1). The well-known Lasso estimator is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/n + 2\lambda\|\boldsymbol{\beta}\|_1), \tag{2.1}$$

where λ is a tuning parameter [33]. Recently, the debiased Lasso estimator has been developed to mitigate the bias of Lasso estimator [34, 37]. The debiased Lasso estimator is defined as

$$\hat{\mathbf{b}} = \left(\hat{b}_1, \ldots, \hat{b}_p\right)^T = \hat{\boldsymbol{\beta}} + n^{-1}\hat{\boldsymbol{\Theta}}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.2}$$

where $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^{p\times p}$ is an estimate for the precision matrix of the predictors and can be obtained via nodewise regression on $\mathbf{X}$ as in [28]. Let $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{X}^T\mathbf{X}$. It has been shown that

$$\sqrt{n}(\hat{\mathbf{b}} - \boldsymbol{\beta}) = n^{-1/2}\hat{\boldsymbol{\Theta}}\mathbf{X}^T\boldsymbol{\varepsilon} - \boldsymbol{\delta} = \mathbf{w} - \boldsymbol{\delta}, \tag{2.3}$$

where

$$\mathbf{w}|\mathbf{X} \sim \mathcal{N}_p(0, \sigma^2\hat{\boldsymbol{\Omega}}), \qquad \hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}^T,$$

and

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T = \sqrt{n}(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{2.4}$$

Under certain conditions, $\|\boldsymbol{\delta}\|_\infty = o_p(1)$, which implies the asymptotic normality of $\hat{\mathbf{b}}$ [6, 20, 21, 34, 37]. We present the set of conditions from [21] as A1)–A3) in Appendix A.1.

In this paper, we obtain test statistics for $\boldsymbol{\beta}$ using the standardized debiased Lasso estimator as

$$z_j = \sqrt{n}\hat{b}_j\sigma^{-1}\hat{\boldsymbol{\Omega}}_{jj}^{-1/2} \quad \text{for} \qquad 1 \le j \le p \tag{2.5}$$

where $\hat{\mathbf{\Omega}}_{jj}$ denotes the (j,j) entry of $\hat{\mathbf{\Omega}}$. Therefore, for each $1 \le j \le p$,

$$z_j = \mu_j + w_j' - \delta_j',$$

where, given $\mathbf{X}$,

$$w_j' = \frac{w_j}{\sigma\sqrt{\hat{\mathbf{\Omega}}_{jj}}} \sim \mathcal{N}(0,1), \quad \delta_j' = \frac{\delta_j}{\sigma\sqrt{\hat{\mathbf{\Omega}}_{jj}}} \quad \text{and} \quad \mu_j = \frac{\sqrt{n}\beta_j}{\sigma\sqrt{\hat{\mathbf{\Omega}}_{jj}}}. \tag{2.6}$$

2.2. *Approximating false negative proportion*

We aim to determine a cut-off value for the realized test statistics to control false negative proportion (FNP) at an user-specified level. For this purpose, we first study the consistent estimation of FNP. For any $t > 0$, define

$$R(t) = \sum_{j=1}^{p} 1_{\{|z_j|>t\}}, \qquad \mathsf{TP}(t) = \sum_{j\in I_1} 1_{\{|z_j|>t\}},$$
$$\mathsf{FN}(t) = \sum_{j\in I_1} 1_{\{|z_j|\le t\}}, \qquad \mathsf{FP}(t) = \sum_{j\in I_0} 1_{\{|z_j|>t\}}.$$

Note that $\mathsf{FN}(t)$ is unobservable as I_1 is unknown, and that the dependence among z_j's also affect $\mathsf{FN}(t)$. It is easy to see that

$$\mathsf{FN}(t) = s - \mathsf{TP}(t) = s - [R(t) - \mathsf{FP}(t)] \tag{2.7}$$

and

$$\mathsf{FNP}(t) = \frac{\mathsf{FN}(t)}{s} = 1 - \frac{R(t) - \mathsf{FP}(t)}{s}. \tag{2.8}$$

Since $R(t)$ is directly observable from the data, the unknown quantities in (2.8) are $\mathsf{FP}(t)$ and s. We propose to substitute $\mathsf{FP}(t)$ in (2.7) by $2(p-s)\Phi(-t)$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard Normal random variable, because z_j is asymptotically standard Normal for $j \in I_0$. Further, we can plug in an estimator $\hat{s}$ for s, which results in the estimator

$$\widehat{\mathsf{FNP}}(t) = 1 - \frac{R(t) - 2(p-\hat{s})\Phi(-t)}{\hat{s}}. \tag{2.9}$$

From the definitions of $\mathsf{FNP}(t)$ and $\widehat{\mathsf{FNP}}(t)$, it can be shown that $\left|\widehat{\mathsf{FNP}}(t) - \mathsf{FNP}(t)\right| = o_P(1)$ is implied by

$$s^{-1}\left|\mathsf{FP}(t) - 2p_0\Phi(-t)\right| = o_P(1) \qquad \text{and} \qquad |\hat{s}/s - 1| = o_P(1). \tag{2.10}$$

Because $\mathsf{FP}(t)$ is the summation of p_0 terms and s can be much smaller than p_0, approximating $\mathsf{FP}(t)/s$ requires more delicate analysis than approximating $\mathsf{FP}(t)/p_0$, which has been studied in the literature for False Discovery Proportion

(FDP) control (e.g. [13]). Also, the dependence among test statistics $\{z_j\}_{j=1}^p$ adds another layer of difficulty.

In this paper, we consider $s = p^{1-\eta}$ for some $\eta \in (0,1)$, so that the number of relevant predictors is of a smaller order than the total number of variables. On the other hand, we consider t values calibrated as $t = t_\xi = \sqrt{2\xi \log p}$ for some $\xi > 0$, so that the scale of t is comparable to that of the extreme value of p independent standard Gaussian variables. Such calibration has been utilized to study the detection of Gaussian mixtures [2, 7, 12], and to analyze variable selection consistency in linear regression [23]. In this paper, we adopt the calibration to study the estimation of FNP in linear regression.

Further, define the precision matrix of the predictors as $\mathbf{\Theta}$ and let

$$s_j = |\{k \neq j : \mathbf{\Theta}_{jk} \neq 0\}| \qquad \text{and} \qquad s_{max} = \max_{1 \leq j \leq p} s_j.$$

Namely, the parameter s_{max} represents the row-sparsity of the precision matrix, which contributes to the strength of the dependence among the test statistics. Define

$$\gamma_1^* = 2\eta - \min\{1, \frac{\log(n/s_{max})}{2\log p}\}, \qquad \gamma_2^* = 2 - 2\eta - \frac{\log n}{2\log p},$$

and

$$\gamma^* = \max\{\gamma_1^*, \gamma_2^*\}. \tag{2.11}$$

The next theoretical result demonstrates the range of t values in which the first equation $s^{-1}\left|\mathsf{FP}(t) - 2p\Phi(-t)\right| = o_P(1)$ in (2.10) is achievable.

Theorem 2.1. *Consider model (1.1) and the test statistics $\{z_j\}_{j=1}^p$ in (2.5). Assume conditions A1) through A3) in Appendix A.1 for the asymptotic normality of $\{z_j\}_{j=1}^p$. Let $s = p^{1-\eta}$ for some $\eta \in (0,1)$ and $t = t_\xi = \sqrt{2\xi \log p}$ for $\xi > 0$. Assume $\xi > \min\{\eta, \gamma^*\}$ for γ^* in (2.11), then*

$$s^{-1}\left|\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi)\right| = o_P(1). \tag{2.12}$$

Because $\mathsf{FP}(t)$ is the summation of p_0 indicator functions and $p_0 \gg s(= p^{1-\eta})$, $\mathsf{FP}(t)/s$ blows up at constant t. Theorem 2.1 says that the approximation of $\mathsf{FP}(t)/s$ is achievable for t at the scale of t_ξ. This is substantially different from the existing study of FDP control, where the approximation of $\mathsf{FP}(t)/p_0$ and $R(t)/p_0$ are studied at constant t.

The condition $\xi > \min\{\eta, \gamma^*\}$ can be decomposed as follows. When $\eta \leq \gamma^*$, we have $\xi > \eta$, and the claim in (2.12) follows by showing that $s^{-1}\mathsf{FP}(t_\xi) = o_p(1) = s^{-1}p_0\Phi(-t_\xi)$. On the other hand, when $\eta > \gamma^*$ and $\gamma^* < \xi \leq \eta$, more delicate analysis is needed to study the variability of $\mathsf{FP}(t_\xi)$. The condition $\xi > \gamma_1^*$ essentially controls the variability of $s^{-1}\mathsf{FP}_{\mathbf{w}'}(t_\xi)$, where $\mathbf{w}'$ is the Gaussian component of $\mathbf{z}$ as in (2.6) and $\mathsf{FP}_{\mathbf{w}'}(t_\xi) = \sum_{j \in I_0} 1_{\{|w_j'| > t_\xi\}}$. The condition $\xi > \gamma_2^*$ controls the cumulative errors caused by the component $\boldsymbol{\delta}'$ of $\mathbf{z}$.

Existing study in [23] has shown optimal phase diagram in (ξ, η) for high-dimensional variable selection. Their work, however, focuses on scenarios with

$\xi > \eta$. We extend the analysis to the more challenging case with $\gamma^* < \xi \leq \eta$, for which we study the variability of $s^{-1}\mathsf{FP}_{w'}(t_\xi)$ under the dependence of test statistics. Recall the covariance matrix $\sigma^2\hat{\mathbf{\Omega}}$ in (2.3). Since $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}^T$ and that $\hat{\mathbf{\Sigma}}$ is not a sparse matrix, $\sigma^2\hat{\mathbf{\Omega}}$ is not sparse or possessing any well-known structures. The study in [23] imposes conditions on the covariance matrix of predictors that essentially prohibit excessive signal cancellations when performing marginal regression. Our condition of dependence, on the other hand, demonstrate the effect of the sparsity of precision matrix (s_{max}) through γ^*. Overall, $\xi > \min\{\eta, \gamma^*\}$ is easier to be satisfied with larger n, smaller p, or smaller s_{max}.

