The Annals of Statistics
2017, Vol. 45, No. 3, 1214–1253
DOI: 10.1214/16-AOS1482

WEAK SIGNAL IDENTIFICATION AND INFERENCE IN PENALIZED MODEL SELECTION

BY PEIBEI SHI[1,*] AND ANNIE QU[1,†]

University of Michigan and Univeristy of Illinois at Urbana-Champaign†*

Weak signal identification and inference are very important in the area of penalized model selection, yet they are underdeveloped and not well studied. Existing inference procedures for penalized estimators are mainly focused on strong signals. In this paper, we propose an identification procedure for weak signals in finite samples, and provide a transition phase in-between noise and strong signal strengths. We also introduce a new two-step inferential method to construct better confidence intervals for the identified weak signals. Our theory development assumes that variables are orthogonally designed. Both theory and numerical studies indicate that the proposed method leads to better confidence coverage for weak signals, compared with those using asymptotic inference. In addition, the proposed method outperforms the perturbation and bootstrap resampling approaches. We illustrate our method for HIV antiretroviral drug susceptibility data to identify genetic mutations associated with HIV drug resistance.

1. Introduction. Penalized model selection methods are developed to select variables and estimate coefficients simultaneously, which is extremely useful in variable selection if the dimension of predictors is large. Some most popular model selection methods include Lasso [26], SCAD [6], adaptive Lasso [33], MCP [31] and the truncated-L_1 penalty method [23]. Asymptotic properties have been established for desirable penalized estimators such as unbiasedness, sparsity and the oracle property. However, established asymptotic theory mainly targets strong-signal coefficient estimators. When signal strength is weak, existing penalized methods are more likely to shrink the coefficient estimator to be 0. For finite samples, the inference of the weak signals is still lacking in the current literature.

In general, identification and inference for weak signal coefficients play an important role in scientific discovery. A more extreme argument is that all useful signals are weak [3], where each individual weak signal might not contribute significantly to a model's prediction, but the weak signals combined together could have significant influence to predict a model. In addition, even though some variables do not have strong signal strength, they might still need to be included in the model by design or by scientific importance.

Received May 2015; revised February 2016.
[1]Supported by NSF Grants DMS-13-08227 and DMS-14-15308.
MSC2010 subject classifications. Primary 62F30, 62J07; secondary 62E15.
Key words and phrases. Model selection, weak signal, finite sample inference, adaptive Lasso.

The estimation of the distribution for the penalized estimator in finite samples is quite challenging when the true coefficients are small. Standard bootstrap methods are not applicable when the parameter is close to zero ([1] and [16]). Recently, Pötscher and Leeb [20] and Pötscher and Schneider [21, 22] show that the distribution of penalized estimators such as Lasso-type estimators are highly nonnormal in finite samples. They also indicate that the distribution of the penalized estimator relies on the true parameter and, therefore, is hard to estimate if the true information is unknown. Their findings confirm that even if a weak signal is selected in the model selection procedure, inference of weak-signal parameters in finite samples is not valid based on the asymptotic theory.

Studies on weak signal identification and inference are quite limited. Among these few studies, Jin, Zhang and Zhang [14] propose a graphlet screening method in high-dimensional variable selection, where all the useful features are assumed to be rare and weak. Their work mainly focuses on signal detection, but not on parameter inference. Zhang and Zhang [32] develop a projection approach to project a high-dimensional model to a low-dimensional problem and construct confidence intervals. However, their inference method is not for the penalized estimator. The most recent related work is by Minnier, Tian and Cai [19], where they propose a perturbation resampling method to draw inference for regularized estimators. However, their approach is more suitable for relatively strong signal rather than weak signal inference.

In this paper, we investigate finite sample behavior for weak signal inference. Mainly we propose an identification procedure for weak signals, and provide a weak signal interval in-between noise and strong signal strengths, where the weak signal's range is defined based on the signal's detectability under the penalized model selection framework. In addition, we propose a new two-step inferential method to construct better inference for the weak signals. In theory, we show that our two-step procedure guarantees that the confidence interval reaches an accurate coverage rate under regularity conditions. Our numerical studies also confirm that the proposed method leads to better confidence coverage for weak signals, compared to existing methods based on asymptotic inference, perturbation methods and bootstrap resampling approaches ([4] and [5]). Note that our method and theory are developed under the orthogonal design assumption.

Our paper is organized as follows. In Section 2, we introduce the general framework for penalized model selection. In Section 3, we propose weak signal definition and identification. The two-step inference procedure and its theoretical property for finite samples are illustrated in Section 4. In Section 5, we evaluate finite sample performance of the proposed method and compare it to other available approaches, and apply these methods for an HIV drug resistance data example. The last section provides a brief summary and discussion.

2. Penalized least square method.

We consider a linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is a $n \times p$ design matrix with $p < n$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n})$. Throughout the entire paper, we assume that all covariates are standardized with $\mathbf{X_j^T X_j} = n$ for $j = 1, \ldots, p$.

The penalized least square estimator is the minimizer of the penalized least square function:

$$L(\boldsymbol{\theta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\theta_j|), \tag{2.1}$$

where $\|\cdot\|$ is the Euclidean norm and $p_\lambda(\cdot)$ is a penalty function controlled by a tuning parameter λ. For example, the adaptive Lasso penalty proposed by [33] has the following form:

$$p_{\mathrm{ALASSO},\lambda}(\theta) = \lambda \frac{|\theta|}{|\hat{\theta}^{\mathrm{LS}}|},$$

where θ is any component of $\boldsymbol{\theta}$, and $\hat{\theta}^{\mathrm{LS}}$ is the least-square estimator of θ. The penalized least square estimator $\hat{\boldsymbol{\theta}}$ is obtained by minimizing (2.1) given a λ, where the best λ can be selected through k-fold cross validation, generalized cross-validation (GCV) [6] or the Bayesian information criterion (BIC) [28].

In this paper, we mainly focus on the adaptive Lasso estimator as an illustration for penalized estimators. Our method, however, is also applicable for other appropriate penalized estimators. Under the orthogonal designed matrix $\mathbf{X}$, the adaptive Lasso estimator has an explicit expression:

$$\hat{\theta}_{\mathrm{ALASSO}} = \left(|\hat{\theta}^{\mathrm{LS}}| - \frac{\lambda}{|\hat{\theta}^{\mathrm{LS}}|}\right)_+ \mathrm{sgn}(\hat{\theta}^{\mathrm{LS}}). \tag{2.2}$$

Assume $\mathcal{A} = \{j : \theta_j \neq 0\}$, $\mathcal{A}^c = \{j : \theta_j = 0\}$, $\mathcal{A}_n = \{j : \hat{\theta}_j \neq 0\}$, $\mathcal{A}_n^c = \{j : \hat{\theta}_j = 0\}$, where $\hat{\boldsymbol{\theta}}$ denotes the penalized estimation. If the tuning parameter λ_n satisfies conditions of $\sqrt{n}\lambda_n \to 0, n\lambda_n \to \infty$, the adaptive Lasso estimator has oracle properties such that $\mathcal{A}_n = \mathcal{A}$ with probability tending to 1 as n goes to infinity. This indicates that the adaptive Lasso is able to successfully classify model parameters into two groups, $\mathcal{A}$ and $\mathcal{A}^c$, if the sample size is large enough. An underlying sufficient condition for such perfect separation asymptotically is that all nonzero signals should be greater than a uniform signal strength, which is proportional to $\sigma/\sqrt{n}$ [6]. In other words, signal strength within a noise level $C\sigma/\sqrt{n}$ should not be detected through a regularized procedure. However, due to an uncertain scale for the constant C, the absolute boundary between noise and signal level is unclear.

Therefore, it is important to define a more informative signal magnitude which is applicable in finite samples. This motivates us to define a transition phase in-between noise level and strong-signal level. In the following, we propose three phases corresponding to noise, weak signal and strong signal, where three different levels are defined based on low, moderate and high detectability of signals, respectively.

3. Weak signal definition and identification.

3.1. *Weak signal definition.* Suppose a model contains both strong and weak signals. Without loss of generality, the parameter vector $\boldsymbol{\theta}$ consists of three components: $\boldsymbol{\theta} = (\boldsymbol{\Theta}^{(S)}, \boldsymbol{\Theta}^{(W)}, \boldsymbol{\Theta}^{(N)})^T$, where $\boldsymbol{\Theta}^{(S)}$, $\boldsymbol{\Theta}^{(W)}$ and $\boldsymbol{\Theta}^{(N)}$ represent strong-signal, weak-signal and noise coefficients. We introduce a degree of detectability to measure different signal strength levels as follows.

For any given penalized model selection method, we define P_d as a probability of selecting an individual variable. For example, for the Lasso approach in (2.2), P_d has an explicit form of θ function given n, σ and λ:

$$P_d(\theta) = P(\hat{\theta}_{\mathrm{ALASSO}} \neq 0 | \theta) = \Phi\left(\frac{\theta - \sqrt{\lambda}}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-\theta - \sqrt{\lambda}}{\sigma/\sqrt{n}}\right). \tag{3.1}$$

Clearly, $P_d(\theta)$ is a symmetric function, and $P_d(\theta) \to 0$ for $\theta = 0$, $P_d(\theta) \to 1$ for any $\theta \neq 0$, as $n \to \infty$. For finite samples, $P_d(\theta)$ is an increasing function of $|\theta|$, and measures the detectability of a signal coefficient, which serves as a good indicator of signal strength such that a stronger signal leads to a larger P_d and vice versa.

In the following, we define a strong signal if P_d is close to 1, a noise variable if P_d is close to 0, and a weak signal if a signal strength is in-between strong and noise levels. Specifically, suppose there are two threshold probabilities, γ^s and γ^w derived from P_d, the three signal-strength levels are defined as

$$\begin{cases} \theta \in \boldsymbol{\Theta}^{(S)}, & \text{if } P_d > \gamma^s, \\ \theta \in \boldsymbol{\Theta}^{(W)}, & \text{if } \gamma^w < P_d \leq \gamma^s, \\ \theta \in \boldsymbol{\Theta}^{(N)}, & \text{if } P_d \leq \gamma^w, \end{cases} \tag{3.2}$$

where $\tau_0 \preceq \gamma^w < \gamma^s \preceq 1$, and $\tau_0 = \min_\theta P_d(\theta) = 2\Phi(-\frac{\sqrt{n\lambda}}{\sigma})$ can be viewed as a false-positive rate of model selection. Theoretically, $\tau_0 \to 0$ when $n \to \infty$ for consistent model selection. In finite samples, τ_0 does not need to be 0, but close to 0.

To see the connection between signal detectability P_d and signal strength, we let ν^γ be a positive solution of $P_d = \gamma$ in (3.1):

$$\gamma = \Phi\left(\frac{\nu^\gamma - \sqrt{\lambda}}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-\nu^\gamma - \sqrt{\lambda}}{\sigma/\sqrt{n}}\right). \tag{3.3}$$

It can be shown that ν^γ is an increasing function of γ. In addition, if the two positive threshold values ν^s and ν^w are solutions of equation (3.3) corresponding to $\gamma = \gamma^s$ and γ^w, then the definition in (3.2) is equivalent to

$$\begin{cases} \theta \in \boldsymbol{\Theta}^{(S)}, & \text{if } |\theta| > \nu^s, \\ \theta \in \boldsymbol{\Theta}^{(W)}, & \text{if } \nu^w < |\theta| \leq \nu^s, \\ \theta \in \boldsymbol{\Theta}^{(N)}, & \text{if } |\theta| \leq \nu^w. \end{cases} \tag{3.4}$$

Figure 1 also illustrates a connection between definition (3.2) and definition (3.4).

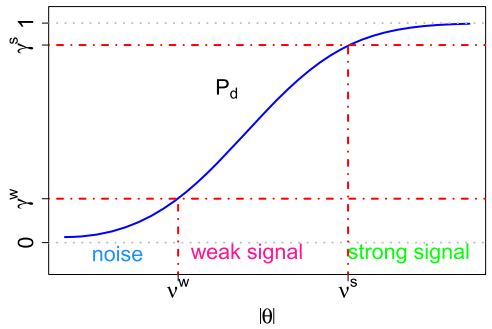

FIG. 1. *Define signal level based on* P_d.

The following lemma provides selections of γ^s and γ^w, which is useful to differentiate weak signals from noise variables. Lemma 1 also infers the order of weak signals, given both γ^s and γ^w are bounded away from 0 and 1.

LEMMA 1 (Selection of γ^s and γ^w). *If assumptions of* $\sqrt{n}\lambda_n \to 0, n\lambda_n \to \infty$ *are satisfied, and if the threshold values of detectability* γ^w *and* γ^s *corresponding to the lower bounds of weak and strong signals satisfy*

$$\max\left\{\epsilon, 2\Phi\left(-\frac{\sqrt{n\lambda_n}}{\sigma}\right)\right\} < \gamma^w < \gamma^s < 1-\epsilon,$$

where ϵ *is a small positive value; then for any* γ *in the weak signal range* (γ^w, γ^s), *we have* $\nu^\gamma/\sqrt{\lambda_n} \to 1$.