To achieve the second equation in (2.10), we modify the estimator introduced in [29] and study its consistency for estimating s in our setting. We refer to the modified estimator as the MR estimator. Recall the standardized debiased Lasso estimator $z_j = \sqrt{n}\hat{b}_j\sigma^{-1}\hat{\mathbf{\Omega}}_{jj}^{-1/2}, 1 \leq j \leq p$. Let $F_p(t) = p^{-1}\sum_{j=1}^p 1_{\{|z_j|>t\}}$ and $\bar{\sigma}(t) = \sqrt{2\bar{\Phi}(t)\left(1-2\bar{\Phi}(t)\right)}$, where $\bar{\Phi}(t) = 1 - \Phi(t)$. The MR estimator for the portion of relevant predictors ($\pi = s/p$) is constructed as

$$\hat{\pi} = \sup_{t>0} \frac{F_p(t) - 2\bar{\Phi}(t) - c_p\bar{\sigma}(t)}{1 - 2\bar{\Phi}(t)}, \tag{2.13}$$

where c_p is a bounding sequence pre-specified as follows. Define

$$G_p(t) = p^{-1}\sum_{j=1}^p 1_{\{|w_j'|>t\}},$$
$$\mathsf{H}(t) = \frac{G_p(t) - 2\bar{\Phi}(t)}{\bar{\sigma}(t)}, \qquad \text{and} \qquad V_p = \sup_{t>0}\mathsf{H}(t). \tag{2.14}$$

Set c_p as the $(1-\alpha_p)$-th quantile of V_p for $\alpha_p = o(1)$, so that $P(V_p > c_p) = \alpha_p \to 0$ as $p \to \infty$. In other words, c_p can be looked upon as an upper bound for V_p probabilistically, and the implement of c_p in (2.13) eventually controls over-estimation on π.

Compared to the original MR estimator in [29], the key modification in (2.13) and (2.14) is the use of $F_p(t)$ and $G_p(t)$, two empirical processes each with dependent random summands. Naturally, this requires different techniques to find $\{c_p\}_{p\geq 1}$. The setting in [29] considers independent p-values that are uniformly distributed under the null hypothesis. Since the limiting distribution of the uniform empirical process with independent summands is known and has an analytic expression, a bounding sequence can be directly found from the distribution in the construction of the original MR estimator. However, in our settings $\{z_j\}_{j=1}^p$ are dependent, and the exact distributions of $\{\hat{b}_j\}_{j=1}^p$ are unspecified.

In fact, $\{\hat{b}_j\}_{j=1}^p$ asymptotically has covariance matrix $\sigma^2\hat{\mathbf{\Omega}} = \sigma^2\hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}^T$. In theory, $|\hat{\mathbf{\Omega}}_{ij} - \mathbf{\Theta}_{ij}| = o_p(1)$ for any (i,j) under conditions A1) through A3) in Appendix A.1. However, $\hat{\mathbf{\Omega}}$ itself is neither diagonal nor sparse, and the approximation errors of all the elements in $\hat{\mathbf{\Omega}}$ add up to influence V_p. Note that V_p is the higher criticism statistic of [12] based on the Gaussian component

$\mathbf{w}'$ of $\mathbf{z}$. Unfortunately, existing techniques for higher criticism statistic under short-range and long-range dependence [18] cannot be applied here because our test statistics with covariance matrix $\sigma^2\hat{\mathbf{\Omega}}$ cannot be partitioned as in [18].

In this paper, we employ a discretization technique adopted from [1] to derive bounds on the variance of a discretized $\{\mathsf{H}(t) : t > 0\}$ and define a discretized version of V_p as

$$V_p^* = \max\left\{\mathsf{H}(t) : t \in \left[\sqrt{\tau_0 \log p}, \sqrt{\tau_1 \log p}\right] \cap \mathbb{N}\right\} \tag{2.15}$$

for two positive constants τ_0 and τ_1 such that $0 < \tau_0 < \tau_1$. Then, a discretized version of the MR estimator is defined as

$$\hat{\pi}^* = \max\left\{\frac{F_p(t) - 2\bar{\Phi}(t) - c_p^*\bar{\sigma}\,(t)}{1 - 2\bar{\Phi}(t)} : t \in \left[\sqrt{\tau_0 \log p}, \sqrt{\tau_1 \log p}\right] \cap \mathbb{N}\right\}, \tag{2.16}$$

where c_p^* is the $(1-\alpha_p)$-th quantile of V_p^* for $\alpha_p = o(1)$. Let

$$\mu_{min} = \min_{j \in I_1} \sqrt{n}\,|\beta_j|\,\sigma^{-1}\mathbf{\Theta}_{jj}^{-1/2} \tag{2.17}$$

as a measure on the minimal effect size of relevant variables. The following theorem demonstrates the consistency of $\hat{\pi}^*$. Its proof is presented in Section 6.2.

Theorem 2.2. *Assume conditions A1) through A3) in Appendix A.1 for the asymptotic normality of* $\{z_j\}_{j=1}^p$. *Let* $s = p^{1-\eta}$ *for some* $\eta \in (0,1)$ *and* $\mu_{min} \geq \sqrt{2(\gamma^* + c)\log p}$ *for* μ_{min} *and* γ^* *in (2.17) and (2.11) and some constant* $c > 0$. *Then* $\hat{\pi}^*$ *with bounding sequence* c_p^* *at the order of* $(s_{\max}/n)^{1/4}\log p$, $\tau_0 \in (2\gamma^*, 2\gamma^* + c)$, *and* $\tau_1 > 2(\gamma^* + c)$ *consistently estimates the proportion* π *of relevant predictors, i.e., for any* $\delta > 0$,

$$P(|\hat{\pi}^*/\pi - 1| < \delta) \to 1$$

and, consequently, $\hat{s} = \hat{\pi}^* p$ *satisfies*

$$P(|\hat{s}/s - 1| < \delta) \to 1.$$

Note that the order of c_p^* shows the effects of sparsity of the precision matrix (s_{max}), sample size (n), and dimensionality (p) on V_p^*. The condition $\mu_{min} \geq \sqrt{2(\gamma^* + c)\log p}$ shows that consistent estimation of s gets easier with smaller s_{max}, larger n, and smaller p.

In summary, Theorem 2.1 and Theorem 2.2 facilitate the two equations in (2.10) for $\mathrm{FNP}(t)$ estimation by $\widehat{\mathsf{FNP}}(t)$. Note that in practice we will need to simulate V_p and c_p to derive the estimated s and $\mathsf{FNP}(t)$. Please refer to Section 3.2 for details of the numerical implementation.

3. FNP control at a user-specified level

In this section, we introduce a new method for FNP control at a user-specified level in high-dimensional regression. We say that a variable selection method

asymptotically controls FNP at a pre-specified level $\epsilon \in (0,1)$ if the FNP of its selection outcome satisfies

$$P\left(\mathsf{FNP} < \epsilon\right) \to 1.$$

Such methods are useful in applications where data dimensions need to be largely reduced for subsequent analyses while controlling false negatives at a tolerable level.

3.1. The FNC-Reg procedure

Based on the approximation results of FNP, we propose the False Negative Control for Regression (FNC-Reg) procedure, which determines the cut-off threshold on the list of ranked $\{|z_j|\}_{j=1}^p$ as

$$t^*(\epsilon) = \sup\left\{t : \widehat{\mathsf{FNP}}(t) \le \epsilon\right\} \tag{3.1}$$

for an user-specified $\epsilon \in (0,1)$. FNC-Reg selects predictors with $|z_j| > t^*(\epsilon)$.

It can be seen that FNC-Reg is a procedure built upon direct estimation of FNP and a user-specified control level of FNP. Given that $\widehat{\mathsf{FNP}}(t)$ is non-increasing with t, FNC-Reg selects the smallest subset of $\{z_j\}_{j=1}^p$ such that the estimated FNP is less than ϵ. Moreover, this procedure depends on user's preference for the control level of FNP. Since $t^*(\epsilon)$ is non-decreasing with ϵ, if users can tolerate missing a higher proportion (larger ϵ) of relevant variables, they may select less variables using the procedure. The selected subset of variables can be much smaller than the full set of variables, which corresponds to no false negatives. The next theorem shows that under certain conditions, the FNC-Reg procedure asymptotically controls the true FNP at the level of ϵ.

Theorem 3.1. *Assume conditions A1) through A3) in Appendix A.1. Assume* $\mu_{min} \ge \sqrt{2(\gamma^* + c)\log p}$ *for* μ_{min} *and* γ^* *in (2.17) and (2.11) and some constant* $c > 0$. *Then* $t^*(\epsilon)$ *determined by (3.1) with* $\hat{s} = \hat{\pi}^* p$ *satisfies*

$$P\left(\mathsf{FNP}(t^*(\epsilon)) \le \epsilon\right) \to 1. \tag{3.2}$$

Consequently, $t^*(\epsilon)$ *determined by (3.1) with* $\hat{s} = \hat{\pi} p$ *and* $c_p = c_p^*$ *also satisfies*

$$P\left(\mathsf{FNP}(t^*(\epsilon)) \le \epsilon\right) \to 1. \tag{3.3}$$

Result in (3.2) shows the FNP control by FNC-Reg when the discretized MR estimator is implemented in (2.9). (3.3) extends the result to FNC-Reg with the MR estimator.

We compare the condition on $\mu_{\min}$ in Theorem 3.1 with the beta-min condition of variable selection consistency. Our condition on $\mu_{\min}$ achieves the order $O(\sqrt{(\log p)/n})$ for β_{min}, which is the optimal oder for variable selection consistency [19, 35]. On the other hand, Our condition on $\mu_{\min}$ specifies the constant term $\sqrt{2\gamma^*}$ with γ^* in (2.11), while existing beta-min conditions for

different methods have various constant terms that are often not fully specified. Therefore, we attempt to compare with the optimal constant term for variable selection consistency in the ideal setting, where the predictors $(X_{i1}, \ldots, X_{ip})$ are generated as i.i.d. samples from $N(0, I_{p\times p})$. Existing study in, for example, [23] has shown that the optimal constant is $\sqrt{2} + \sqrt{2(1-\eta)}$, i.e. smaller η (larger s) makes it harder to perfectly separate all the signals from noise. Then, it follows that $\sqrt{2\gamma^*} < \sqrt{2} + \sqrt{2(1-\eta)}$ for any $\eta \in (0,1)$. The above analysis shows that in the ideal setting, our condition on $\mu_{\min}$ is weaker than the optimal beta-min condition for variable selection consistency, and that FNP control can be achieved by FNC-Reg when relevant and irrelevant variables may not be perfectly separated.

3.2. Numerical implementation of FNC-Reg

We provide a computational algorithm to implement the proposed FNC-Reg procedure. First, the estimation of s relies on the bounding sequence c_p, which is pre-fixed as the $(1-\alpha_p)$-th quantile of V_p. In numerical implementation, we suggest to simulate V_p and c_p as follows. We simulate the data under the global null hypothesis that no relevant predictors exist and calculate the corresponding standardized debiased Lasso estimator $\tilde{z}_j$. Note that $\tilde{z}_j$ is asymptotically distributed as w'_j under the global null hypothesis. We order $\tilde{z}_j$'s by their absolute values such that $|\tilde{z}_{(1)}| > |\tilde{z}_{(2)}| > \ldots > |\tilde{z}_{(p)}|$ and calculate

$$\tilde{V}_p = \max_{1<j<p/2} \frac{j/p - 2\bar{\Phi}(|\tilde{z}_{(j)}|)}{\bar{\sigma}\left(|\tilde{z}_{(j)}|\right)}. \tag{3.4}$$

Repeat the above 1000 times and determine $\tilde{c}_p$ as the $(1-1/\sqrt{\log p})$-th quantile of the empirical distribution of $\tilde{V}_p^{(1)}, \ldots, \tilde{V}_p^{(1000)}$. Consequently, given the ordered test statistics $|z_{(1)}| > |z_{(2)}| > \ldots > |z_{(p)}|$, calculate

$$\tilde{\pi} = \max_{1<j<p/2} \frac{j/p - 2\bar{\Phi}(|z_{(j)}|) - \tilde{c}_p\bar{\sigma}\left(|z_{(j)}|\right)}{1 - 2\bar{\Phi}(|z_{(j)}|)}. \tag{3.5}$$

Algorithm 1 FNC-Reg

1. Derive the debiased Lasso estimator $\hat{\mathbf{b}}$ as in (2.2).
2. Standardize $\hat{\mathbf{b}}$ and obtain $z_j = \sqrt{n}\hat{b}_j\sigma^{-1}\hat{\mathbf{\Omega}}_{jj}^{-1/2}$ for $1 \le j \le p$. Order the $\{z_j\}_{j=1}^p$ as $|z_{(1)}| > |z_{(2)}| > \ldots > |z_{(p)}|$.
3. Calculate the bounding sequence $\tilde{c}_p$ as the $(1-1/\sqrt{\log p})$-th quantile of the empirical distribution of $\tilde{V}_p$ in (3.4).
4. Obtain $\tilde{\pi}$ by (3.5) and $\hat{s} = \tilde{\pi}p$.
5. Calculate $\widehat{\mathsf{FNP}}(|z_{(j)}|)$ for $j = 1, \ldots, p$ by (2.9).
6. Obtain $t^*(\epsilon) = \max\{|z_{(j)}| : \widehat{\mathsf{FNP}}(|z_{(j)}|) \le \epsilon\}$ for a user-specified $\epsilon > 0$.
7. Select predictors with $|z_j| \ge t^*(\epsilon)$.

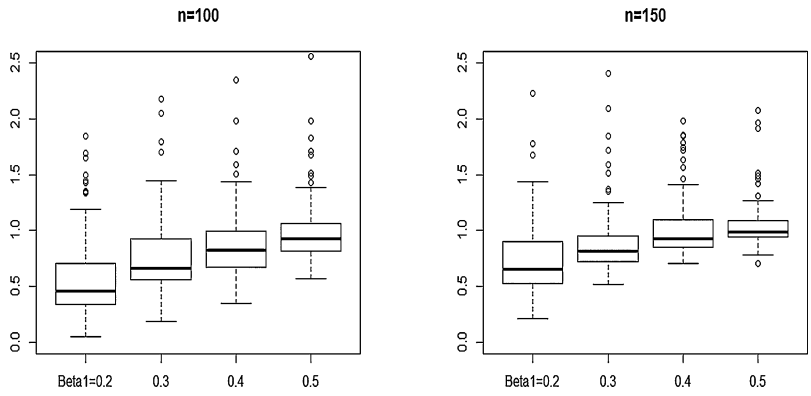

FIG 1. *Box-plots of $\hat{s}/s$ with $p = 200, s = 10$, and β_1 increasing from 0.2 to 0.5. The left plot has $n = 100$ and the right plot has $n = 150$.*

4. Numerical analysis

Examples in this section have the response y simulated by the regression model (1.1) with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \mathbf{I})$. Each row of $\mathbf{X}$ is simulated from $\mathcal{N}_p(0, \boldsymbol{\Sigma})$. We use the Ergös-Rényi random graph in [8] to generate the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ with $s_{max} \sim \mathsf{Binomial}(p, \theta)$, such that the nonzero elements of $\boldsymbol{\Theta}$ are randomly located in each of its rows with magnitudes randomly generated from the uniform distribution $\mathsf{Uniform}[0.4, 0.8]$. The nonzero coefficients are set at $\beta_1, \ldots, \beta_s$ with the same values. The debiased Lasso estimates are obtained by applying the R package *hdi* [11].

4.1. Estimating s

We compare the estimated $\hat{s}$ with the true s in two settings. The first setting has $p = 200, n = 100, s = 10, \theta = 0.02$, and $\beta_1 = 0.2 - 0.5$. The second setting increases sample size n to 150. As claimed in Theorem 2.2, the accuracy of $\hat{s}$ increases with the magnitude of non-zero coefficients and the sample size. Figure 1 presents the box-plots of the ratio $\hat{s}/s$ from 100 replications. When β_1 or sample size is small, $\hat{s}$ tends to under-estimate the true s. As β_1 increases from 0.2 to 0.5 or n increases from 100 to 150, $\hat{s}/s$ concentrates more around 1.

4.2. FNP control

We apply the FNC-Reg algorithm presented in Section 3.2 to the simulated data with $p = 200$, $n = 150$, $s = 10$, $\theta = 0.02$, and $\beta_1 = 0.2 - 0.5$. Table 1 has ϵ fixed at 0.1 and reports the mean value of $\mathsf{FNP}(t^*)$ as β_1 increased from 0.2 to 0.5. We also calculated the associated false discovery proportion ($\mathsf{FDP}(t^*) = \mathsf{FP}(t^*)/R(t^*)$) to reveal the price in incurring false positives for FNP control. Further, we calculate the F-measure, which summarizes FNP and FDP by the

harmonic mean of (1-FNP) and (1-FDP) [30]. F-measure takes a value between 0 and 1, and higher value corresponds to better summarized performance.

Because we are not aware of any existing methods that directly control FNP in high-dimensional regression, we present the corresponding results of two other methods that perform variable selection based on different criteria. These results help to better understand the results of FNC-Reg. The first method is Lasso whose solution is obtained using the R package *hdi*, in which λ is determined by cross validation. The second method is Knockoff, which has been developed to control false discovery rate (FDR) at a user-specified level in high-dimensional regression [3, 9]. We use the "knockoff.filter" function in default from the R package *knockoff*, which creates model-X second-order Gaussian knockoffs as introduced in [9]. The nominal level is set at 0.1.

TABLE 1
The mean values and standard deviations (in brackets) of FNP, FDP, and the F-measure from 100 replications for FNC-Reg, Lasso, and Knockoff.

β_1	Method	FNP	FDP	F-measure
0.2	FNC-Reg	0.37 (0.16)	0.35 (0.30)	0.58 (0.20)
	Lasso	0.29 (0.12)	0.51 (0.16)	0.56 (0.16)
	Knockoff	0.91 (0.23)	0.06 (0.16)	0.08 (0.21)
0.3	FNC-Reg	0.19 (0.11)	0.20 (0.24)	0.77 (0.17)
	Lasso	0.15 (0.09)	0.43 (0.11)	0.67 (0.09)
	Knockoff	0.60 (0.44)	0.11 (0.15)	0.37 (0.41)
0.4	FNC-Reg	0.10 (0.09)	0.17 (0.25)	0.84 (0.19)
	Lasso	0.06 (0.07)	0.42 (0.12)	0.71 (0.10)
	Knockoff	0.34 (0.45)	0.12 (0.14)	0.61 (0.42)
0.5	FNC-Reg	0.04 (0.09)	0.13 (0.22)	0.90 (0.18)
	Lasso	0.02 (0.10)	0.38 (0.12)	0.74 (0.14)
	Knockoff	0.21 (0.49)	0.09 (0.11)	0.74 (0.38)

It can be seen from Table 1 that as β_1 increases, the FNP of FNC-Reg decreases, which agrees with the theoretical insight provided by the condition on μ_{min} in Theorem 3.1. In the challenging scenarios where β_1 is very small, the FNP of FNC-Reg mostly exceeds the nominal level of 0.1, which is due to the under-estimation of $\hat{s}$ and FNC-Reg's tendency to select less variables to capture the under-estimated number of signals. Furthermore, both FNP and FDP of FNC-Reg get smaller for larger β_1, suggesting that FNC-Reg automatically adapt to and benefit from increasing signal intensity for both false negative and false positive control. Table 1 also shows that Lasso has lower FNP and much higher FDP than FNC-Reg, which agrees with Lasso's known tendency of over-selection when $p > n$. On the other hand, Knockoff has FDP reasonably controlled at the nominal level of 0.1 but much higher FNP than those of FNC-Reg and Lasso. In terms of the F-measure that summarizes FNP and FDP, FNC-Reg seems to outperform the other two methods under different β_1 values.

We further illustrate the adaptivity of FNC-Reg to the user-specified control level of FNP. For various values of ϵ, we calculate the relative frequency of the event $\{\mathsf{FNP}(t^*(\epsilon)) \leq \epsilon\}$. Table 2 summarizes the results for different settings with $\epsilon = 0.1, 0.2, 0.3$ and $\beta_1 = 0.3, 0.5, 0.7$. It can be seen that the relative

frequency of $\mathsf{FNP} \leq \epsilon$ for FNC-Reg increases with β_1, which is consistent with the theoretical insight in Theorem 3.1. On the other hand, for a fixed β_1, the relative frequency of $\mathsf{FNP} \leq \epsilon$ and the FDP of FNC-Reg decreases with ϵ, which agrees with our expectation for FNC-Reg as more liberal control of FNP incurs less price in false positives. Note that the results of Lasso and Knockoff do not change with the varying ϵ.