Although Lemma 1 implies that ν within the weak signal range converges to zero asymptotically, the weak signal and noise variables have different orders. Specifically, Lemma 1 indicates that if the regularity condition $n\lambda_n \to \infty$ is satisfied, then a weak signal goes to zero more slowly than a noise variable. This is due to the fact that the weak signal has the same order as $\sqrt{\lambda_n}$, which goes to zero more slowly than the order of noise level $n^{-1/2}$. To simplify notation, the tuning parameter λ_n is denoted as λ throughout the rest of the paper.

The definitions in (3.2) and (3.4) are particularly meaningful in finite samples since ν^γ depends on n, λ, σ and γ. That is, the weak signals are relative and depend on the sample size, the signal to noise ratio and the tuning parameter selection. In other words, weak signals $\boldsymbol{\Theta}^{(W)}$ might be asymptotically trivial since the three levels automatically degenerate into two levels: zero and nonzero coefficients. However, weak signals should not be ignored in finite samples and serve as a transition phase between noise variables $\boldsymbol{\Theta}^{(N)}$ and strong signals $\boldsymbol{\Theta}^{(S)}$.

3.2. *Weak signal identification.* In this section, we discuss how to identify weak signals more specifically. We propose a two-step procedure to recover possible weak signals which might be missed in a standard model selection procedure, and distinguish weak signals from strong signals.

The key component of the proposed procedure is to utilize the estimated probability of detection $\widehat{P}_d$. Since the true information of parameter θ is unknown, P_d cannot be calculated directly using (3.1). We propose to estimate P_d by plugging in the least-square estimator $\hat{\theta}_{\mathrm{LS}}$ in (3.1). The expectation of the estimator $\hat{P}_d$ remains an increasing function of $|\theta|$, that is,

$$\widehat{P}_d = \Phi\Big(\frac{\hat{\theta}_{\mathrm{LS}} - \sqrt{\lambda}}{\sigma/\sqrt{n}}\Big) + \Phi\Big(\frac{-\hat{\theta}_{\mathrm{LS}} - \sqrt{\lambda}}{\sigma/\sqrt{n}}\Big) \tag{3.5}$$

and

$$E(\widehat{P}_d) = \Phi\Big(\frac{\sqrt{n}}{\sqrt{2}\sigma}(\theta - \sqrt{\lambda})\Big) - \Phi\Big(-\frac{\sqrt{n}}{\sqrt{2}\sigma}(\theta + \sqrt{\lambda})\Big).$$

In the following, the weak signal is identified through replacing P_d, (γ^w, ν^w) and (γ^s, ν^s) in (3.2) by $\widehat{P}_d$, (γ_1, ν_1) and (γ_2, ν_2), where (γ_1, ν_1) and (γ_2, ν_2) satisfy equation (3.3). We denote the identified noise, weak and strong signal set as $(\widehat{\mathbf{S}}^{(N)}, \widehat{\mathbf{S}}^{(W)}, \widehat{\mathbf{S}}^{(S)})$, where

$$\begin{aligned}
\widehat{\mathbf{S}}^{(N)} &= \{i : |\hat{\theta}_{\mathrm{LS},i}| \le \nu_1, i = 1, \ldots, p\} = \{i : \widehat{P}_{d,i} \le \gamma_1\},\\
\widehat{\mathbf{S}}^{(W)} &= \{i : \nu_1 < |\hat{\theta}_{\mathrm{LS},i}| \le \nu_2, i = 1, \ldots, p\} = \{i : \gamma_1 < \widehat{P}_{d,i} \le \gamma_2\}, \quad \text{and}\\
\widehat{\mathbf{S}}^{(S)} &= \{i : |\hat{\theta}_{\mathrm{LS},i}| > \nu_2, i = 1, \ldots, p\} = \{i : \widehat{P}_{d,i} > \gamma_2\}.
\end{aligned}$$

The details of selecting ν_1 and ν_2 are given below.

Note that in finite samples, there is no ideal threshold value ν_1 which can separate signal variables and noise variables perfectly, as there is a trade-off between recovering weak signals and including noise variables. Here, ν_1 is selected to control a signal's false-positive rate τ. Specifically, $\nu_1 = z_{\tau/2}\frac{\sigma}{\sqrt{n}}$ for any given tolerant false-positive rate τ since it can be shown that $P(i \notin \widehat{\mathbf{S}}^{(N)} | \theta_i = 0) = \tau$; see Lemma 3 in the Appendix. Here, we choose the false-positive rate τ to be larger than the τ_0, since we intend to recover most of the weak signals. This is very different from standard model selection which mainly focuses on model selection consistency, but neglects detection of weak signals.

The low threshold value ν_2 for strong signals is selected to ensure that a strong signal can be identified with high probability. We choose $\nu_2 = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, and it can be verified that the estimated detection rate $\widehat{P}_d$ for any identified strong signal stays above $1 - \alpha$. In fact, based on (3.5), $\widehat{P}_d$ satisfies the inequality $P_d > E(\widehat{P}_d)$ when the true signal is strong. Figure 2 illustrates the relationship between P_d and $E(\widehat{P}_d)$. Therefore, there is a high probability that the true detection rate P_d is larger than $1 - \alpha$ when $\hat{P}_d > 1 - \alpha$.

In summary, the main focus of weak signal identification is to recover weak signals as much as possible, at the cost of having a false-positive rate τ in finite

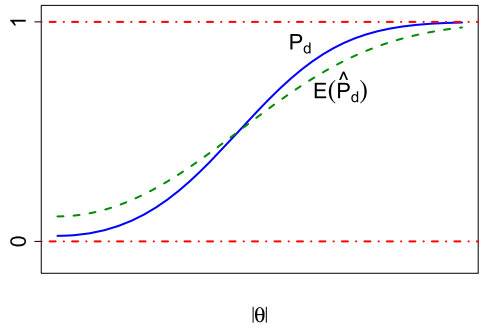

FIG. 2. *P_d and $E(\widehat{P}_d)$.*

samples. This is in contrast to standard model selection procedures which emphasize consistent model selection with a close-to-zero false-positive rate, but at the cost of not selecting most weak signals.

To better understand the difference and connection between the proposed weak signal identification procedure and the standard model selection procedure, we provide Figure 3 for illustration. Let $P_{d,0}(\theta)$ (dashed line) and $P_{d,1}(\theta)$ (dotted line) denote the probabilities of selecting θ in the standard model selection and the proposed weak signal identification, respectively, where $P_{d,0}(\theta) = P(|\hat{\theta}_{\mathrm{LS}}| > \sqrt{\lambda})$, and $P_{d,1}(\theta) = P(\nu_1 < |\hat{\theta}_{\mathrm{LS}}| < \sqrt{\lambda})$. Then the total selection probability $P_{d,2}(\theta)$ (solid line) for the proposed method is $P_{d,2}(\theta) = P_{d,0}(\theta) + P_{d,1}(\theta) = P(|\hat{\theta}_{\mathrm{LS}}| > \nu_1)$. Figure 3 indicates that the proposed procedure recovers weak signals better than the standard model selection procedure, but at a cost of a small false-positive rate of including some noise variables. These two procedures have similar detection power for strong signals.

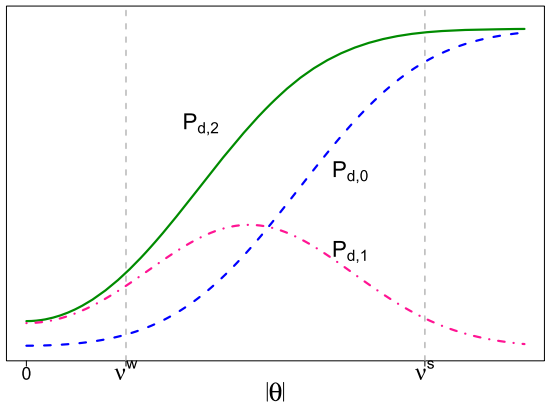

FIG. 3. *Signal's detectability in two-step procedure.*

4. Weak signal inference.

4.1. *Two-step inference method.* In this section, we propose a two-step inference procedure which consists of an asymptotic-based confidence interval for strong signals, and a least-square confidence interval for the identified weak signals. In the following, the proposed procedure is based on the orthogonal design assumption.

The asymptotic-based inference method has been developed for the SCAD estimator [6]. Zou [33] also provides the asymptotic distribution of the adaptive Lasso estimator $\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}$ for nonzero parameters, where $\mathcal{A}_n = \{1, 2, \ldots, q\}$. In finite samples, the adaptive Lasso estimator $\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}$ is biased due to the shrinkage estimation. The bias term of $\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}$ and the covariance matrix estimator of $\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}$ are given by

$$\hat{\boldsymbol{b}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}) = \left(\frac{1}{n}\mathbf{X}_{\mathcal{A}_n}^T\mathbf{X}_{\mathcal{A}_n}\right)^{-1}\left(p'_\lambda(|\hat{\theta}_1|)\operatorname{sgn}(\hat{\theta}_1), \ldots, p'_\lambda(|\hat{\theta}_q|)\operatorname{sgn}(\hat{\theta}_q)\right)^T, \tag{4.1}$$

and

$$\widehat{\operatorname{Cov}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}_n}) = \{\mathbf{X}_{\mathcal{A}_n}^T\mathbf{X}_{\mathcal{A}_n} + n\lambda\boldsymbol{\Omega}\}^{-1}\mathbf{X}_{\mathcal{A}_n}^T\mathbf{X}_{\mathcal{A}_n}\{\mathbf{X}_{\mathcal{A}_n}^T\mathbf{X}_{\mathcal{A}_n} + n\lambda\boldsymbol{\Omega}\}^{-1}\hat{\sigma}^2, \tag{4.2}$$

where $\boldsymbol{\Omega} = \operatorname{diag}\{\frac{\hat{w}_1}{|\hat{\theta}_1|}, \ldots, \frac{\hat{w}_q}{|\hat{\theta}_q|}\}$, and $\hat{w}_i = 1/|\hat{\theta}_{\mathrm{LS},i}|$. Although the bias term is asymptotically negligible, it is important to correct the biased term to get more accurate confidence intervals in finite samples.

Consequently, if the ith variable is identified as a strong signal in $\widehat{\mathbf{S}}^{(S)}$, a $100(1-\alpha)\%$ confidence interval for θ_i can be constructed as

$$\hat{\theta}_i + \hat{b}_{\mathrm{AL},i} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{AL},i}, \tag{4.3}$$

where $\hat{b}_{\mathrm{AL},i}$ and $\hat{\sigma}_{\mathrm{AL},i}$ are the corresponding biased component in (4.1) and the square root of the diagonal variance component in (4.2), respectively. Under the orthogonal design, they are equivalent to

$$\hat{b}_{\mathrm{AL},i} = \frac{\lambda}{|\hat{\theta}_{\mathrm{LS},i}|}\cdot\operatorname{sgn}(\hat{\theta}_i) \quad \text{and} \tag{4.4}$$

$$\hat{\sigma}_{\mathrm{AL},i} = \left(1 + \frac{\lambda}{|\hat{\theta}_i||\hat{\theta}_{\mathrm{LS},i}|}\right)^{-1}\cdot\hat{\sigma}/n. \tag{4.5}$$

The above inference procedure performs well for strong signals ([6, 33] and [12]). However, this procedure does not apply well to weak signals. This is because weak signals are often missed in standard model selection procedures and, therefore, there is no confidence interval constructed for any estimator shrunk to 0. Moreover, even if a weak signal is selected, the variance estimator in (4.2) tends to underestimate its true standard error, and consequently the confidence interval based on (4.3) is under-covered. Here, we propose an alternative confidence interval for a weak signal in $\widehat{\mathbf{S}}^{(W)}$ by utilizing the least-square information as follows.

The proposed inference for weak signals is motivated in that the bias-corrected confidence interval in (4.3) is close to the least-square confidence interval when a signal is strong. Therefore, it is natural to construct a least-square confidence interval for a weak signal to solve the problem of excessive shrinkage for weak signal estimators.

If the ith variable is identified as a weak signal in $\widehat{\mathbf{S}}^{(W)}$, we construct a $100(1-\alpha)\%$ least-square confidence interval for θ_i as

$$\hat{\theta}_{\mathrm{LS},i} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{LS},i}, \tag{4.6}$$

where $\hat{\theta}_{\mathrm{LS},i}$ and $\hat{\sigma}_{\mathrm{LS},i}$ are the components of the least-square estimator and the square root of the diagonal component of the covariance matrix estimator:

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$
$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}) = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2.$$

Under the orthogonal design, $\hat{\theta}_{\mathrm{LS},i}$ and $\hat{\sigma}_{\mathrm{LS},i}$ are

$$\hat{\theta}_{\mathrm{LS},i} = \mathbf{X}_i^T\mathbf{y}/n,$$
$$\hat{\sigma}_{\mathrm{LS},i} = \hat{\sigma}/n.$$

In summary, if a nonzero signal is detected, combining (4.3) and (4.6), we provide a new two-step confidence interval for the ith variable as follows:

$$\{\hat{\theta}_{\mathrm{LS},i} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{LS},i}\}\mathbf{1}_{\{i\in\widehat{\mathbf{S}}^{(W)}\}} + \{\hat{\theta}_i + \hat{b}_{\mathrm{AL},i} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{AL},i}\}\mathbf{1}_{\{i\in\widehat{\mathbf{S}}^{(S)}\}}.$$

Here, we propose different confidence interval constructions for weak and strong signals, and the proposed inference is a mixed procedure combining (4.3) and (4.6). Our inference procedure performs similarly to the asymptotic inference for strong signals, but outperforms the existing inference procedures in that the proposed confidence interval provides more accurate coverage for weak signals. Note that if a signal strength is too weak, neither existing methods nor our method can provide reasonably good inferences. Nevertheless, our method still provides a better inference than the asymptotic-based method across all signal levels.

4.2. *Finite sample theories*. In this section, we establish finite sample theory on coverage rate for the proposed two-step inference method, and compare it with the coverage rate of the asymptotic-based inference method. The asymptotic properties for penalized estimators have been investigated by [6, 8, 33, 35] and many others. When the sample size is sufficiently large and the signal strength is strong, the asymptotic inference is quite accurate in capturing the information of the penalized estimators. For instance, the covariance estimator of the penalized estimates in (4.2) is a consistent estimator [8]. However, the sandwich estimator of the covariance only performs well for strong signals, not for weak signals in finite samples.

Therefore, it is important to investigate the finite sample property of the penalized estimator, and especially the weak signal estimators for the proposed method.

We construct the exact coverage rates of the $100(1-\alpha)\%$ confidence intervals for the proposed method and the asymptotic method when the sample size is finite. The derivation for finite sample theory is very different from the asymptotic theory. In addition, since the coverage rate function is not monotonic, we need to compare the difference of the two coverage rates piecewisely.

Given a confidence level parameter α, the following regularity conditions are required for selecting the false-positive rate τ:

(C1) $\tau \geq \alpha$,
(C2) $\frac{\alpha+\tau}{2} < \Phi(-\frac{1}{2}z_{\alpha/2})$, which is equivalent to $\tau < 2\Phi(-\frac{1}{2}z_{\alpha/2}) - \alpha$.

Condition (C1) is to ensure that the second step of the proposed method is able to identify weak signals. Condition (C2) provides a range of τ, so the false positive-rate is not too large. In addition, we also assume that λ satisfies the criterion

$$\sqrt{\lambda} \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{4.7}$$

The criterion in (4.7) implies that our focus is the case when $\sqrt{\lambda} \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where excessive shrinkage might affect weak signal selection. It can be verified that $\alpha \geq \tau_0$ if λ is in this range, and this guarantees that $\tau > \tau_0$.

In the following, for any parameter θ and parameter ν associated with a different level of tuning, we introduce three probability functions, P_s, CR_a and CR_b as follows. Let P_s be the detection power of θ:

$$P_s(\theta, \nu) = \Phi\left(\frac{\theta - \nu}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-\theta - \nu}{\sigma/\sqrt{n}}\right).$$

We define CR_a as the coverage probability based on the asymptotic inference approach when $|\hat{\theta}_{\mathrm{LS}}|$ is larger than ν:

$$\mathrm{CR}_a(\theta, \nu) = \begin{cases} \left\{P_s(\theta, \nu) - 2\Phi\left(-z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sigma}\right)\right\} & \\ \quad \times I_{\{\nu \leq z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\}}, & \text{if } |\theta| \leq \left|\nu - z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right|, \\ \Phi\left(z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sigma}\right) - \Phi\left(\dfrac{\sqrt{n}(\nu-\theta)}{\sigma}\right), & \\ \quad \text{if } \left|\nu - z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right| \leq |\theta| \leq \nu + z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sqrt{n}}, & \\ 1 - 2\Phi\left(-z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sigma}\right), & \text{if } |\theta| > \nu + z_{\alpha/2}\dfrac{\tilde{\sigma}(\theta)}{\sqrt{n}}, \end{cases}$$

where $\tilde{\sigma}(\theta) = (1 + \frac{\lambda}{\theta^2})^{-1}\sigma$; and CR_b is the coverage probability based on the least-square confidence interval when $|\hat{\theta}_{\mathrm{LS}}|$ is larger than ν:

$$\mathrm{CR}_b(\theta, \nu) = \begin{cases} \{P_s(\theta, \nu) - \alpha\} \cdot I_{\{\nu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}}, & \text{if } |\theta| \le \left|\nu - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right|, \\ 1 - \dfrac{\alpha}{2} - \Phi\left(\dfrac{\sqrt{n}(\nu - \theta)}{\sigma}\right), & \\ \quad \text{if } \left|\nu - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right| \le |\theta| \le \nu + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, & \\ 1 - \alpha, & \text{if } |\theta| > \nu + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}. \end{cases}$$

The explicit expressions of coverage rates based on the asymptotic and the proposed two-step methods are provided in the following lemma.

LEMMA 2. *Suppose n, σ and tuning parameter λ are given, the coverage rate $\mathrm{CR}_1(\theta)$ of the $100(1-\alpha)\%$ confidence interval for any coefficient θ based on the asymptotic inference is*

$$\mathrm{CR}_1(\theta) = \frac{\mathrm{CR}_a(\theta, \nu_0)}{P_s(\theta, \nu_0)}, \tag{4.8}$$

where $\nu_0 = \sqrt{\lambda}$. Given any τ, the coverage rate $\mathrm{CR}(\theta)$ of the $100(1-\alpha)\%$ confidence interval for any coefficient θ using the proposed two-step inference method is given by

$$\mathrm{CR}(\theta) = \frac{\mathrm{CR}_b(\theta, \nu_1) + \mathrm{CR}_a(\theta, \nu_2) - \mathrm{CR}_b(\theta, \nu_2)}{P_s(\theta, \nu_1)}, \tag{4.9}$$

where $\nu_0 = \sqrt{\lambda}$, $\nu_1 = z_{\tau/2}\frac{\sigma}{\sqrt{n}}$, and $\nu_2 = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

The derivations of $\mathrm{CR}_1(\theta)$ and $\mathrm{CR}(\theta)$ are provided in the proof of Lemma 2 in the Appendix. In fact, $\mathrm{CR}_1(\theta)$ is the conditional coverage probability based on the asymptotic confidence interval, given that θ is selected using tuning parameter λ. Similarly, $\mathrm{CR}(\theta)$ is the conditional coverage probability of the proposed confidence interval in (4.7), given that θ is selected based on the two-step procedure. The expression of $\mathrm{CR}(\theta)$ in (4.9) can be interpreted as the summation of two sub-components, where the first component $\frac{\mathrm{CR}_b(\theta,\nu_1)-\mathrm{CR}_b(\theta,\nu_2)}{P_s(\theta,\nu_1)}$, corresponds to the conditional coverage probability of the least-square confidence interval when $\nu_1 < |\hat{\theta}_{\mathrm{LS}}| < \nu_2$, and the second component $\frac{\mathrm{CR}_a(\theta,\nu_2)}{P_s(\theta,\nu_1)}$, is the conditional coverage probability of the asymptotic-based confidence interval when $|\hat{\theta}_{\mathrm{LS}}| > \nu_2$.

In addition, we show in the supplement [24] that both $\mathrm{CR}_1(\theta)$ and $\mathrm{CR}(\theta)$ are piecewise smooth functions, and require one to compare two coverage rates at each interval separately. We introduce the boundary points associated with $\mathrm{CR}_1(\theta)$ and

$\mathrm{CR}(\theta)$ as follows. Let c_1, c_2, c_3 and c_4 be the solutions of $\theta = \sqrt{\lambda} - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, and $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, respectively. Here, c_1 and c_2 are the boundary points of piecewise intervals for $\mathrm{CR}_1(\theta)$ in (4.8), and c_3 and c_4 are the boundary points of piecewise intervals for $\mathrm{CR}(\theta)$ in (4.9). It can be shown that the orders of c_1, c_2, c_3 and c_4 satisfy $c_1 < c_3 < c_2 < c_4$. More specific ranges for c_1, c_2, c_3 and c_4 are provided in Lemma 4 of the Appendix. Since there are no explicit solutions for these boundary points, we rely on the orders of these boundary points to examine the difference between $\mathrm{CR}_1(\theta)$ and $\mathrm{CR}(\theta)$.

In the following, we define $\Delta(\theta) = \mathrm{CR}(\theta) - \mathrm{CR}_1(\theta)$ as a difference function between $\mathrm{CR}(\theta)$ and $\mathrm{CR}_1(\theta)$. Theorem 1 and Theorem 2 provide the uniform low bounds of $\Delta(\theta)$ for different ranges of λ when $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \sqrt{\lambda} < (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$ and $\sqrt{\lambda} \ge (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$. The mathematical details of the proofs are provided in the Appendix and supplement materials.

THEOREM 1. *Under conditions* (C1)–(C2), *if* λ *satisfies* $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \sqrt{\lambda} < (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$, *the piecewise lower bounds for* $\Delta(\theta)$ *are provided as follows*:

(a) *when* $\theta \in [0, c_1]$, $\Delta(\theta) \ge 1 - \frac{\alpha}{\tau} > 0$;
(b) *when* $\theta \in [c_1, \nu_0]$, $\Delta(\theta) \ge \frac{2}{1+\alpha} - 2\Phi(\frac{1}{2}z_{\alpha/2}) > 0$;
(c) *when* $\theta \in [\nu_0, +\infty)$, $\Delta(\theta)$ *satisfies either* $\Delta(\theta) \ge 0$ *or* $-\frac{\alpha}{2} < \Delta(\theta) < 0$.

In addition, a more specific lower bound for $\Delta(\theta)$ *on* $[\nu_0, +\infty)$ *is given by*

$$\Delta(\theta) \ge \begin{cases} -4\left(1 - \dfrac{\alpha}{2}\right)\Phi\left(-\dfrac{3}{2}z_{\alpha/2}\right), & \text{if } \theta \in \left[\nu_0, \min\{\nu_3, c_3\}\right], \\ \text{See Table 1}, & \text{if } \theta \in \left[\min\{\nu_3, c_3\}, \max\{\nu_3, c_2\}\right], \\ -\dfrac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2}) - \dfrac{\alpha(1-\alpha)}{2-\alpha}, & \text{if } \theta \in \left[\max\{\nu_3, c_2\}, c_4\right], \\ -(1-\alpha)\dfrac{\Phi(-\frac{3}{2}z_{\alpha/2})}{\Phi(\frac{3}{2}z_{\alpha/2})^2}, & \text{if } \theta \in [c_4, \nu_4], \\ -(1-\alpha)\dfrac{\Phi(-2z_{\alpha/2})}{\Phi(2z_{\alpha/2})^2}, & \text{if } \theta \in [\nu_4, \infty), \end{cases}$$

where $\nu_3 = (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$, *and* $\nu_4 = \sqrt{\lambda} + 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. *Table 1 provides the lower bounds for* $\Delta(\theta)$ *on interval* $[\min\{\nu_3, c_3\}, \max\{\nu_3, c_2\}]$ *under three cases*.

THEOREM 2. *Under conditions* (C1)–(C2), *if* λ *satisfies* $\sqrt{\lambda} \ge (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$, *the lower bounds for* $\Delta(\theta)$ *are provided as follows*:

1. *When* $\theta \in [0, \min\{\nu_3, c_1\}]$, $\Delta(\theta) \ge 1 - \frac{\alpha}{\tau} > 0$.

TABLE 1
Specific bounds of $\Delta(\theta)$ *on interval* $[\min\{\nu_3, c_3\}, \max\{\nu_3, c_2\}]$

Case 1:	$\theta \in [c_3, \nu_3]$	$\theta \in [\nu_3, c_2]$
$c_3 < \nu_3 < c_2$	$-2\Phi(-\frac{3}{2}z_{\alpha/2})$	$-\frac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2}) - \frac{\alpha(1-\alpha)}{2-\alpha}$
Case 2:	$\theta \in [c_3, c_2]$	$\theta \in [c_2, \nu_3]$
$c_3 < c_2 < \nu_3$	$-2(1-\alpha)\Phi(-\frac{3}{2}z_{\alpha/2})$	$-\frac{1-\alpha}{[\Phi(\frac{1}{2}z_{\alpha/2})]^2}\Phi(-2z_{\alpha/2})$
Case 3:	$\theta \in [\nu_3, c_2]$	
$\nu_3 < c_3 < c_2$	$-\frac{\alpha}{2}$	

2. *When* $\theta \in [\min\{\nu_3, c_1\}, \nu_0]$, *see Table* 2.
3. *When* $\theta \in [\nu_0, +\infty)$, $\Delta(\theta) \geq 0$ *or* $-\frac{\alpha}{2} < \Delta(\theta) < 0$.