TABLE 2
The relative frequencies of the event $\{\mathsf{FNP} < \epsilon\}$ *and mean values of FDP from 100 replications for FNC-Reg.*

	$\beta_1 = 0.3$			$\beta_1 = 0.5$			$\beta_1 = 0.7$		
	$\epsilon = 0.1$	0.2	0.3	$\epsilon = 0.1$	0.2	0.3	$\epsilon = 0.1$	0.2	0.3
$1(\mathsf{FNP} \leq \epsilon)$	0.38	0.53	0.71	0.72	0.86	0.91	0.98	0.98	0.98
FDP	0.20	0.15	0.09	0.13	0.08	0.05	0.14	0.10	0.07

5. Conclusion and discussion

We propose a new variable selection method, FNC-Reg, to efficiently control false negatives in linear regression. Different from existing methods and theory for power analysis and Sure Screening, our procedure directly estimates the FNP of a decision rule and selects the smallest subset of variables that has the estimated FNP less than a user-specified control level. FNP control is specifically challenging when relevant variables cannot be consistently separated from irrelevant ones due to limited sample size and effect size. We develop new techniques to analyze FNP control in the challenging setting and to cope with difficulties caused by the dependence of test statistics.

FNC-Reg possesses two types of adaptivity property. First, it adapts to the user's preference level on the control of FNP. When a user can tolerate a less stringent control on FNP, he or she can input a larger ϵ in the FNC-Reg procedure and select less variables with less false positives. Secondly, the proposed method is adaptive to the unknown effect size. Note that the implementation of the procedure does not requires the information of effect size. Nevertheless, the result of the procedure automatically improves in both FNP and FDP as effect size increases.

Our theoretical study presents a weaker condition on μ_{min} for FNP control by FNC-Reg than the beta-min condition for variable selection consistency. It is also of interest to understand the result of FNC-Reg if the condition on $\mu_{\min}$ may not be satisfied. Assume that among the s signal variables only s_1 of them satisfy $\mu_j \geq \sqrt{2(\gamma^* + c)\log p}$ for some constant $c > 0$. Then, similar arguments as in the proof of Theorem 2.2 can be applied to show that $P((1-\delta)s_1 < \hat{s} < (1+\delta)s) \to 1$ for any $\delta > 0$. Note that $\hat{s}$ does not consistently estimate s anymore, nor is it a consistent estimator for s_1. Such $\hat{s}$ tends to under-estimate s, which can cause the proposed method to select less variables to capture the under-estimated number of signals. Because FNC-Reg ranks the test statistics by their significance and select variables from the top, one can make a statement

about FNP control for the signals with effect sizes larger than the observed cut-off position. Such interpretation of results remains valid whether the condition on μ_{min} holds or not.

Last but not least, we adopt the debiased Lasso estimator as the test statistic in the paper to demonstrate the new analytic framework of FNP control. We expect that the proposed framework can incorporate other test statistics in linear regression and promote further developments in false negative control based variable selection.

6. Proofs

This section contains the proofs of Theorem 2.1, Theorem 2.2, and Theorem 3.1. Auxiliary lemmas are provided in the appendices. We will frequently use the Mill's ratio, i.e.,

$$\bar{\Phi}(x) = x^{-1}\phi(x)(1+o(1)) \quad \text{for} \quad x \to +\infty,$$

without mentioning it at each instance. All arguments will be conditional on $\mathbf{X}$, and the symbol C denotes a generic, finite constant whose values can be different at different occurrences.

6.1. Proof of Theorem 2.1

The proof is composed of two parts. The first part assumes $\xi > \eta$ and the second part assumes $\xi \leq \eta$.

Consider the first part with $\xi > \eta$. It suffices to show $s^{-1}\mathsf{FP}(t_\xi) = o_P(1)$ and $s^{-1}p_0\Phi(-t_\xi) = o(1)$ with $\xi > \eta$. By Mill's ratio,

$$s^{-1}p\Phi(-t_\xi) \leq Cp^{\eta-\xi}/\sqrt{\log p} = o(1)$$

when $\xi > \eta$. On the other hand, for a fixed constant $a > 0$,

$$P\left(s^{-1}\mathsf{FP}(t_\xi) > a\right) \leq \frac{E(\mathsf{FP}(t_\xi))}{as} = \frac{p_0 \max_{j\in I_0} P(|z_j| > t_\xi)}{as}.$$

The following lemma help quantify the order of $P(|z_j| > t_\xi)$ for $j \in I_0$, and its proof is provided in Section A.3.

Lemma 6.1. *Assume A1) through A3). Define*

$$d_p = C_1\left(\sqrt{s}\log p/\sqrt{n} + \min\{s, s_{max}\}\log p/\sqrt{n}\right)$$

for some constant $C_1 \geq \max\left\{1, 2(\sigma\sqrt{C_{\min}})^{-1}\right\}$. *Then*

$$\left|P(|w_j' - \delta_j'| > t_\xi) - P(|w_j'| > t_\xi)\right| \leq Cp^{-\xi}d_p + Cp^{-2}.$$

Recall $s = p^{1-\eta}$ with $0 < \eta < 1$. Then Lemma 6.1 implies

$$P\left(s^{-1}\mathsf{FP}\left(t_\xi\right) > a\right) \leq \frac{Cp_0\left(p^{-\xi} + p^{-\xi}d_p + p^{-2}\right)}{p^{1-\eta}} = o\left(1\right),$$

where the last step is by $d_p = o(1)$ under A3) and $\xi > \eta$. Then $s^{-1}\mathsf{FP}\left(t_\xi\right) = o_P\left(1\right)$. This justifies the claim of Theorem 2.1 for $\xi > \eta$.

Next, we present the second part of the proof with $\xi \leq \eta$. Define

$$D_p = s^{-2}(p^{1-\xi} + p^{2-\xi}\lambda_1\sqrt{s_{max}})\log p.$$

By the order of λ_1 in A2) and condition $\xi > \gamma_1^*$, $D_p = o(1)$. Then it is sufficient to show

$$P(s^{-1}\left|\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi)\right| > \sqrt{D_p}) \to 0.$$

Perform the decomposition

$$\begin{aligned} s^{-1}\left|\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi)\right| \quad &\leq \quad s^{-1}\left|\mathsf{FP}(t_\xi) - \mathsf{E}(\mathsf{FP}(t_\xi))\right| \\ &+ \quad s^{-1}\left|\mathsf{E}(\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi))\right|. \end{aligned}$$

Then it is sufficient to show

$$s^{-1}\left|\mathsf{FP}(t_\xi) - \mathsf{E}(\mathsf{FP}(t_\xi))\right| = o_p(\sqrt{D_p}) \tag{6.1}$$

and

$$s^{-1}\left|\mathsf{E}(\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi))\right| = o(\sqrt{D_p}). \tag{6.2}$$

Consider (6.1) first. By Chebyshev's inequality,

$$P(s^{-1}\left|\mathsf{FP}(t_\xi) - \mathsf{E}(\mathsf{FP}(t_\xi))\right| > \sqrt{D_p}) \leq \frac{\mathsf{Var}(\mathsf{FP}(t_\xi))}{s^2 D_p}.$$

We derive the order of $\mathsf{Var}(\mathsf{FP}(t_\xi))$. By Lemma 6.1, $P(|w_j' - \delta_j'| > t_\xi) = P(|w_j'| > t_\xi)(1 + o(1))$ given $d_p = o(1/\sqrt{\log p})$ from A3) and $\xi \leq \eta < 1$, then direct calculation gives

$$\mathsf{Var}(\mathsf{FP}(t_\xi)) = \mathsf{Var}(\mathsf{FP}_{w'}(t_\xi))(1 + o(1)),$$

where $\mathsf{FP}_{w'}(t_\xi) = \sum_{j\in I_0} 1_{\{|w_j'|>t_\xi\}}$. The following lemma is proved in Section A.4.

Lemma 6.2. *Assume A1) and A2) and let* $t_\xi = \sqrt{2\xi\log p}$ *for any* $\xi > 0$. *Then*

$$\mathsf{Var}\left(\sum_{j=1}^{p} 1_{\{|w_j'|>t_\xi\}}\right) = O\left(p^{1-\xi} + p^{2-\xi}\lambda_1\sqrt{s_{max}}\right).$$

The above gives

$$\frac{\mathsf{Var}(\mathsf{FP}(t_\xi))}{s^2 D_p} = \frac{\mathsf{Var}(\mathsf{FP}_{w'}(t_\xi))}{s^2 D_p}(1 + o(1)) = o(1),$$

so that

$$P(s^{-1}\,|\mathsf{FP}(t_\xi) - \mathsf{E}(\mathsf{FP}(t_\xi))| > \sqrt{D_p}) \to 0.$$

Next consider (6.2). By Lemma 6.1

$$\begin{aligned} s^{-1}\,|\mathsf{E}(\mathsf{FP}(t_\xi) - 2p_0\Phi(-t_\xi))| &\leq s^{-1}\sum_{j\in I_0}\left|P(|w_j' - \delta_j'| > t_\xi) - P(|w_j'| > t_\xi)\right| \\ &\leq Cs^{-1}p^{1-\xi}d_p + Cs^{-1}p^{-1}. \end{aligned}$$

Recall $s = p^{1-\eta}$ and the definitions of d_p and D_p. Note that $d_p > s\log p/\sqrt{n}$, then direct calculation gives

$$s^{-1}p^{1-\xi}d_p = o(\sqrt{D_p})$$

under condition $\xi > \gamma_2^*$, and

$$s^{-1}p^{-1} = o(\sqrt{D_p})$$

with $\xi \leq \eta < 1$. (6.2) follows consequently. This concludes the second part of the proof with $\xi \leq \eta$.

6.2. Proof of Theorem 2.2

First, we have the following lemma showing the order of the bounding sequence c_p^* for V_p^*. The proof is provided in Section A.5.