Theorem 1 and Theorem 2 indicate that the proposed method outperforms the asymptotic-based method, with a uniform lower bound for $\Delta(\theta)$ when $\theta \in [0, \nu_0]$. More specifically, the lower bound of $\Delta(\theta)$ depends on α and τ for case (i) ($\theta \in [0, c_1]$) in Theorem 1 and case (i) ($\theta \in [0, \min\{\nu_3, c_1\}]$) in Theorem 2. Since we select τ to be larger than α, it is clear that $\Delta(\theta)$ is bounded above zero. For case (ii) ($\theta \in [c_1, \nu_0]$) in Theorem 1 and case (ii) ($\theta \in [\min\{\nu_3, c_1\}, \nu_0]$) in Theorem 2, the lower bound of $\Delta(\theta)$ only depends on α. In fact, the minimum value of $\frac{2}{1+\alpha} - 2\Phi(\frac{1}{2}z_{\alpha/2})$ is larger than 0.22 if $\alpha \in [0.05, 0.1]$, based on Theorem 1. This also confirms that the proposed method provides a confidence region with at least 22% improvement in coverage rate than the one based on the asymptotic method. The lower bounds of case (ii) in Theorem 2 can be interpreted in a similar way.

In addition, both Theorem 1 and Theorem 2 imply that when $\theta \in (\nu_0, +\infty)$ with a moderately large coefficient, the proposed method performs better than, or close to, the asymptotic method. In summary, the two-step inference method provides more accurate coverage than the one based on the asymptotic inference, and is also more effective for the weak signal region.

In Theorem 1, since the order relationships among c_2, c_3 and ν_3 change for different ranges of tuning parameters and choices of false positive rate τ, it leads to the three cases in Table 1. Similarly, the order relationships among ν_3 and c_1 also change for different choices of λ and τ in Theorem 2, leading to the two cases

TABLE 2
Specific bounds of $\Delta(\theta)$ *on interval* $[\min\{\nu_3, c_1\}, \nu_0]$

Case 4:	$\theta \in [\nu_3, \nu_0]$	
$\nu_3 < c_1$	$2 - \alpha - 2\Phi(\frac{1}{2}z_{\alpha/2})$	
Case 5:	$\theta \in [c_1, \nu_3]$	$\theta \in [\nu_3, \nu_0]$
$c_1 < \nu_3$	$\Phi(-\frac{1}{2}z_{\alpha/2}) - \frac{\alpha}{2}$	$2 - \alpha - 2\Phi(\frac{1}{2}z_{\alpha/2})$

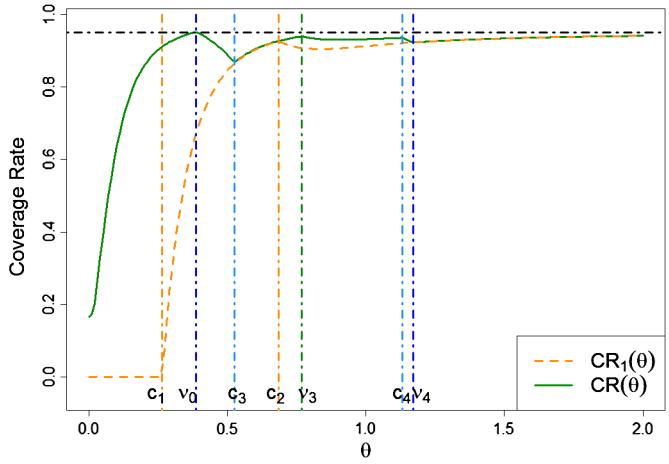

FIG. 4. $\mathrm{CR}(\theta)$ *versus* $\mathrm{CR}_1(\theta)$ (*an example*: *Case* 1).

in Table 2. Figure 4 illustrates an example for case 1. Figures for the other four cases are provided in the supplemental material.

5. Finite sample performance.

5.1. *Simulation studies*. To examine the empirical performance of the proposed inference procedure, we conduct simulation studies to evaluate the accuracy of the confidence intervals described in Section 4.1. We generate 400 simulated data with a sample size of n under the linear model $y = \mathbf{X}\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$, where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ and $\mathbf{X}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I_n})$. We allow covariates $\mathbf{X}$ to be correlated with an AR(1) correlation structure, and the pairwise correlation $\mathrm{cor}(\mathbf{X}_i, \mathbf{X}_j) = \rho^{|i-j|}$. We choose $(n, p, \sigma) = (100, 20, 2)$ and $(400, 50, 2)$, and $\rho = 0, 0.2$ or 0.5 for each setting. In addition, the p-dimensional coefficient vector $\boldsymbol{\theta} = (1, 1, 0.5, \theta, 0, \ldots, 0)$, which consists of two strong signals of 1's, one moderate strong signal of 0.5, one varying-signal θ, and $(p-4)$ null variables. We let the coefficient θ vary between 0 (null) to 1 (strong signal) to examine the confidence coverages across different signal strength levels.

We construct 95% confidence intervals for an identified signal based on (4.7). We implement the glmnet package in R [9] to obtain the adaptive Lasso estimator. We choose the tuning parameter λ based on the Bayesian information criterion (BIC), because of its consistency property to select the true model [29]. Here, we follow a strategy by [28] to select the BIC tuning parameter for the adaptive Lasso penalty (details are provided in Appendix A.2). The standard deviation $\hat{\sigma}$ is estimated based on the scaled Lasso method [25], using the "scalreg" package in R. We replace $\hat{\theta}_i$ by its bias-corrected form $\hat{\theta}_i + \hat{b}_{\mathrm{AL},i}$ in (4.5) when estimating $\hat{\sigma}_{\mathrm{AL}}$, which achieves better estimation of the true standard deviation. For comparison,

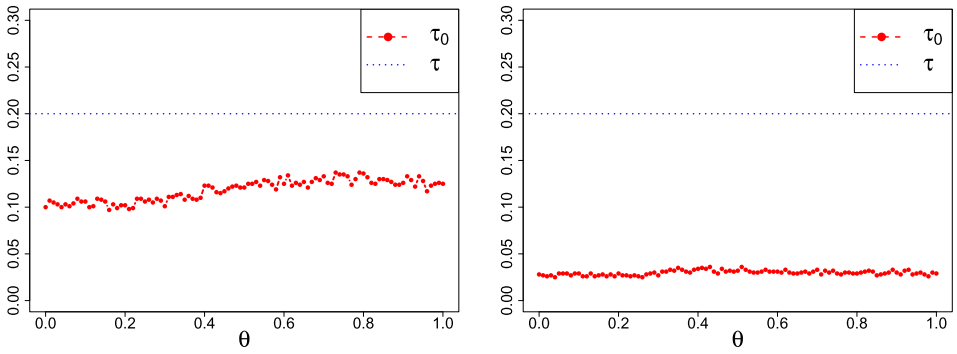

FIG. 5. *False positive rate for simulation setting* 1 (*left*) *and* 2 (*right*).

we also construct standard confidence intervals based on the asymptotic formula in (4.3), along with the bootstrap method [5], the smoothed bootstrap method [4], the perturbation method [19] and the de-biased Lasso method [13]. The de-biased method is implemented using the R codes provided by Montanari's website. For both regular bootstrap and smoothed bootstrap methods, the number of bootstrap sampling is set to be 4000 [4]. For the perturbation method, the resampling time is set to be 500 according to [19].

In addition, the coverage rate for the OLS estimator is included as a benchmark since there is no shrinkage in OLS estimation and the confidence interval is the most accurate. Here, the OLS estimator $\hat{\theta}_{\mathrm{LS}}$ given in (4.6) is estimated from the full model. We used the estimator from the full model because our method assumes that the covariates are orthogonally designed. Under this assumption, the least square estimator under a submodel is the same estimator as that under the full model. If covariates are correlated, the estimator under the correctly specified submodel is more efficient than the one under the full model. However, we cannot guarantee that the selected submodel is correctly specified. If the submodel is misspecified, then the $\hat{\theta}_{\mathrm{LS}}$ could be biased, which could lead to inaccurate inference for the coefficients of the selected variables. Note that selection of the wrong model is likely, especially when weak signals exist.

Figure 5 illustrates the relationship between τ_0 and τ when $\rho = 0.2$ for two model settings $(n, p, \sigma) = (100, 20, 2)$ and $(400, 50, 2)$, where $\tau_0 = 2\Phi(-\frac{\sqrt{n\lambda}}{\sigma})$ based on Section 3.1. We choose τ larger than τ_0 according to Section 3, that is, the false-positive rate in the weak signal recovery procedure is slightly larger than the false-positive rate in the model selection procedure. In these two model settings, τ_0 are around 0.1 and 0.03, respectively; here, we select τ as 0.2. In practice, the selection of τ is flexible, and can be determined by a tolerance level for including noise variables.

Figure 6 and Figure 7 provide the coverage probabilities for θ varying between 0 and 1 when $\rho = 0.2$ in two model settings. In each figure, ν^s and ν^w are the average

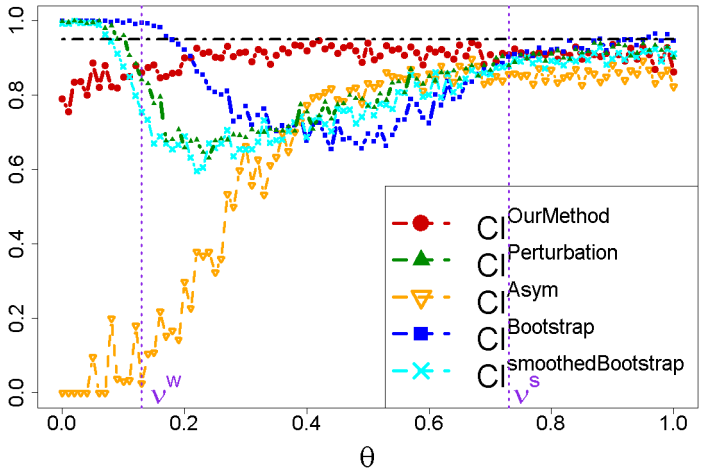

FIG. 6. *95% confidence interval's coverage rates for simulation setting* 1 *when* $\rho = 0.2$.

threshold coefficients corresponding to the detection powers $P_d = 0.9$ and 0.1, respectively. When the signal strength is close to zero, neither of the coverage rates using our method and the asymptotic method are accurate. However, the proposed method is still better than the asymptotic one, since the asymptotic coverage rate is close to zero; while the bootstrap and perturbation methods tend to provide over-coverage confidence intervals. The proposed method becomes more accurate as the magnitude of signal θ increases, and also outperforms all the other methods especially in the weak signal region. For example, in setting 1, the coverage rate of the proposed method is quite close to 95% when the signal strength is larger than

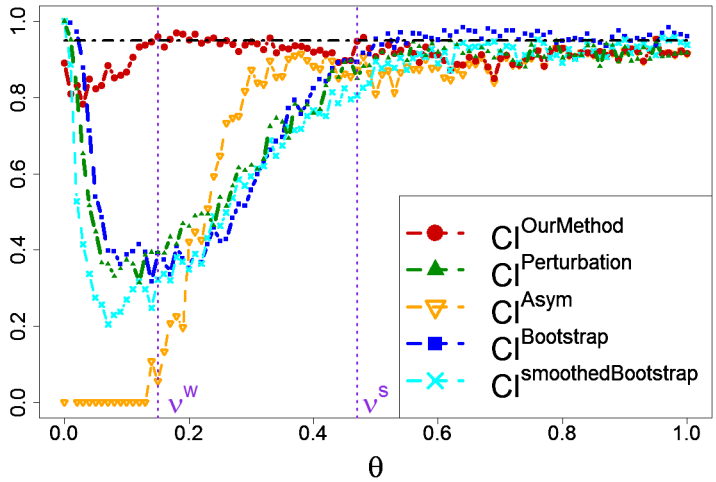

FIG. 7. *95% confidence interval's coverage rates for simulation setting* 2 *when* $\rho = 0.2$.

0.4. On the other hand, the resampling methods and asymptotic inference provide low coverage rates for signal strength below 0.8. When signal strength is above 0.8, the coverages from all methods are accurate and close to 95%.