Lemma 6.3. *Assume conditions A1) through A3) in Appendix A.1. Consider V_p^* as in (2.15). Then c_p^* at the order of $(s_{max}/n)^{1/4}\log p$ satisfies $P(V_p^* > c_p^*) \to 0$ as $p \to \infty$.*

Now, recall $F_p(t) = p^{-1}\sum_{j=1}^p 1_{\{|z_j|>t\}}$ and define

$$\bar{\Phi}_p(t) = p^{-1}\sum_{j=1}^p 1_{\{|\mu_j + w_j'|>t\}}.$$

Consider the decomposition

$$\hat{\pi}^* = \max_{t\in\mathbb{T}}\left\{\frac{F_p(t) - \bar{\Phi}_p(t)}{1 - 2\bar{\Phi}(t)} + \frac{\bar{\Phi}_p(t) - 2\bar{\Phi}(t) - c_p^*\bar{\sigma}\,(t)}{1 - 2\bar{\Phi}(t)}\right\}, \tag{6.3}$$

where $\mathbb{T}$ is defined in (A.10). The first summand within the parentheses on the right hand side (RHS) of (6.3) can be safely ignored when bounding $\hat{\pi}^*/\pi$ as asserted by the following Lemma 6.4.

Lemma 6.4. *Assume $t = t_\xi = \sqrt{2\xi\log p}$ with $\xi > \gamma^*$. Then*

$$\pi^{-1}\left|F_p(t) - \bar{\Phi}_p(t)\right|\left(1 - 2\bar{\Phi}(t)\right)^{-1} = o_P(1).$$

Define

$$\hat{\pi}^{**} = \max_{t\in\mathbb{T}} \frac{\bar{\Phi}_p(t) - 2\bar{\Phi}(t) - c_p^*\bar{\sigma}\,(t)}{1 - 2\bar{\Phi}(t)}.$$

Then it suffices to show

$$P(1-\delta < \hat{\pi}^{**}/\pi < 1) \to 1. \tag{6.4}$$

We first show that $\hat{\pi}^{**}$ is an asymptotic lower bound of π. Recall the definition of V_p^* as

$$V_p^* = \max_{t\in\mathbb{T}} \frac{p^{-1}\sum_{j=1}^{p} 1_{\{|w_j'|>t\}} - 2\bar{\Phi}(t)}{\bar{\sigma}(t)}.$$

Since

$$\bar{\Phi}_p(t) \le p^{-1}s + p^{-1}\sum_{j\in I_0} 1_{\{|w_j'|>t\}} = \pi + (1-\pi)p_0^{-1}\sum_{j\in I_0} 1_{\{|w_j'|>t\}},$$

then

$$\begin{aligned}
& P(\hat{\pi}^{**} > \pi) \\
\le\ & P\left(\max_{t\in\mathbb{T}}\left\{(1-\pi)\left(p_0^{-1}\textstyle\sum_{j\in I_0} 1_{\{|w_j'|>t\}} - 2\bar{\Phi}(t)\right) - c_p^*\bar{\sigma}(t)\right\} > 0\right) \\
\le\ & P\left(\max_{t\in\mathbb{T}}\left\{p_0^{-1}\textstyle\sum_{j\in I_0} 1_{\{|w_j'|>t\}} - 2\bar{\Phi}(t) - c_{p_0}^*\bar{\sigma}(t)\right\} > 0\right) \\
\le\ & P\left(V_{p_0}^* > c_{p_0}^*\right),
\end{aligned}$$

where the second inequality follows since c_p^* is non-decreasing in p and $c_p^*(1-\pi)^{-1} > c_{p_0}^*$. However, Lemma 6.3 asserts $P(V_{p_0}^* > c_{p_0}^*) \to 0$. So,

$$P(\hat{\pi}^{**} > \pi) \le P(V_{p_0}^* > c_{p_0}^*) \to 0. \tag{6.5}$$

Next, we show that $\hat{\pi}^{**}$ is an asymptotic upper bound of $(1-\delta)\pi$ for any $\delta > 0$. Let $\mathsf{FP}_{w'}(t) = \sum_{j\in I_0} 1_{\{|w_j'|>t\}}$ and rewrite

$$\bar{\Phi}_p(t) = \frac{\pi}{s}\sum_{j\in I_1} 1_{\{|\mu_j + w_j'|>t\}} + \frac{1-\pi}{p_0}\mathsf{FP}_{w'}(t).$$

Since $\hat{\pi}^{**} > \bar{\Phi}_p(t) - 2\bar{\Phi}(t) - c_p^*\bar{\sigma}\,(t)$ for any $t \in \mathbb{T}$, then

$$\begin{aligned}
\frac{\hat{\pi}^{**}}{\pi} - 1 \quad > \quad & \left(s^{-1}\textstyle\sum_{j\in I_1} 1_{\{|\mu_j + w_j'|>t\}} - 1\right) - 2\bar{\Phi}(t) \\
& + \frac{1-\pi}{\pi}\left(p_0^{-1}\mathsf{FP}_{w'}(t) - 2\bar{\Phi}(t)\right) - \frac{1}{\pi}c_p^*\bar{\sigma}(t)
\end{aligned} \tag{6.6}$$

for any any $t \in \mathbb{T}$. Now set t in the inequality (6.6) to be

$$t_\tau = \sqrt{2\tau\log p} \ \text{ with } \tau = \gamma^* + c/2, \tag{6.7}$$

where γ^* is defined in (2.11). We will show that each term on the RHS of (6.6) is $o_P(1)$.

Firstly, $c_p^* = O\left((s_{max}/n)^{1/4}\log p\right)$ set in Lemma 6.3 implies the last term at t_τ

$$\pi^{-1}c_p^*\bar{\sigma}(t_\tau) = O(p^{\eta-\tau/2}(s_{max}/n)^{1/4}\log p) = o(1).$$

The second term $2\bar{\Phi}(t_\tau) = O(p^{-\tau}/\sqrt{\log p}) = o(1)$.

Consider the third term at t_τ. Similar arguments for (2.12) can be applied to show $s^{-1}|\mathsf{FP}_{w'}(t_\tau) - 2p_0\bar{\Phi}(t_\tau)| = o_P(1)$. Then

$$\frac{1-\pi}{\pi}\left|p_0^{-1}\mathsf{FP}_{w'}(t) - 2\bar{\Phi}(t)\right| \le Cs^{-1}|\mathsf{FP}_{w'}(t_\tau) - 2p_0\bar{\Phi}(t_\tau)| = o_P(1).$$

For the first term of (6.6), let $A_1(t) = s^{-1}\sum_{j\in I_1} 1_{\{|\mu_j+w'_j|\le t\}}$. Then the first term is $s^{-1}\sum_{j\in I_1} 1_{\{|\mu_j+w'_j|>t_\tau\}} - 1 = A_1(t_\tau)$. The following lemma shows $A_1(t_\tau) = o_P(1)$, and its proof is provided in Section A.7.

Lemma 6.5. *Let* $A_1(t) = s^{-1}\sum_{j\in I_1} 1_{\{|\mu_j+w'_j|\le t\}}$ *and assume*

$$\mu_{min} \ge \sqrt{2(\gamma^*+c)\log p}.$$

Then $A_1(t_\tau) = o_P(1)$ *for* t_τ *in (6.7).*

Thus, we have shown

$$P(\hat{\pi}^{**}/\pi - 1 < -\delta) \to 0 \tag{6.8}$$

for any $\delta > 0$. Consequently, (6.4) follows from (6.5) and (6.8).

6.3. Proof of Theorem 3.1

Recall $\mathsf{FNP}(t) = 1 - s^{-1}R(t) + s^{-1}FP(-t)$ and $\widehat{\mathsf{FNP}}(t) = 1 - \hat{s}^{-1}R(t) + 2\hat{s}^{-1}p\Phi(-t)$ for $t \ge 0$. Recall the definition of $t^*(\epsilon)$ and simplify the notation by $t^* = t^*(\epsilon)$. We have the following Lemma 6.6, whose proof is provided in Section A.8.

Lemma 6.6. *Assume* $\mu_{min} \ge \sqrt{2(\gamma^*+c)\log p}$. *If* t^* *satisfies* $P(t^* \ge t_\tau) \to 1$ *for* t_τ *in (6.7), then*

$$|\widehat{\mathsf{FNP}}(t^*) - \mathsf{FNP}(t^*)| = o_P(1). \tag{6.9}$$

Now we aim to show $P(t^* \ge t_\tau) \to 1$. The proof of the following Lemma 6.7 is presented in Section A.9.

Lemma 6.7. *Assume* $\mu_{min} \ge \sqrt{2(\gamma^*+c)\log p}$. *Then, for* t_τ *in (6.7),*

$$\mathsf{FNP}(t_\tau) = o_P(1). \tag{6.10}$$

Note that a special case of (6.9) is $|\widehat{\mathsf{FNP}}(t_\tau) - \mathsf{FNP}(t_\tau)| = o_P(1)$, which holds when t^* is set to be t_τ. Then (6.10) implies $\widehat{\mathsf{FNP}}(t_\tau) = o_P(1)$, and $P(t^* \geq t_\tau) \to 1$ follows from the definition of t^*.

On the other hand, the definition of t^* implies $\widehat{\mathsf{FNP}}(t^*) < \epsilon$ almost surely, then Lemma 6.6 implies $P(\mathsf{FNP}(t^*) < \epsilon) \to 1$ as stated in (3.2).

Next, we show (3.3). Denote

$$\widehat{\mathsf{FNP}}_{\hat{\pi}}(t) = 1 - \frac{R(t) - 2p\Phi(-t)}{\hat{\pi}p} \quad \text{and} \quad \widehat{\mathsf{FNP}}_{\hat{\pi}^*}(t) = 1 - \frac{R(t) - 2p\Phi(-t)}{\hat{\pi}^*p}.$$

By the definition of $\hat{\pi}$ and $\hat{\pi}^*$ and $c_p = c_p^*$, it is easy to see that $\hat{\pi} \geq \hat{\pi}^*$ and, consequently,

$$\widehat{\mathsf{FNP}}_{\hat{\pi}}(t) \geq \widehat{\mathsf{FNP}}_{\hat{\pi}^*}(t)$$

for any $t > 0$. Denote

$$t^*_{\hat{\pi}} = \sup\left\{t : \widehat{\mathsf{FNP}}_{\hat{\pi}}(t) \leq \epsilon\right\} \quad \text{and} \quad t^*_{\hat{\pi}^*} = \sup\left\{t : \widehat{\mathsf{FNP}}_{\hat{\pi}^*}(t) \leq \epsilon\right\}.$$

Then $t^*_{\hat{\pi}} \leq t^*_{\hat{\pi}^*}$ almost surely.