The results for correlated covariates settings are provided in Table 3. For each setting, we select two different values of θ whose detection probabilities P_d are between 0.1 and 0.9. Here, the first θ is relatively weaker, and the second one is at the boundary of strong signal. For all these settings, the asymptotic inference, bootstrap and perturbation methods provide confidence intervals far below 95% when signals are weak. In general, our method provides a stable inference even when the correlation coefficient increases, and the coverage rate for weak signals is between 90–96% when $\rho = 0.5$. The asymptotic-based inference has the lowest coverage rates among all, and performs extremely poorly when ρ is larger. The coverage rates based on the perturbation method are all below 75% for weak signals. Note that the coverage rate improvement using the smoothed bootstrap method is not significant compared to the standard bootstrap method. In addition, for $n = 100$, $p = 50$, the bootstrap and smooth bootstrap methods face a singular-designed matrix problem due to small sample size, which does not produce any simulation results 7–10% of the time. The average coverage rates provided in Table 3 might not be valid and are marked with *.

Table 4 provides the CI lengths of all methods for both weak and strong signals. In general, the proposed method provides narrower confidence intervals and better coverage rates than the perturbation and bootstrap methods, and shorter confidence intervals with comparable coverage rates as the de-biased Lasso method for strong signals. For example, when $(n, p, \rho) = (100, 20, 0)$ and $\theta = 0.75$, the coverage rate of our method is 94.4%, compared to 87.6% based on the perturbation method, 91.4% based on the bootstrap method, and 93.8% based on the de-biased method. The corresponding CI length of our method equals 0.770, which is smaller than the 0.911 from the perturbation method, 1.020 from the bootstrap method, and 1.101 from the de-biased method. Furthermore, our CI is also shorter compared to the least square CI for strong signals in general.

For weak signals, our CI is wider than the perturbation and bootstrap methods. This is because both the perturbation and bootstrap methods provide inaccurate coverage rates which tend to be smaller than 95%. For example, when $(n, p, \rho) = (100, 20, 0)$ and $\theta = 0.3$, the coverage rate of our method is 94.4%, compared to 67.3% based on the perturbation method, 74.5% based on the bootstrap method, and 94.5% based on the de-biased Lasso method. The corresponding CI length of our method is 0.862, which is wider than the 0.652 from the perturbation method and the 0.786 from the bootstrap method, but is still shorter than the 1.071 from the de-biased Lasso method.

Figure 8 also presents the probabilities of assigning each signal category for a given θ value, where the probabilities for identified strong signal $[P(i \in \widehat{\mathbf{S}}^{(S)})]$, weak signal $[P(i \in \widehat{\mathbf{S}}^{(W)})]$ and null variable $[P(i \in \widehat{\mathbf{S}}^{(N)})]$ are denoted as solid,

TABLE 3
Coverage probabilities of confidence regions when $\sigma = 2$

			$p = 20$			$p = 50$		
n	θ		$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
100	0.3	CR^{Our}	94.4	92.6	91.1	92.1	91.1	95.3
		CR^{Asym}	61.5	61.2	38.3	33.3	18.5	21.4
		CR^{Ptb}	67.3	68.6	74.2	68.6	64.8	58.6
		CR^{Bs}	74.5	77.0	88.7	100.0*	100.0*	100.0*
		CR^{smBS}	68.1	65.4	74.3	95.1*	93.5*	92.3*
		CR^{OLS}	93.2	93.2	94.0	94.5	93.0	95.0
		$CR^{Lasso\text{-}debiased}$	94.5	95.5	97.0	94.8	94.0	96.5
	0.75	CR^{Our}	94.4	92.9	91.9	93.8	92.5	93.6
		CR^{Asym}	89.6	87.4	75.1	85.3	77.5	63.9
		CR^{Ptb}	87.6	90.9	86.4	90.0	93.8	78.9
		CR^{Bs}	91.4	90.7	88.0	98.9*	98.8*	100.0*
		CR^{smBS}	89.3	89.1	89.2	91.3*	95.3*	91.1*
		CR^{OLS}	92.8	96.0	94.0	95.5	95.5	94.0
		$CR^{Lasso\text{-}debiased}$	93.8	96.5	96.8	96.3	96.0	96.3
200	0.2	CR^{Our}	94.6	94.7	93.3	95.3	93.7	91.3
		CR^{Asym}	52.0	51.6	38.1	15.9	22.4	18.0
		CR^{Ptb}	61.4	65.3	69.4	48.1	44.2	49.3
		CR^{Bs}	58.5	58.1	72.8	56.1	61.0	63.5
		CR^{smBS}	54.7	50.5	62.6	46.0	48.8	46.4
		CR^{OLS}	95.2	94.2	95.8	95.2	95.5	95.8
		$CR^{Lasso\text{-}debiased}$	93.5	95.0	96.5	95.5	96.0	94.8
	0.6	CR^{Our}	95.5	93.0	91.5	93.7	91.6	90.3
		CR^{Asym}	88.8	86.2	76.6	88.5	82.7	65.0
		CR^{Ptb}	90.2	92.6	86.1	84.2	88.0	89.6
		CR^{Bs}	90.7	91.7	88.4	86.9	89.4	82.7
		CR^{smBS}	88.4	89.5	91.2	80.7	84.2	81.4
		CR^{OLS}	96.2	96.0	96.5	95.2	93.8	94.2
		$CR^{Lasso\text{-}debiased}$	95.5	95.3	96.3	95.0	95.0	94.0
400	0.15	CR^{Our}	93.6	94.6	93.4	97.0	96.3	90.5
		CR^{Asym}	31.7	33.8	44.8	9.1	11.6	10.7
		CR^{Ptb}	33.6	51.0	60.3	33.6	39.3	44.7
		CR^{Bs}	35.3	54.2	57.9	35.3	39.1	38.6
		CR^{smBS}	27.4	49.9	52.2	27.4	32.2	31.3
		CR^{OLS}	92.5	96.8	96.0	94.8	95.5	94.2
		$CR^{Lasso\text{-}debiased}$	90.8	93.3	91.5	94.0	96.5	93.0
	0.4	CR^{Our}	94.8	92.2	92.2	92.7	92.1	92.8
		CR^{Asym}	94.0	91.2	85.5	91.7	88.7	72.6
		CR^{Ptb}	79.5	89.3	89.2	79.5	75.8	70.7
		CR^{Bs}	87.5	89.3	80.3	87.5	82.4	70.3
		CR^{smBS}	79.8	87.2	84.3	79.8	76.6	71.0
		CR^{OLS}	95.8	93.2	93.5	94.5	94.8	93.0
		$CR^{Lasso\text{-}debiased}$	90.0	92.5	93.5	95.8	94.3	93.8

Note: The values are multiplied by 100. *Indicates that the bootstrap and smooth bootstrap methods encounter a singular-designed matrix problem (7–10% times), and only partial simulation results are used for calculation.

TABLE 4
Average widths of confidence intervals when $\sigma = 2$

			$p = 20$			$p = 50$		
n	θ		$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
100	0.3	$\text{len CR}^{\text{Our}}$	0.862	0.889	1.098	1.083	1.131	1.403
		$\text{len CR}^{\text{Asym}}$	0.594	0.593	0.582	0.542	0.541	0.519
		$\text{len CR}^{\text{Ptb}}$	0.652	0.691	0.803	0.679	0.672	0.829
		$\text{len CR}^{\text{Bs}}$	0.786	0.818	0.954	1.717*	1.756*	2.206*
		$\text{len CR}^{\text{smBS}}$	0.626	0.654	0.746	0.998*	1.014*	1.296*
		$\text{len CR}^{\text{OLS}}$	0.864	0.900	1.094	1.106	1.135	1.401
		$\text{len CR}^{\text{Lasso-debiased}}$	1.071	1.081	1.304	1.179	1.220	1.487
	0.75	$\text{len CR}^{\text{Our}}$	0.770	0.794	1.031	0.956	1.045	1.306
		$\text{len CR}^{\text{Asym}}$	0.659	0.659	0.648	0.612	0.592	0.579
		$\text{len CR}^{\text{Ptb}}$	0.911	0.929	1.129	0.973	0.971	1.154
		$\text{len CR}^{\text{Bs}}$	1.020	1.054	1.262	1.786*	1.886*	2.276*
		$\text{len CR}^{\text{smBS}}$	0.902	0.944	1.114	1.073*	1.134*	1.354*
		$\text{len CR}^{\text{OLS}}$	0.866	0.899	1.094	1.118	1.151	1.399
		$\text{len CR}^{\text{Lasso-debiased}}$	1.101	1.126	1.414	1.193	1.269	1.529
200	0.2	$\text{len CR}^{\text{Our}}$	0.580	0.603	0.743	0.628	0.659	0.812
		$\text{len CR}^{\text{Asym}}$	0.412	0.420	0.427	0.391	0.370	0.379
		$\text{len CR}^{\text{Ptb}}$	0.436	0.444	0.526	0.381	0.356	0.445
		$\text{len CR}^{\text{Bs}}$	0.458	0.504	0.586	0.470	0.504	0.600
		$\text{len CR}^{\text{smBS}}$	0.384	0.434	0.498	0.356	0.388	0.454
		$\text{len CR}^{\text{OLS}}$	0.581	0.603	0.745	0.635	0.658	0.815
		$\text{len CR}^{\text{Lasso-debiased}}$	0.635	0.681	0.817	0.838	0.832	0.845
	0.6	$\text{len CR}^{\text{Our}}$	0.517	0.567	0.700	0.556	0.614	0.765
		$\text{len CR}^{\text{Asym}}$	0.473	0.467	0.471	0.437	0.433	0.421
		$\text{len CR}^{\text{Ptb}}$	0.673	0.699	0.864	0.715	0.727	0.859
		$\text{len CR}^{\text{Bs}}$	0.714	0.734	0.902	0.814	0.820	0.992
		$\text{len CR}^{\text{smBS}}$	0.708	0.738	0.894	0.732	0.748	0.892
		$\text{len CR}^{\text{OLS}}$	0.584	0.604	0.743	0.641	0.661	0.818
		$\text{len CR}^{\text{Lasso-debiased}}$	0.689	0.728	0.933	0.865	0.862	0.866
400	0.15	$\text{len CR}^{\text{Our}}$	0.397	0.416	0.516	0.418	0.434	0.537
		$\text{len CR}^{\text{Asym}}$	0.293	0.296	0.311	0.283	0.283	0.278
		$\text{len CR}^{\text{Ptb}}$	0.309	0.312	0.391	0.226	0.247	0.261
		$\text{len CR}^{\text{Bs}}$	0.322	0.314	0.372	0.244	0.272	0.290
		$\text{len CR}^{\text{smBS}}$	0.308	0.298	0.352	0.214	0.242	0.254
		$\text{len CR}^{\text{OLS}}$	0.401	0.416	0.515	0.419	0.434	0.537
		$\text{len CR}^{\text{Lasso-debiased}}$	0.412	0.434	0.438	0.436	0.432	0.430
	0.4	$\text{len CR}^{\text{Our}}$	0.368	0.401	0.492	0.381	0.419	0.516
		$\text{len CR}^{\text{Asym}}$	0.327	0.329	0.337	0.306	0.303	0.306
		$\text{len CR}^{\text{Ptb}}$	0.507	0.531	0.638	0.545	0.552	0.602
		$\text{len CR}^{\text{Bs}}$	0.530	0.542	0.654	0.560	0.574	0.670
		$\text{len CR}^{\text{smBS}}$	0.578	0.596	0.712	0.578	0.600	0.698
		$\text{len CR}^{\text{OLS}}$	0.401	0.418	0.515	0.419	0.434	0.536
		$\text{len CR}^{\text{Lasso-debiased}}$	0.432	0.464	0.477	0.442	0.440	0.435

Note: *Indicates that the bootstrap and smooth bootstrap methods encounter a singular-designed matrix problem (7–10% times), and only partial simulation results are used for calculation.

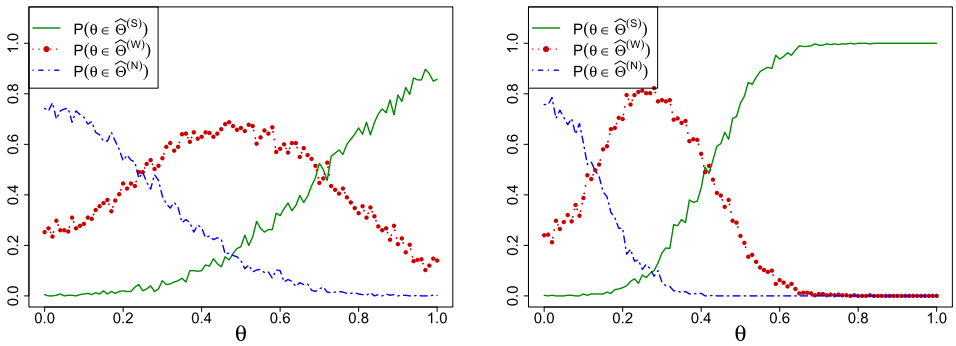

FIG. 8. *Empirical probabilities of identifying each signal level. Left*: *setting* 1. *Right*: *setting* 2.

dotted and dashed lines, respectively. Here, i corresponds to the index of coefficient θ. The probability of each identified signal category relies on signal strength. Specifically, when a signal is close to zero, it is likely to be identified as zero most of the time, with the highest $P(i \in \widehat{\mathbf{S}}^{(N)})$; when a signal falls into the weak signal region, $P(i \in \widehat{\mathbf{S}}^{(W)})$ becomes dominant; and when θ increases to be a strong signal, $P(i \in \widehat{\mathbf{S}}^{(S)})$ also gradually increases and reaches to 1.

5.2. *HIV data example.* In this section, we apply HIV drug resistance data (http://hivdb.stanford.edu/) to illustrate the proposed method. The HIV drug resistance study aims to identify the association of protease mutations with susceptibility to the antiretroviral drug. Since antiretroviral drug resistance is a major obstacle to the successful treatment of HIV-1 infection, studying the generic basis of HIV-1 drug resistance is crucial for developing new drugs and designing an optimal therapy for patients. The study was conducted on 702 HIV-infected patients, where 79 out of 99 protease codons in the viral genome have mutations. Here, the drug resistance is measured in units of IC_{50}.

We consider a linear model:

$$\mathbf{y} = \sum_{i=1}^{p} \mathbf{X}_i \theta_i + \boldsymbol{\varepsilon}, \tag{5.1}$$

where the response variable $\mathbf{y}$ is the log-transformation of nonnegative IC_{50}, and the model predictors $\mathbf{X}_i$ are binary variables indicating the mutation presence for each codon. For each predictor, 1 represents mutation and 0 represents no mutation. The total number of candidate codons $p = 79$. We are interested in examining which codon mutations have effect on drug resistance.

We apply the proposed two-step inference method to identify codons' mutation presence which have strong or mild effects on HIV drug resistance. We use the GLMNET in R to obtain the adaptive Lasso estimator for the linear model in (5.1), where the initial weight of each coefficient is based on the OLS estimator. The

tuning parameter λ is selected by the Bayesian information criterion, and σ is estimated similarly as in [33]. To control the noise variable selection, we choose $\tau = 0.05$. According to the proposed identification procedure in Section 3.2, we calculate two threshold values ν_1 and ν_2 as 0.061 and 0.136, which correspond to two threshold probabilities, $\gamma_1 = 0.327$ and $\gamma_2 = 0.975$, for identifying weak and strong signals, respectively.

We constructed 95% confidence intervals using the proposed method and the perturbation approach [19] for the chosen variables. Both the standard bootstrap and smoothed bootstrap methods are not applicable to the HIV data. Since mutation is rather rare and only a few subjects present mutations for most codons, it is highly probable that a predicator is sampled with all 0 indicators from the Bootstrap resamples. Consequently, the gram matrix from the Bootstrap resampling procedure is singular, and we cannot obtain Bootstrap estimators.

In the first step, we apply the adaptive Lasso procedure which selects 17 codons; in the second step, our method identifies additional 11 codons associated with drug resistance. Among 28 codons we identified, 13 of them are identified as strong signals and 15 of them as weak signals. Approach in [19] identified 18 codons, where the 13 signals (codon 10, 30, 32, 33, 46, 47, 48, 50, 54, 76, 84, 88 and 90) are the same as our strong-signal codons, and their remaining 5 signals (codon 37, 64, 71, 89 and 93) are among our 15 identified weak signals. In previous studies, [30] identifies all 13 strong signals using a permutation test for the regression coefficients obtained from Lasso; while [15] collect drug resistance mutation information based on multiple research studies, and discover 9 strong-signal codons (10, 32, 46, 47, 50, 54, 76, 84, 90) which are relevant to drug resistance. Neither of these approaches distinguish between strong-signal and weak-signal codons.

Figure 9 presents a graphical summary showing the half-width of the constructed confidence intervals based on our method and the perturbation approach, where strong signals are labeled in blue, and weak signals are labeled in red. To make full comparisons for both strong and weak signals, Figure 9 includes confidence intervals for all selected variables based on our method, even if some of them are not selected by approach in [19]. Table 5 also provides the average half-widths of confidence intervals in each signal category. In general, our method provides shorter lengths of confidence intervals for all strong signals, and longer lengths of confidence intervals for weak signals compared to the perturbation approach. This is not surprising, since the variables with weak coefficients associated with the response variable are relatively weaker, and likely result in wider confidence intervals to ensure a more accurate coverage. These findings are consistent with our simulation studies.

In summary, our approach recovers more codons than other existing approaches. One significance of our method lies in its capability of identifying a pool of strong signals which have strong evidence association with HIV drug resistance, and a pool of possible weak signals which might be mildly associated with drug resistance. In many medical studies, it is important not to miss statistically weak signals, which could be clinically valuable predictors.

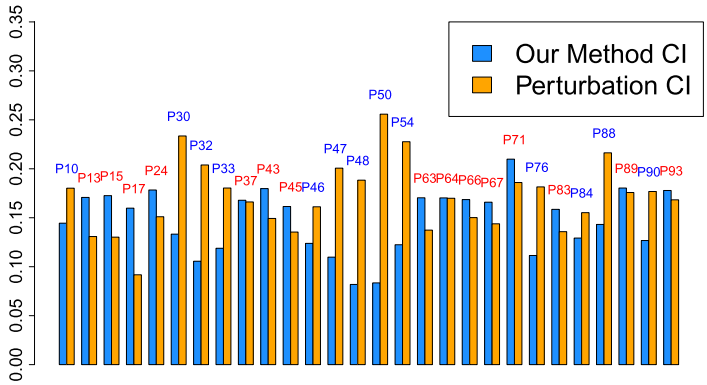

FIG. 9. *Half-width of confidence intervals of selected signals for HIV data.*

6. Summary and discussion. In this paper, we propose weak signal identification under the penalized model selection framework, and develop a new two-step inferential method which is more accurate in providing confidence coverage for weak signal parameters in finite samples. The proposed method is applicable for true models involving both strong and weak signals. The primary concern regarding the existing model selection procedure is that it applies excessive shrinkage in order to achieve model selection consistency. However, this results in low detection power for weak signals in finite samples. The essence of the proposed method is to apply a mild tuning in identifying weak signals. Therefore, there is always a trade-off between a signal's detection power and the false-discovery rate. In our approach, we intend to recover weak signals as much as possible, without sacrificing too much model selection consistency by including too many noise variables.

The two-step inference procedure imposes different selection criteria and confidence interval construction for strong and weak signals. Both theory and numerical studies indicate that the combined approach is more effective compared to the asymptotic inference approach, and bootstrap sampling and other resampling methods. In our numerical studies, we notice that the resampling methods do not provide good inference for weak signals. Specifically, the coverage probability of bootstrap confidence interval is over-covered and exceeds the $(1-\alpha)100\%$ confidence level when the true parameter is close to 0. This is not surprising, as [1]

TABLE 5
Average half-width of the CIs

	All selected variables	**Strong signals**	**Weak signals**
CI^{Our}	0.147	0.118	0.173
CI^{Ptb}	0.171	0.197	0.148

shows that the bootstrap procedure is inconsistent for boundary problems, such as in our case where the boundary parameters are in the order of $1/\sqrt{n}$.

Our method is related to post-selection inference in that we select variables if the corresponding estimated coefficients are not shrunk to zero [2, 17], and then construct confidence intervals for those nonzero coefficients. This is quite different from the hypothesis testing approach which constructs valid confidence intervals for all variables first, then selects variables based on p-values or confidence intervals. One important work among the hypothesis testing approaches is the de-biased Lasso approach [13], which corrects the bias introduced by the Lasso procedure. The de-biased approach constructs valid confidence intervals and p-values for all variables, which is quite powerful when $p > n$ or the gram matrix is singular. However, this approach selects variables through the p-value. In contrast, we select variables if the corresponding estimated coefficients are not shrunk to zero, and construct confidence intervals for those nonzero coefficients. These two approaches are fundamentally different since some coefficients might not be statistically significant; however, the corresponding variables can still contribute to model prediction and should be included in the model. This difference is also reflected in our simulation studies in that the de-biased method has much lower detection rates for true signals in general, and especially when the signals are relatively weak.

In the proposed two-step inference procedure, although we utilize information from the least-square estimators, our approach is very different from applying the least-square inference directly without a model selection step. The nonpenalization method is not feasible when the dimension of covariates is very large, for example, to examine or visualize thousands of conference intervals without model selection. Therefore, it is essential to make a sound statistical inference in conjunction with the variable selection, simultaneously. Our approach has several advantages: (1) It is able to recover possible weak signals which are missed due to excessive shrinkage in model selection, in addition to distinguishing weak signals from strong signals. (2) Our inferences are constructed for selected variable coefficients only. We eliminate noise variables first, and this is different from [13, 19, 27] and [32], which construct CIs for all variables. Consequently, the CI widths we construct for strong signals are much narrower compared to the least squared approach or de-biased method, given that the coverage rates are all accurate. This indicates that our procedure is more effective compared to the approaches which do not perform model selection first. This finding is not surprising since the full model including all the noise variables likely leads to less efficient inference in general. (3) For the weak signal CI's, our numerical studies show that the proposed two-step approach provides CIs comparable to the least square's, but has a much better coverage rate compared to the asymptotic, perturbation and resampling approaches.

In this paper, we develop our method and theory under the orthogonal design assumption. However, our numerical studies indicate that the proposed method is still valid when correlations among covariates are weak or moderate. It would

be interesting to extend the current method to nonorthogonal designed covariates problems.

In addition, it is important to explore weak-signal inference for high-dimensional model settings when the dimension of covariates exceeds the sample size. Note that when the dimension of covariates exceeds the sample size, the least square estimator is no longer feasible and cannot be used as the initial weights for the adaptive Lasso. One possible solution is to replace the full model $\hat{\theta}_{\mathrm{LS}}$ by the marginal regression estimator. In order to do this, we assume that the true model satisfies the partial orthogonality condition, such that the covariates with zero coefficients and those with nonzero coefficients are weakly correlated, and the nonzero coefficients are bounded away from zero at certain rates. Under these assumptions, the estimator of the marginal regression coefficient satisfies the following property, such that the corresponding estimator is not too large for the zero coefficient, and not too small for the nonzero coefficient [11]. This allows us to obtain a reasonable estimator to assign weights in the adaptive Lasso. The same idea has been adopted in [11], where marginal regression estimators are used to assign weights in the adaptive Lasso for sparse high-dimensional data. Huang et al. [11] and [10] show that the adaptive Lasso using the marginal estimator as initial weights yields model selection consistency under the partial orthogonality condition. Alternatively, we can first reduce the model size using the marginal screening approach ([7] and [18]), and then apply our method to the reduced size model. The marginal screening method ensures that we can reduce the model size to be smaller than the sample size, and thus the least square estimator $\hat{\theta}_{\mathrm{LS}}$ can be obtained from a much smaller model.

Finally, the variance estimation of the penalized estimator for weak signal is still very challenging, and worthy of future research. In the proposed method, we use the least-square estimator to provide inference for weak signals. However, the variance of the least-square estimator $\hat{\sigma}_{\mathrm{LS}}$ is inflated when p is close to n. This is likely due to the gram matrix being close to singular when p gets close to n for the least-square estimation. We believe that the de-biased method [13, 27] could be very useful when the gram matrix is singular or close to singular, and it would be interesting to explore a future direction approximating a singular gram matrix to obtain parameter estimation and variance estimation as good as the de-biased method and, therefore, to improve the precision of the confidence intervals for the proposed method.

APPENDIX: NOTATION AND PROOFS

A.1. Notation. $\nu_0 = \sqrt{\lambda}$, $\nu_1 = z_{\tau/2}\frac{\sigma}{\sqrt{n}}$, $\nu_2 = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, $\nu_3 = (z_{\alpha/2} + z_{\tau/2})\frac{\sigma}{\sqrt{n}}$, $\nu_4 = \sqrt{\lambda} + 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

A.2. Tuning parameter selection. BIC criteria function is

$$\mathrm{BIC}(\lambda) = \log(\hat{\sigma}_\lambda^2) + \hat{q}_\lambda \frac{\log(n)}{n},$$

where $\hat{\sigma}_\lambda$ is the estimated standard deviation based on λ, and $\hat{q}_\lambda$ is the number of covariates in the model.