Recall $\mathsf{FNP}(t^*_{\hat{\pi}^*}) < \epsilon$ with probability tending to 1 as stated in (3.2) and the fact that $\mathsf{FNP}(t)$ is non-decreasing in t, then $\mathsf{FNP}(t^*_{\hat{\pi}}) \leq \mathsf{FNP}(t^*_{\hat{\pi}^*}) < \epsilon$ with probability tending to 1. Therefore (3.3) holds.

Appendix A: Appendix

The notations we will use throughout the appendices are collected as follows. For a matrix $\mathbf{M}$, the q-norm $\|\mathbf{M}\|_q = \left(\sum_{i,j} |\mathbf{M}_{ij}|^q\right)^{1/q}$ for $q > 0$, ∞-norm $\|\mathbf{M}\|_\infty = \max_{i,j} |\mathbf{M}_{ij}|$, and $\|\mathbf{M}\|_{1,\infty}$ the maximum of the 1-norm of each row of $\mathbf{M}$. If $\mathbf{M}$ is symmetric, $\sigma_i(\mathbf{M})$ denotes the ith largest eigenvalue of $\mathbf{M}$.

A.1. Debiased Lasso

The matrix $\hat{\mathbf{\Theta}} \in \mathbb{R}^{p\times p}$ appearing in the debiased Lasso estimator for $\boldsymbol{\beta}$ in the main text is obtained as follows. Let $\mathbf{X}_{-j}$ denote the matrix obtained by removing the jth column of $\mathbf{X}$. For each $j = 1, \ldots, p$, let

$$\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left(n^{-1}\|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\gamma}\|_2^2 + 2\lambda_j\|\boldsymbol{\gamma}\|_1\right) \tag{A.1}$$

with components $\hat{\gamma}_{j,k}, k = 1, \ldots, p$ and $k \neq j$, and define

$$\hat{\tau}_j^2 = n^{-1}\left\|\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_j\right\|_2^2 + 2\lambda_j\left\|\hat{\boldsymbol{\gamma}}_j\right\|_1.$$

Then

$$\hat{\mathbf{\Theta}} = \operatorname{diag}\left(\hat{\tau}_1^{-2}, \cdots, \hat{\tau}_p^{-2}\right)\begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}.$$

Recall $\sqrt{n}(\hat{\mathbf{b}} - \boldsymbol{\beta}) = \mathbf{w} - \boldsymbol{\delta}$, where $\mathbf{w} \sim \mathcal{N}_p(0, \sigma^2\hat{\boldsymbol{\Omega}})$ conditional on $\mathbf{X}$. To quantify the magnitude of $\boldsymbol{\delta}$, we adopt and rephrase Theorem 3.13 of [21] for unknown $\boldsymbol{\Sigma}$ as follows. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $s_j = |\{k \neq j : \boldsymbol{\Theta}_{jk} \neq 0\}|$ and $s_{max} = \max_{1\leq j\leq p} s_j$.

A1) Gaussian random design: the rows of $\mathbf{X}$ are i.i.d. $\mathcal{N}_p(0, \boldsymbol{\Sigma})$ for which $\boldsymbol{\Sigma}$ satisfies:

- **A1a)** $\max_{1\leq j\leq p} \boldsymbol{\Sigma}_{jj} \leq 1$.
- **A1b)** $0 < C_{min} \leq \sigma_1(\boldsymbol{\Sigma}) \leq \sigma_p(\boldsymbol{\Sigma}) \leq C_{max} < \infty$ for constants C_{min} and C_{max}.
- **A1c)** $\rho(\Sigma, C_0 s) \leq \rho$ for some constant $\rho > 0$, where $C_0 = 32C_{max}C_{min}^{-1} + 1$,
$$\rho(\mathbf{A}, k) = \max_{T\subseteq[p],|T|\leq k} \left\| (\mathbf{A}_{T,T})^{-1} \right\|_{1,\infty}$$
for a square matrix $\mathbf{A}$, $[p] = \{1, ..., p\}$, $\mathbf{A}_{T,T}$ is a sub-matrix formed by taking entries of $\mathbf{A}$ whose row and column indices respectively form the same subset T.

A2) Tuning parameters: for the Lasso in (2.1), $\lambda = 8\sigma\sqrt{n^{-1}\log p}$; for nodewise regression in (A.1), $\lambda_j = \tilde{\kappa}\sqrt{n^{-1}\log p}, j = 1, \ldots, p$ for a suitably large universal constant $\tilde{\kappa}$.

A3) Sparsities of $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$: $s = o\left(n/(\log p)^2\right)$, $\max\{s, s_{max}\} = o(n/\log p)$, $\min\{s, s_{max}\} = o(\sqrt{n}/\log p)$.

Lemma A.1. *Assume A1) and A2). Then there exist positive constants c and c' depending only on C_{min}, C_{max} and $\tilde{\kappa}$ such that, for $\max\{s, s_{max}\} < cn/\log p$, the probability that*

$$\|\boldsymbol{\delta}\|_\infty \leq c'\rho\sigma\sqrt{\frac{s}{n}}\log p + c'\sigma\min\{s, s_{max}\}\frac{\log p}{\sqrt{n}} \tag{A.2}$$

is at least $1 - 2pe^{-16^{-1}ns^{-1}C_{min}} - pe^{-cn} - 6p^{-2}$. Further, assume A3), then $\|\boldsymbol{\delta}\|_\infty = o_P(1)$.

Note that the above result relaxed the ultra-sparse condition $s = o(\sqrt{n}/\log p)$ in [34] to $s = o\left(n/(\log p)^2\right)$ as shown in A3).

Recall the standardized debiased Lasso estimate $z_j = \sqrt{n}\hat{b}_j\sigma^{-1}\hat{\boldsymbol{\Omega}}_{jj}^{-1/2}$ for $1 \leq j \leq p$. Namely, $z_j = \mu_j + w'_j - \delta'_j$, $w'_j = \frac{w_j}{\sigma\sqrt{\hat{\boldsymbol{\Omega}}_{jj}}} \sim \mathcal{N}(0, 1)$, $\delta'_j = \frac{\delta_j}{\sigma\sqrt{\hat{\boldsymbol{\Omega}}_{jj}}}$, $\mu_j = \frac{\sqrt{n}\beta_j}{\sigma\sqrt{\hat{\boldsymbol{\Omega}}_{jj}}}$ for each j. The $\sigma\hat{\boldsymbol{\Omega}}_{jj}^{1/2}$'s are refereed to as standardizers. Let $\boldsymbol{\delta}' = \left(\delta'_1, \ldots, \delta'_p\right)^T$. We quote from [22] some results on $\sigma\hat{\boldsymbol{\Omega}}_{jj}^{1/2}$ for $1 \leq j \leq p$ and the $\|\cdot\|_1$-norms of the covariance matrices for $\mathbf{w} = (w_1, \ldots, w_p)^T$ and $\mathbf{w}' = \left(w'_1, \ldots, w'_p\right)^T$.

Lemma A.2. *Assume A2) and $s_{max} = o(n/\log p)$. Then $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}\|_\infty = o_P(1)$. If further A1b) holds, then $\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty = O_P(\lambda_1)$, both $\min_{1\leq j\leq p}\hat{\boldsymbol{\Omega}}_{jj}$ and*

$\max_{1\le j\le p}\hat{\mathbf{\Omega}}_{jj}$ *are uniformly bounded (in* p*) away from* 0 *and* ∞ *with probability tending to* 1*, and* $\left\|\boldsymbol{\delta}'\right\|_\infty \le (\sigma\sqrt{C_{min}})^{-1}\left\|\boldsymbol{\delta}\right\|_\infty$ *with probability tending to* 1.

Lemma A.3. *Let* $\hat{\mathbf{K}}$ *be the correlation matrix of* $\mathbf{w}$*. Assume A1) and A2). Then*

$$p^{-2}\|\sigma^2\hat{\mathbf{\Omega}}\|_1 = O_P\left(\lambda_1\sqrt{s_{max}}\right) \quad \text{and} \quad \|\hat{\mathbf{K}}\|_1 = O(\sigma^2\|\hat{\mathbf{\Omega}}\|_1). \tag{A.3}$$

A.2. Hermite polynomials and Mehler expansion

The following is quoted from [22]. Let $\phi(x) = (2\pi)^{-1/2}\exp\left(-x^2/2\right)$ and

$$f_\rho(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2+y^2-2\rho xy}{2\left(1-\rho^2\right)}\right)$$

for $\rho\in(-1,1)$. For a nonnegative integer k, let $H_k(x) = (-1)^k\frac{1}{\phi(x)}\frac{d^k}{dx^k}\phi(x)$ be the kth Hermite polynomial; see [16] for such a definition. Then Mehler's expansion [27] gives

$$f_\rho(x,y) = \left(1+\sum\nolimits_{k=1}^{\infty}\frac{\rho^k}{k!}H_k(x)H_k(y)\right)\phi(x)\phi(y). \tag{A.4}$$

Further, Lemma 3.1 of [10] asserts

$$\left|e^{-y^2/2}H_k(y)\right| \le C_0\sqrt{k!}k^{-1/12}e^{-y^2/4} \quad \text{for any } y\in\mathbb{R} \tag{A.5}$$

for some constant $C_0 > 0$.