We choose the tuning parameter λ based on the BIC because of its consistency property to select the true model [29]. Here, we follow the strategy in [28] to select the BIC tuning parameter for the adaptive Lasso penalty. Specifically, for a given λ,

$$\mathrm{BIC}(\lambda) = (\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})^T \widehat{\boldsymbol{\Sigma}}_\lambda^{-1} (\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\mathrm{LS}}) + \hat{q}_\lambda \log(n)/n,$$

where $\hat{\boldsymbol{\theta}}_\lambda$ is the adaptive Lasso estimator with a tuning parameter λ provided in (2.1); $\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} = (n\hat{\sigma}^2)^{-1}\{\mathbf{X}^T\mathbf{X} + n\lambda\,\mathrm{diag}\{I(\hat{\theta}_{\lambda,j} \neq 0)/|\hat{\theta}_{\lambda,j}\hat{\theta}_{\mathrm{LS},j}|\}_{j=1}^p\}$; $\hat{\sigma}$ is a consistent estimator of σ based on the scaled Lasso procedure [25]; and $\hat{q}_\lambda$ is the number of nonzero coefficients of $\hat{\boldsymbol{\theta}}_\lambda$, a simple estimator for the degree of freedom [34].

A.3. Proof of Lemma 1. For any γ satisfies $\epsilon < \gamma < 1 - \epsilon$, we show that v^γ that solves $P_d = \gamma$ follows $\frac{v^\gamma}{\sqrt{\lambda_n}} \to 1$, as $n \to \infty$.

P_d can be rewritten as $P_d = \Phi(\frac{\sqrt{n\lambda_n}}{\sigma}(\frac{v^\gamma}{\sqrt{\lambda_n}} - 1)) - \Phi(-\frac{\sqrt{n\lambda_n}}{\sigma}(1 + \frac{v^\gamma}{\sqrt{\lambda_n}}))$. Given $n\lambda_n \to \infty$, if $\lim_{n\to+\infty} \frac{v^\gamma}{\sqrt{\lambda_n}} > 1$, then $P_d(v^\gamma) \to 1$, as $n \to \infty$; else if $\lim_{n\to+\infty} \frac{v^\gamma}{\sqrt{\lambda_n}} < 1$, then $P_d(v^\gamma) \to 0$, as $n \to \infty$. Since $P_d(v^\gamma) = \gamma \in (\epsilon, 1-\epsilon)$, we have $\lim_{n\to+\infty} \frac{v^\gamma}{\sqrt{\lambda_n}} = 1$, as $n \to \infty$.

Therefore, both v^s and v^w satisfy $\frac{v^s}{\sqrt{\lambda_n}} \to 1$ and $\frac{v^w}{\sqrt{\lambda_n}} \to 1$.

A.4. Proof of Lemma 2. Define $\mathrm{CI}_a : \{\theta : |\hat{\theta}_{\mathrm{LS}} - \theta| < z_{\alpha/2}\tilde{\sigma}(\theta)/\sqrt{n}\}$, and $\mathrm{CI}_b : \{\theta : |\hat{\theta}_{\mathrm{LS}} - \theta| < z_{\alpha/2}\sigma/\sqrt{n}\}$. The confidence interval in (4.7) can be rewritten as

$$\mathrm{CI}_a \cdot I_{\{|\hat{\theta}_{\mathrm{LS}}| \geq v_2\}} + \mathrm{CI}_b \cdot I_{\{v_1 < |\hat{\theta}_{\mathrm{LS}}| < v_2\}}.$$

Based on $\mathrm{CI}_a, \mathrm{CI}_b$, we define functions $\mathrm{CR}_a(\theta, v), \mathrm{CR}_b(\theta, v)$ in the following manners:

$$\mathrm{CR}_a(\theta, v) = P\big(\theta \in \mathrm{CI}_a, |\hat{\theta}_{\mathrm{LS}}| > v\big), \tag{A.1}$$

$$\mathrm{CR}_b(\theta, v) = P\big(\theta \in \mathrm{CI}_b, |\hat{\theta}_{\mathrm{LS}}| > v)\sigma/\sqrt{n}\big). \tag{A.2}$$

Besides, define $P_s(\theta, v)$ as $P(|\hat{\theta}_{\mathrm{LS}}| > v)$, which equals

$$P_s(\theta, v) = \Phi\left(\frac{\theta - v}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-\theta - v}{\sigma/\sqrt{n}}\right). \tag{A.3}$$

The explicit expression of $\mathrm{CR}_a(\theta, v)$ is derived based on three cases:

(i) If $\nu < \theta - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, then

$$\mathrm{CR}_a(\theta,\nu) = P\left(|\hat{\theta}^{\mathrm{LS}} - \theta| \le z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right) = 1 - 2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right).$$

(ii) If $|\theta - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}| < \nu < \theta + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, then

$$\mathrm{CR}_a(\theta,\nu) = P\left(\nu < \hat{\theta}^{\mathrm{LS}} < \theta + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right) = \Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right) - \Phi\left(\frac{\nu-\theta}{\sigma/\sqrt{n}}\right).$$

(iii) If $\nu < -\theta + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, then

$$\begin{aligned}\mathrm{CR}_a(\theta,\nu) &= P\left(\theta - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}} < \hat{\theta}^{\mathrm{LS}} < -\nu\right) + P\left(\nu < \hat{\theta}^{\mathrm{LS}} < \theta + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right)\\ &= P_s(\theta,\nu) - 2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right).\end{aligned}$$

The expression of $\mathrm{CR}_b(\theta,\nu)$ can be derived in a similar way. Therefore, $\mathrm{CR}_a(\theta,\nu)$ and $\mathrm{CR}_b(\theta,\nu)$ have the explicit expressions as

$$\mathrm{CR}_a(\theta,\nu) = \begin{cases} \left(P_s(\theta,\nu) - 2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\right) & \\ \quad \times I_{\{\nu < z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\}}, & \text{if } |\theta| < \left|\nu - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right|, \\ \Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right) - \Phi\left(\frac{\nu-\theta}{\sigma/\sqrt{n}}\right), & \\ \quad \text{if } \left|\nu - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\right| \le |\theta| \le \nu + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}, & \\ 1 - 2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right), & \text{if } |\theta| > \nu + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}, \end{cases}$$

and

$$\mathrm{CR}_b(\theta,\nu) = \begin{cases} \left(P_s(\theta,\nu) - \alpha\right)1_{\{\nu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}}, & \text{if } |\theta| < \left|\nu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right|, \\ 1 - \frac{\alpha}{2} - \Phi\left(\frac{\nu-\theta}{\sigma/\sqrt{n}}\right), & \\ \quad \text{if } \left|\nu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right| < |\theta| < \nu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, & \\ 1 - \alpha, & \text{if } |\theta| > \nu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{cases}$$

The equations in (A.1)–(A.3) are used to provide explicit expressions for $\mathrm{CR}_1(\theta)$ and $\mathrm{CR}(\theta)$ in Lemma 2. More specifically,

$$\begin{aligned}\mathrm{CR}_1(\theta) &= P(\theta \text{ in asymptotic-based CI } |\theta \text{ is selected in model selection})\\ &= P\left(|\hat{\theta}_{\mathrm{LS}} - \theta| < z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}\Big||\hat{\theta}_{\mathrm{LS}}| > \sqrt{\lambda}\right)\\ &= \frac{\mathrm{CR}_a(\theta, \nu_0)}{P_s(\theta, \nu_0)},\end{aligned}$$

where $\nu_0 = \sqrt{\lambda}$. Similarly,

$$\begin{aligned}&\mathrm{CR}(\theta)\\ &\quad= P\big(\theta \in \text{ CI as in (4.7) } |\theta \text{ is selected by the two-step procedure}\big)\\ &\quad= \frac{P(\theta \in \mathrm{CI}_a, |\hat{\theta}_{\mathrm{LS}}| \geq \nu_2) + P(\theta \in \mathrm{CI}_b, \nu_1 < |\hat{\theta}_{\mathrm{LS}}| < \nu_2)}{P(|\hat{\theta}_{\mathrm{LS}}| > \nu_1)}\\ &\quad= \frac{P(\theta \in \mathrm{CI}_a, |\hat{\theta}_{\mathrm{LS}}| \geq \nu_2) + P(\theta \in \mathrm{CI}_b, |\hat{\theta}_{\mathrm{LS}}| > \nu_2) - P(\theta \in \mathrm{CI}_b, |\hat{\theta}_{\mathrm{LS}}| > \nu_1)}{P(|\hat{\theta}_{\mathrm{LS}}| > \nu_1)}\\ &\quad= \frac{\mathrm{CR}_a(\theta, \nu_2) + \mathrm{CR}_b(\theta, \nu_1) - \mathrm{CR}_b(\theta, \nu_2)}{P_s(\theta, \nu_1)}.\end{aligned}$$

A.5. Lemmas.

LEMMA 3. *If we select* $\nu_1 = z_{\tau/2}\frac{\sigma}{\sqrt{n}}$, *then the false positive rate of weak signal's identification procedure equals* τ.

PROOF. By definition, the false positive rate equals $P(i \in \widehat{\mathbf{S}}^{(W)} \cup \widehat{\mathbf{S}}^{(S)}|\theta_i = 0) = P(|\hat{\theta}_{\mathrm{LS},i}| > \nu_1|\theta_i = 0) = 2\Phi(-\frac{\sqrt{n}}{\sigma}\nu_1) = \tau$. □

LEMMA 4. *Under conditions* (C1)–(C2), *when* λ *satisfies conditions* $\sqrt{\lambda} > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$:

(a) *if* c_1 *is the solution to* $\theta = \sqrt{\lambda} - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, *then* $c_1 \in ((z_{\alpha/2} - z_{\tau/2})\frac{\sigma}{\sqrt{n}}, \sqrt{\lambda})$;

(b) *if* c_2 *is the solution to* $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, *then* $c_2 \in (\sqrt{\lambda} + \frac{1}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$;

(c) *if* c_3 *is the solution to* $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, *then* $c_3 \in (\sqrt{\lambda}, \sqrt{\lambda} + \frac{1}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$;

(d) *if* c_4 *is the solution to* $\theta = \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, *then* $c_4 \in (\sqrt{\lambda} + \frac{3}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \sqrt{\lambda} + 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.

In addition, the order relationships of c_1, c_2, c_3 *and* c_4 *follow*: $c_1 < c_3 < c_2 < c_4$.

LEMMA 5. *Given* $\theta \in (\sqrt{\lambda} + \frac{1}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \sqrt{\lambda} + 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$, *then* $\theta > c_2$ *if and only if* $\theta > \sqrt{\lambda} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$, *and* $\theta > c_4$ *if and only if* $\theta > \sqrt{\lambda} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$.

LEMMA 6 (Monotonicity of $\mathrm{CR}_1(\theta)$). *Suppose* $\theta > 0$, $\mathrm{CR}_1(\theta)$ *is a piecewise monotonic function on* $[0, c_2]$. *More specifically,* $\mathrm{CR}_1(\theta)$ *is a nondecreasing function on* $[0, c_1]$, *an increasing function on* $[c_1, c_2]$.

LEMMA 7. *For any fixed parameter value* $\nu > 0$, *the function*

$$\frac{\mathrm{CR}_b(\theta,\nu)}{P_s(\theta,\nu)} = \begin{cases} \left(1 - \dfrac{\alpha}{P_s(\theta,\nu)}\right) 1_{\{\nu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}}, & \text{if } |\theta| < \left|\nu - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right|, \\ \dfrac{1 - \frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu - \theta))}{P_s(\theta,\nu)}, & \\ \quad \text{if } \left|\nu - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right| < |\theta| < \nu + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, & \\ \dfrac{1-\alpha}{P_s(\theta,\nu)}, & \text{if } |\theta| > \nu + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \end{cases}$$

is:

(i) *nondecreasing, when* $|\theta| \le |\nu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|$;
(ii) *increasing, when* $|\nu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}| < |\theta| < \nu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$;
(iii) *decreasing, when* $|\theta| \ge \nu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

LEMMA 8. *The formulas for* $\mathrm{CR}_1(\theta)$ *and* $\mathrm{CR}(\theta)$ *in Lemma* 2 *can also be expressed as*

$$\mathrm{CR}_1(\theta) = \frac{\mathrm{CR}_a(\theta,\nu_0)}{P_s(\theta,\nu_0)}, \tag{A.4}$$

$$\mathrm{CR}(\theta) = \begin{cases} \dfrac{\mathrm{CR}_b(\theta,\nu_1)}{P_s(\theta,\nu_1)}, & \text{if } |\theta| < \nu_0, \\ \dfrac{\mathrm{CR}_b(\theta,\nu_1)}{P_s(\theta,\nu_1)} - \dfrac{\mathrm{CR}_b(\theta,\nu_2)}{P_s(\theta,\nu_1)}, & \text{if } \nu_0 \le |\theta| \le c_3, \\ \dfrac{\mathrm{CR}_b(\theta,\nu_1)}{P_s(\theta,\nu_1)} + \dfrac{\mathrm{CR}_a(\theta,\nu_2)}{P_s(\theta,\nu_1)} - \dfrac{\mathrm{CR}_b(\theta,\nu_2)}{P_s(\theta,\nu_1)}, & \text{if } |\theta| > c_3. \end{cases} \tag{A.5}$$

Then $\mathrm{CR}_1(\theta) = J_1(\theta)$, $\mathrm{CR}(\theta) = J_2(\theta) - J_3(\theta) + J_4(\theta)$, *where the four functions* $J_1(\theta)$, $J_2(\theta)$, $J_3(\theta)$ *and* $J_4(\theta)$ *are defined as*

$$J_1(\theta) = \frac{\mathrm{CR}_a(\theta,\nu_0)}{P_s(\theta,\nu_0)}, \qquad J_2(\theta) = \frac{\mathrm{CR}_b(\theta,\nu_1)}{P_s(\theta,\nu_1)},$$

$$J_3(\theta) = \frac{\mathrm{CR}_b(\theta,\nu_2)}{P_s(\theta,\nu_1)}, \qquad J_4(\theta) = \frac{\mathrm{CR}_a(\theta,\nu_2)}{P_s(\theta,\nu_1)}.$$

A.6. Proof of Theorem 1. (a) When $\theta \in [0, c_1]$, we have $\Delta(\theta) \geq 1 - \frac{\alpha}{\tau} > 0$. First, it is obvious that $\mathrm{CR}_1(\theta) = 0$ when $\theta \in [0, c_1]$. By Lemma 7, $\mathrm{CR}(\theta)$ is increasing on $[0, \nu_0]$, and $\mathrm{CR}(\theta) = 1 - \frac{\alpha}{\tau}$ when $\theta = 0$. Thus, $\mathrm{CR}(\theta) - \mathrm{CR}_1(\theta) \geq 1 - \frac{\alpha}{\tau}$ for $\theta \in [0, c_1]$, which provides the first lower bound in Theorem 1. Note that here we also use $c_1 < \nu_0$ by Lemma 4.