A.3. Proof of Lemma 6.1

By assumption A3), $d_p = o(1)$, $s \ll n/\log p$, and $n \gg \log p$. Then Lemma A.1 implies

$$P(\|\boldsymbol{\delta}\|_\infty \ge d_p) \le 2pe^{-c_* n/s} + pe^{-Cn} + 6p^{-2} \le Cp^{-2}$$

where $c_* = C_{\min}/16$. By Lemma A.2 and the definition of d_p, we have

$$P(\|\boldsymbol{\delta}'\|_\infty \ge d_p) \le Cp^{-2}.$$

Now consider $P(|w_j' - \delta_j'| > t_\xi)$, which is bounded as follows.

$$P(|w_j'| > t_\xi + |\delta_j'|) \le P(|w_j' - \delta_j'| > t_\xi) \le P(|w_j'| > t_\xi - |\delta_j'|).$$

The rightmost term

$$\begin{aligned} P(|w_j'| > t_\xi - |\delta_j'|) &\le P(|w_j'| > t_\xi - |\delta_j'|, \|\boldsymbol{\delta}'\|_\infty \le d_p) + P(\|\boldsymbol{\delta}'\|_\infty > d_p) \\ &\le P(|w_j'| > t_\xi - d_p) + Cp^{-2}. \end{aligned}$$

On the other hand, the leftmost term

$$\begin{aligned} P(|w'_j| > t_\xi + |\delta'_j|) &\geq P(|w'_j| > t_\xi + |\delta'_j|, \|\boldsymbol{\delta}'\|_\infty \leq d_p) \\ &\geq P(|w'_j| > t_\xi + d_p, \|\boldsymbol{\delta}'\|_\infty \leq d_p) \\ &= P(|w'_j| > t_\xi + d_p) - P(|w'_j| > t_\xi + d_p, \|\boldsymbol{\delta}'\|_\infty > d_p) \\ &\geq P(|w'_j| > t_\xi + d_p) - P(\|\boldsymbol{\delta}'\|_\infty > d_p) \\ &\geq P(|w'_j| > t_\xi + d_p) - Cp^{-2}. \end{aligned}$$

Summing up the above gives

$$\left|P(|w'_j - \delta'_j| > t_\xi) - P(|w'_j| > t_\xi)\right| \leq C\phi(t_\xi)d_p + Cp^{-2},$$

and the claim in Lemma 6.1 follows.

A.4. Proof of Lemma 6.2

For $i \neq j$, let ρ_{ij} be the correlation between w'_i and w'_j and

$$C_{ij,\xi} = \mathsf{Cov}\left(1_{\{|w'_i| \leq t_\xi\}}, 1_{\{|w'_j| \leq t_\xi\}}\right).$$

Then, by Lemma A.2, ρ_{ij} is also the correlation between w_i and w_j. Further,

$$\mathsf{Var}\left(\sum\nolimits_{j=1}^{p} 1_{\{|w'_j| > t_\xi\}}\right) \leq \sum_{j=1}^{p} \mathsf{Var}\left(1_{\{|w'_j| \leq t_\xi\}}\right) + \sum_{i \neq j} C_{ij,\xi}. \tag{A.6}$$

By Mill's ratio,

$$\sum_{j=1}^{p} \mathsf{Var}\left(1_{\{|w'_j| \leq t_\xi\}}\right) \leq 2p\Phi\left(-t_\xi\right)\left(1 - 2\Phi\left(-t_\xi\right)\right) = O(p^{1-\xi}). \tag{A.7}$$

It is left to bound $\sum_{i \neq j} C_{ij,\xi}$ in (A.6).

Define $c_{1,\xi} = -t_\xi$ and $c_{2,\xi} = t_\xi$. Fix a pair of (i, j) such that $i \neq j$ and $|\rho_{ij}| \neq 1$. Now we will use the results in Section A.2. Since $C_{ij,\xi}$ is finite and the series in Mehler's expansion in (A.4) as a trivariate function of (x, y, ρ) is uniformly convergent on each compact set of $\mathbb{R} \times \mathbb{R} \times (-1, 1)$ as justified by [36], we can interchange the order of the summation and integration and obtain

$$\begin{aligned} C_{ij,\xi} &= \int_{c_{1,\xi}}^{c_{2,\xi}} \int_{c_{1,\xi}}^{c_{2,\xi}} f_{\rho_{ij}}(x, y)\, dxdy - \int_{c_{1,\xi}}^{c_{2,\xi}} \phi(x)dx \int_{c_{1,\xi}}^{c_{2,\xi}} \phi(y)dy \\ &= \sum_{k=1}^{\infty} \frac{\rho_{ij}^k}{k!} \int_{c_{1,\xi}}^{c_{2,\xi}} H_k(x)\phi(x)dx \int_{c_{1,\xi}}^{c_{2,\xi}} H_k(y)\phi(y)dy. \end{aligned}$$

Since $H_{k-1}(x)\,\phi(x) = \int_{-\infty}^{x} H_k(y)\,\phi(y)\,dy$ for $x \in \mathbb{R}$, then

$$C_{ij,\xi} = \sum_{k=1}^{\infty} \frac{\rho_{ij}^k}{k!} \left[H_{k-1}(c_{2,\xi})\phi(c_{2,\xi}) - H_{k-1}(c_{1,\xi})\phi(c_{1,\xi})\right]^2$$

$$\leq \quad 2\sum_{k=1}^{\infty}\frac{|\rho_{ij}|^k}{k!}\left\{\left[H_{k-1}\left(c_{2,\xi}\right)\phi\left(c_{2,\xi}\right)\right]^2+\left[H_{k-1}\left(c_{1,\xi}\right)\phi\left(c_{1,\xi}\right)\right]^2\right\}.$$

Inequality (A.5) implies, for some finite constant $C_0>0$,

$$\left[H_{k-1}\left(c_{2,\xi}\right)\phi\left(c_{2,\xi}\right)\right]^2+\left[H_{k-1}\left(c_{1,\xi}\right)\phi\left(c_{1,\xi}\right)\right]^2\leq C_0^2(k-1)!(k-1)^{-1/6}e^{-t_\xi^2/2}.$$

Therefore,

$$\begin{aligned}\left|\sum\nolimits_{i\neq j}C_{ij,\xi}\right| &\leq C\sum_{1\leq i<j\leq p}|\rho_{ij}|\sum_{k=1}^{\infty}k^{-7/6}\left|\rho_{ij}\right|^{k-1}e^{-t_\xi^2/2}\\ &\leq Cp^{-\xi}\sum_{1\leq i<j\leq p}|\rho_{ij}|=O(p^{-\xi}\|\hat{\mathbf{K}}\|_1).\end{aligned}\tag{A.8}$$

Combining (A.6) with (A.7) and (A.8) gives

$$\begin{aligned}\mathsf{Var}\left(\sum\nolimits_{j=1}^{p}1_{\{|w_j'|>t_\xi\}}\right) &= O(p^{1-\xi})+O(p^{-\xi}\|\hat{\mathbf{K}}\|_1)\\ &= O(p^{1-\xi})+O(p^{2-\xi}\lambda_1\sqrt{s_{max}}),\end{aligned}$$

where the last inequality follows from Lemma A.3, i.e., $\|\hat{\mathbf{K}}\|_1=O_P(p^2\lambda_1\sqrt{s_{max}})$.

A.5. Proof of Lemma 6.3

Recall

$$\bar{\sigma}\left(t\right)=\sqrt{2\bar{\Phi}(t)\left(1-2\bar{\Phi}(t)\right)},\quad \mathsf{H}(t)=\left(\bar{\sigma}\left(t\right)\right)^{-1}\left(p^{-1}\sum\nolimits_{j=1}^{p}1_{\{|w_j'|>t\}}-2\bar{\Phi}(t)\right).$$

Then $E(\mathsf{H}(t))=0$ since $w_j'\sim\mathcal{N}_1(0,1)$ for all j.

For any $t_\xi=\sqrt{2\xi\log p}$ such that $\lim_{p\to\infty}t_\xi=\infty$, Lemma 6.2 implies

$$\begin{aligned}\mathsf{Var}\left(\mathsf{HC}(t_\xi)\right) &= p^{-2}\bar{\sigma}_p^{-2}\left(t_\xi\right)\mathsf{Var}\left(\sum\nolimits_{j=1}^{p}1_{\{|w_j'|>t_\xi\}}\right)\\ &\leq Cp^{\xi-2}\sqrt{\log p}\left(p^{1-\xi}+p^{2-\xi}\sqrt{\log p}\sqrt{s_{max}/n}\right)\\ &= O\left(\sqrt{s_{max}/n}\log p\right).\end{aligned}\tag{A.9}$$

Let

$$\mathbb{T}=\left[\sqrt{\tau_0\log p},\sqrt{\tau_1\log p}\right]\cap\mathbb{N}\tag{A.10}$$

for which $0<\tau_0<\tau_1$. So, each $t\in\mathbb{T}$ can be written as $t=t_\xi=\sqrt{2\xi\log p}$ for some $\xi>0$ and $\lim_{p\to\infty}t_\xi=\infty$. Recall $V_p^*=\max\{\mathsf{H}(t):t\in\mathbb{T}\}$. Therefore, (A.9) implies

$$\begin{aligned}P(V_p^*>c_p^*) &\leq C\left(c_p^*\right)^{-2}\sqrt{\log p}\max_{t\in\mathbb{T}}\mathsf{Var}\left(\mathsf{H}(t)\right)\\ &\leq C\left(c_p^*\right)^{-2}\cdot(\log p)^{3/2}\cdot\sqrt{s_{max}/n}.\end{aligned}$$

However, $c_p^*=O\left((s_{max}/n)^{1/4}\log p\right)$. Thus, $P(V_p^*>c_p^*)=o\left(1\right)$ as desired.

A.6. Proof of Lemma 6.4

Since $\max_{t\in\mathbb{T}}\left(1-2\bar{\Phi}(t)\right)\geq 4^{-1}$ for all p sufficiently large. It suffices to show

$$\pi^{-1}\left|F_p(t)-\bar{\Phi}_p(t)\right|=o_P(1) \tag{A.11}$$

for $t=t_\xi=\sqrt{2\xi\log p}$ with $\xi>\gamma^*$. Perform the decomposition

$$\begin{aligned}\left|F_p(t)-\bar{\Phi}_p(t)\right| &\leq \left|F_p(t)-E(F_p(t))\right| \\ &+ \left|\bar{\Phi}_p(t)-E(\bar{\Phi}_p(t))\right|+\left|E(F_p(t))-E(\bar{\Phi}_p(t))\right|.\end{aligned}$$

Similar arguments for (6.1) can be applied to show

$$\pi^{-1}\left|F_p(t)-E(F_p(t))\right|=o_p(1)=\pi^{-1}\left|\bar{\Phi}_p(t)-E(\bar{\Phi}_p(t))\right|,$$

and similar arguments for (6.2) imply

$$\pi^{-1}\left|E(F_p(t))-E(\bar{\Phi}_p(t))\right|=o(1).$$

Summing up the above gives (A.11).

A.7. Proof of Lemma 6.5

We will show $A_1(t_\tau)=o_P(1)$. Fix a constant $a>0$,

$$P\left(A_1(t_\tau)>a\right)\leq\frac{1}{as}\sum_{j\in I_1}P\left(|\mu_j+w_j'|\leq t_\tau\right)\leq\frac{1}{a}\max_{j\in I_1}P\left(|\mu_j+w_j'|\leq t_\tau\right) \tag{A.12}$$

and for each $j\in I_1$

$$P\left(|\mu_j+w_j'|\leq t_\tau\right)=1-\bar{\Phi}\left(t_\tau-\mu_j\right)-\Phi\left(-t_\tau-\mu_j\right), \tag{A.13}$$

We only need to uniformly bound the RHS of (A.13).

Recall $\mu_{min}=\min_{j\in I_1}\sqrt{n}\left|\beta_j\right|\sigma^{-1}\sqrt{\mathbf{\Sigma}_{jj}}$ and $\mu_{min}\geq\sqrt{2(\gamma^*+c)\log p}$. Let

$$\mu_{min}=\sqrt{2r\log p},$$

then $r\geq\tau+c/2$. Further, by Lemma A.2, the ratio $\mu_{min}/\min_{j\in I_1}|\mu_j|$ is uniformly bounded (in p) away from 0 and ∞. Then, two cases happen for each $j\in I_1$: (i) both $t_\tau-\mu_j\to-\infty$ and $-t_\tau-\mu_j\to-\infty$ when $\mu_j>0$; (b) both $t_\tau-\mu_j\to+\infty$ and $-t_\tau-\mu_j\to+\infty$ when $\mu_j<0$. However, in either case,

$$\min_{j\in I_1}\min\left\{\left|t_\tau-\mu_j\right|,\left|-t_\tau-\mu_j\right|\right\}\geq\sqrt{2\tilde{c}\log p},$$

where $\tilde{c}=2^{-1}\left(\sqrt{2\tau+c}-\sqrt{2\tau}\right)^2>0$. Therefore,

$$\max_{j\in I_1}P\left(|\mu_j+w_j'|\leq t_\tau\right)\leq 4\bar{\Phi}\left(\sqrt{2\tilde{c}\log p}\right)=O\left(p^{-\tilde{c}}\right). \tag{A.14}$$

Combining (A.14) with (A.12) gives

$$P\left(A_1(t_\tau)>a\right)\leq a^{-1}O\left(p^{-\tilde{c}}\right)=o(1),$$

which is the desired claim on $A_1(t_\tau)$.

A.8. Proof of Lemma 6.6

Recall $\widehat{\mathsf{FNP}}(t) = 1 - \hat{s}^{-1}(R(t) - 2(p-\hat{s})\Phi(-t))$. We only need to show

$$\begin{aligned} & \left|s^{-1}(R(t^*) - \mathsf{FP}(t^*)) - \hat{s}^{-1}(R(t^*) - 2(p-\hat{s})\Phi(-t^*))\right| \\ \leq\ & \left|\mathsf{TP}(t^*)\left(s^{-1} - \hat{s}^{-1}\right)\right| + \left|\hat{s}^{-1}(\mathsf{FP}(t^*) - 2p_0\Phi(-t^*)\right| \\ +\ & \left|2\hat{s}^{-1}\Phi(-t^*)(p - \hat{s} - p_0)\right| = o_P(1). \end{aligned} \tag{A.15}$$

Since $\mu_{min} \geq \sqrt{2(\gamma^* + c)\log p}$, then Theorem 2.2 implies

$$P\left(1 - \delta \leq \hat{s}s^{-1} \leq 1\right) \to 1 \tag{A.16}$$

for any $\delta > 0$. Let $\delta' = \frac{\delta}{1-\delta}$. Then, (A.16) is equivalent to

$$P\left(0 \leq \hat{s}^{-1} - s^{-1} \leq \delta' s^{-1}\right) \to 1.$$

Pick a $\delta > 0$ such that $\delta < \frac{a}{1+a}$. Then $\delta < 1$ and $\delta' s^{-1}\mathsf{TP}(t^*) < a$ almost surely. Therefore,

$$\begin{aligned} & P\left(\left|\mathsf{TP}(t^*)\left(s^{-1} - \hat{s}^{-1}\right)\right| > a\right) \\ \leq\ & P\left(\left|\mathsf{TP}(t^*)\left(s^{-1} - \hat{s}^{-1}\right)\right| > a, \left|s^{-1} - \hat{s}^{-1}\right| \leq \delta' s^{-1}\right) \\ +\ & P\left(\left|s^{-1} - \hat{s}^{-1}\right| > \delta' s^{-1}\right) \\ \leq\ & P\left(\mathsf{TP}(t^*)\,\delta' s^{-1} \geq a\right) + o(1) \\ =\ & 0 + o(1), \end{aligned}$$

i.e., the first term in (A.15) $= o_P(1)$. The remaining two terms in (A.15) are also of $o_P(1)$ by (A.16) and Theorem 2.1. This concludes the proof.

A.9. Proof of Lemma 6.7

Recall $A_1(t) = s^{-1}\sum_{j\in I_1} 1_{\{|\mu_j + w_j'| \leq t\}}$. Lemma 6.5 implies $A_1(t_\tau) = o_P(1)$, given $\mu_{min} \geq \sqrt{2(\gamma^* + c)\log p}$. Now we show $\mathsf{FNP}(t_\tau) = o_P(1)$. Clearly,

$$P(\mathsf{FNP}(t_\tau) > a) \leq \frac{1}{a}\max_{j\in I_1} P\left(|\mu_j + w_j' + \delta_j'| \leq t_\tau\right).$$

However, $\max_{1\leq i\leq p}|\delta_j'| = o_P(1)$ and $w_j' \sim \mathcal{N}_1(0,1)$ for each j together imply

$$\max_{j\in I_1}\left|P\left(|\mu_j + w_j' + \delta_j'| \leq t_\tau\right) - P\left(|\mu_j + w_j'| \leq t_\tau\right)\right| = o(1).$$

Combining the above with (A.14) gives $\max_{j\in I_1} P\left(|\mu_j + w_j' + \delta_j'| \leq t_\tau\right) = o(1)$, and $\mathsf{FNP}(t_\tau) = o_P(1)$ holds.

References

[1] Arias-Castro, E., E. J. Cand's, and Y. Plan (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Statist. 39*(5), 2533–2556. MR2906877

[2] Arias-Castro, E. and A. Ying (2019). Detection of sparse mixtures: Higher criticism and scan statistic. *Electron. J. Statist. 13*(1), 208–230. MR3899951

[3] Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist. 43*(5), 2055–2085. MR3375876

[4] Bogdan, M., E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). Slope—adaptive variable selection via convex optimization. *Ann. Appl. Stat. 9*(3), 1103–1140. MR3418717

[5] Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data Methods, Theory and Applications*. Springer. MR2807761

[6] Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist. 45*(2), 615–646. MR3650395

[7] Cai, T. T., X. J. Jeng, and J. Jin (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc. Ser. B 73*(5), 629–662. MR2867452

[8] Cai, T. T., W. Liu, and H. H. Zhou (2016, 04). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist. 44*(2), 455–488. MR3476606

[9] Candes, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc. Ser. B 80*(3), 551–577. MR3798878

[10] Chen, X. and R. Doerge (2016). A strong law of large numbers related to multiple testing normal means. arXiv:1410.4276v3.

[11] Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statist. Sci. 30*(4), 533–558. MR3432840

[12] Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 962–994. MR2065195

[13] Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc. 107*(499), 1019–1035. MR3010887

[14] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. Ser. B 70*(5), 849–911. MR2530322

[15] Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101–148. MR2640659

[16] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Volume II. Wiley, NewYork, NY. MR0270403

[17] G'sell, M., S. Wager, A. Chouldechova, and R. Tibshirani (2016). Sequential selection procedures and false discovery rate control. *J. R. Statist. Soc. Ser. B 78*(2), 423–444. MR3454203

[18] Hall, P. and J. Jin (2008). Properties of higher criticism under strong dependence. *Ann. Statist. 36*(1), 381–402. MR2387976

[19] Javanmard, A. and A. Montanari (2013). Model selection for high-dimensional regression under the generalized irrepresentability condition. In *Advances in neural information processing systems*, pp. 3012–3020. MR3265038

[20] Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res. 15*(1), 2869–2909. MR3277152

[21] Javanmard, A. and A. Montanari (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Statist. 46*(6A), 2593–2622. MR3851749

[22] Jeng, X. J. and X. Chen (2019). Predictor ranking and false discovery proportion control in high-dimensional regression. *J. Multivariate Anal. 171*, 163–175. MR3892907

[23] Ji, P. and J. Jin (2012). Ups delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist. 40*(1), 73–103. MR3013180

[24] Ji, P. and Z. Zhao (2014). Rate optimal multiple testing procedure in high-dimensional regression. arXiv:1404.2961.

[25] Liu, Z., V. J. Berrocal, A. J. Bartsch, and T. D. Johnson (2016). Presurgical fmri data analysis using a spatially adaptive conditionally autoregressive model. *Bayesian analysis (Online) 11*(2), 599. MR3472004

[26] Loring, D., K. Meador, J. D. Allison, J. Pillai, T. Lavin, G. P. Lee, A. Balan, and V. Dave (2002). Now you see it, now you don't: statistical and methodological considerations in fmri. *Epilepsy & Behavior 3*(6), 539–547.

[27] Mehler, G. F. (1866). Ueber die entwicklung einer funktion von beliebeg vielen variablen nach laplaceschen funktionen hoherer ordnung. *J. Reine Angew. Math. 66*, 161–176. MR1579340

[28] Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist. 34*(3), 1436–1462. MR2278363

[29] Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist. 34*(1), 373–393. MR2275246

[30] Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *J. of Machine Learning Technologies 2*(1), 37–63.

[31] Silva, M., A. See, W. Essayed, A. Golby, and Y. Tie (2018). Challenges and techniques for presurgical brain mapping with functional mri. *NeuroImage: Clinical 17*, 794–803.

[32] Su, W. and E. Candés (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist. 44*, 1038–1068. MR3485953

[33] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. Ser. B 58*(1), 267–288. MR1379242

[34] van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202. MR3224285

[35] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory 55*(5), 2183–2202. MR2729873

[36] Watson, G. N. (1933). Notes on generating functions of polynomials: (2) hermite polynomials. *J. Lond. Math. Soc. s1-8*(3), 194–199. MR1574123

[37] Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. Ser. B 76*(1), 217–242. MR3153940

[38] Zhang, X. and G. Cheng (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc. 112*(518), 757–768. MR3671768