(b) When $\theta \in [c_1, \nu_0]$, we have $\Delta(\theta) \geq \frac{2}{1+\alpha} - 2\Phi(\frac{1}{2}z_{\alpha/2}) > 0$. By definition,

$$\mathrm{CR}_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_0 - \theta))}{P_s(\theta, \nu_0)},$$

$$\mathrm{CR}(\theta) = \frac{1 - \frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_1 - \theta))}{P_s(\theta, \nu_1)}.$$

In the following, we show that $\frac{\partial \mathrm{CR}_1(\theta)}{\partial \theta} > \frac{\partial \mathrm{CR}(\theta)}{\partial \theta}$, so $\mathrm{CR}(\theta) - \mathrm{CR}_1(\theta)$ is decreasing when $\theta \in [c_1, \nu_0]$. The first-order derivatives of $\mathrm{CR}_1(\theta)$ and $\mathrm{CR}(\theta)$ are

$$\frac{\partial \mathrm{CR}_1(\theta)}{\partial \theta} = \left[\frac{z_{\alpha/2}}{\sigma}\phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\tilde{\sigma}(\theta)' + \frac{\sqrt{n}}{\sigma}\phi\left(\frac{\sqrt{n}}{\sigma}(\nu_0 - \theta)\right)\right]P_s(\theta, \nu_0)^{-1} \tag{A.6}$$

$$- \left[\Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\nu_0 - \theta)\right)\right]P_s(\theta, \nu_0)^{-2} \tag{A.7}$$

$$\frac{\partial \mathrm{CR}(\theta)}{\partial \theta} = \frac{\sqrt{n}}{\sigma}\phi\left(\frac{\sqrt{n}}{\sigma}(\nu_1 - \theta)\right)P_s(\theta, \nu_1)^{-1} \tag{A.8}$$

$$- \left[1 - \frac{\alpha}{2} - \Phi\left(\frac{\sqrt{n}}{\sigma}(\nu_1 - \theta)\right)\right]P_s(\theta, \nu_1)^{-2}, \tag{A.9}$$

where each first-order derivative is composed of two parts. We show the inequality of each part separately. First, (A.6) > (A.8), which is sufficient by showing

$$\phi\left(\frac{\sqrt{n}}{\sigma}(\nu_0 - \theta)\right)P_s(\theta, \nu_0)^{-1} > \phi\left(\frac{\sqrt{n}}{\sigma}(\nu_1 - \theta)\right)P_s(\theta, \nu_1)^{-1}.$$

This is equivalent to show

$$\frac{\Phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta + \nu_1))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_1))} > \frac{\Phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta + \nu_0))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_0))}. \tag{A.10}$$

The inequality in (A.10) can be proved based on monotinicity of two functions $\frac{\Phi(x)}{\phi(x)}$ and $\frac{\Phi(-x-y)}{\phi(x-y)}$. Specifically, it can be shown that $\frac{\Phi(x)}{\phi(x)}$ is an increasing function of $x \in \mathbb{R}$, and $\frac{\Phi(-x-y)}{\phi(x-y)}$ is a decreasing function of $y \in \mathbb{R}^+$, for any fixed value of

$x > 0$. More specifically, since $\nu_1 < \nu_0$, we have

$$\frac{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))} > \frac{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0))} \quad \text{and}$$

$$\frac{\Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))} > \frac{\Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))}{\phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0))},$$

based on which the inequality in (A.10) holds.

Next, we show that (A.8) < (A.9), which is equivalent with

$$\left[1-\frac{\alpha}{2}-\Phi\left(\frac{\sqrt{n}}{\sigma}(\nu_1-\theta)\right)\right]P_s(\theta,\nu_1)^{-2} > \left[\Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)-\Phi\left(\frac{\sqrt{n}}{\sigma}(\nu_1-\theta)\right)\right]P_s(\theta,\nu_0)^{-2}.$$

It can be shown by

$$\frac{1-\frac{\alpha}{2}-\Phi(\frac{\sqrt{n}}{\sigma}(\nu_1-\theta))}{P_s(\theta,\nu_1)^2} > \frac{1-\frac{\alpha}{2}-\Phi(\frac{\sqrt{n}}{\sigma}(\nu_0-\theta))}{P_s(\theta,\nu_0)^2} > \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})-\Phi(\frac{\sqrt{n}}{\sigma}(\nu_1-\theta))}{P_s(\theta,\nu_0)^2}.$$

Based on the above arguments, we can conclude that for $\theta \in [c_1, \nu_0]$:

$$\frac{\partial \mathrm{CR}_1(\theta)}{\partial\theta} > \frac{\partial \mathrm{CR}(\theta)}{\partial\theta}.$$

Therefore, $\min_{\theta\in[c_1,\nu_0]}\Delta(\theta) = \mathrm{CR}(\nu_0) - \mathrm{CR}_1(\nu_0)$. More specifically,

$$\mathrm{CR}(\nu_0) = \frac{\Phi(\frac{\nu_0-\nu_1}{\sigma/\sqrt{n}})-\frac{\alpha}{2}}{\Phi(\frac{\nu_0-\nu_1}{\sigma/\sqrt{n}})+\Phi(\frac{-\nu_0-\nu_1}{\sigma/\sqrt{n}})} > \frac{1-\alpha}{1+\alpha},$$

$$\mathrm{CR}_1(\nu_0) = \frac{\Phi(\frac{1}{2}z_{\alpha/2})-\frac{\alpha}{2}}{\frac{1}{2}+\Phi(-\frac{2\nu_0}{\sigma/\sqrt{n}})} < 2\Phi\left(\frac{1}{2}z_{\alpha/2}\right)-1,$$

thus,

$$\mathrm{CR}(\nu_0)-\mathrm{CR}_1(\nu_0) > \frac{2}{1+\alpha}-2\Phi\left(\frac{1}{2}z_{\alpha/2}\right),$$

which provides the second lower bound in Theorem 1.

(c) When $\theta \in [\nu_0, +\infty)$, we have $\Delta(\theta)$ satisfies either $\Delta(\theta) \geq 0$ or $-\frac{\alpha}{2} < \Delta(\theta) < 0$. The proof of case 1 is provided here, and proof of the other two cases are similar and are provided in supplementary materials. In case 1, it satisfies $c_3 <$

$\nu_3 < c_2$. We conduct derivations for sub-intervals $[\nu_0, c_3], [c_3, \nu_3], [\nu_3, c_2], [c_2, c_4], [c_4, \nu_4]$ and $[\nu_4, +\infty)$, separately.

When $\theta \in [\nu_0, c_3]$, we have

$$J_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_0}{\sigma/\sqrt{n}})},$$

$$J_2(\theta) = \frac{1-\frac{\alpha}{2} - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})}, \quad \text{and}$$

$$J_3(\theta) = \frac{1-\frac{\alpha}{2} - \Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})},$$

thus

$$\begin{aligned}\Delta(\theta) &= J_2(\theta) - J_1(\theta) - J_3(\theta)\\ &= \frac{\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_0}{\sigma/\sqrt{n}})}.\end{aligned}$$

Further,

$$\begin{aligned}\Delta(\theta) &= \frac{\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_0}{\sigma/\sqrt{n}})}\\ &> \frac{\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\\ &> \frac{\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\\ &= \frac{[\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) - [\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\\ &\quad - \frac{[\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})]\Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\\ &= \Delta_1(\theta) - \Delta_2(\theta),\end{aligned}$$

where the second inequality uses that $z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}} \leq \nu_2 - \theta$ when $\theta \leq c_3$ by Lemma 5. Here, $\Delta_1(\theta)$ and $\Delta_2(\theta)$ are defined as

$$\Delta_1(\theta) = \frac{[\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) - [\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})},$$

$$\Delta_2(\theta) = \frac{[\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})]}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right).$$

First, it is straightforward to show that $\Delta_1(\theta) > 0$. Second, $\Delta_2(\theta)$ can be bounded from above by some small value. In fact,

$$\begin{aligned}\Delta_2(\theta) &< \frac{[\Phi(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})]}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})}\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right) \\ &< 4\left[\Phi\left(\frac{\nu_2-\theta}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\nu_0-\theta}{\sigma/\sqrt{n}}\right)\right]\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right) \\ &< 4\left[1-\frac{\alpha}{2} - \Phi\left(-\frac{1}{2}z_{\alpha/2}\right)\right]\Phi\left(-\frac{3}{2}z_{\alpha/2}\right),\end{aligned}$$

where we use that $\nu_2 - \theta < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, $-\frac{1}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \nu_0 - c_3 < \nu_0 - \theta$ and $\Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}) < \Phi(\frac{-\nu_0-\nu_1}{\sigma/\sqrt{n}}) < \Phi(-\frac{3}{2}z_{\alpha/2})$ when $\nu_0 < \theta < c_3$. Combining the lower bounds for $\Delta_1(\theta)$ and $\Delta_2(\theta)$, we have

$$\Delta(\theta) > -4\left[1-\frac{\alpha}{2} - \Phi\left(-\frac{1}{2}z_{\alpha/2}\right)\right]\Phi\left(-\frac{3}{2}z_{\alpha/2}\right).$$

In fact, the lower bound on the right-hand side is quite close to zero.

When $\theta \in [c_3, \nu_3]$,

$$\Delta(\theta) = J_2(\theta) - J_1(\theta) - J_3(\theta) + J_4(\theta),$$

where

$$J_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_0-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))},$$

$$J_2(\theta) = \frac{1-\frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_1-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$

$$J_3(\theta) = \frac{1-\frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_3-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$

$$J_4(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_3-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}.$$

Therefore,

$$\begin{aligned}
\Delta(\theta) &= \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_1-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0-\theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_0}{\sigma/\sqrt{n}})} \\
&= \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_0}{\sigma/\sqrt{n}})} \\
&> \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})} \\
&\quad + \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})} \\
&= \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})} \cdot \left(-\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right)\right) \\
&\quad + \Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\left[\frac{1}{\Phi(\frac{\theta-\nu_0}{\sigma/\sqrt{n}})} - \frac{1}{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})}\right] \\
&> \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}})} \cdot \left(-\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right)\right) \\
&> -2\Phi\left(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}}\right) > -2\Phi\left(-\frac{3}{2}z_{\alpha/2}\right),
\end{aligned}$$

the second inequality holds since

$$\Phi\left(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}\right) > \frac{1}{2}, \qquad 0 < \frac{\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{[\Phi(\frac{\theta-\nu_1}{\sigma/\sqrt{n}}) + \Phi(\frac{-\theta-\nu_1}{\sigma/\sqrt{n}})]} < 1,$$

and the last inequality holds since $-\theta-\nu_1 < -(z_{\alpha/2}+z_{\tau/2})\frac{\sigma}{\sqrt{n}} < -\frac{3}{2}z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, when $\theta \geq c_3 > \sqrt{\lambda} \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

When $\theta \in [\nu_3, c_2]$, $\Delta(\theta) = J_2(\theta) - J_1(\theta) + J_4(\theta) - J_3(\theta)$, where

$$J_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_0-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))},$$

$$J_2(\theta) = \frac{1-\alpha}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$

$$J_3(\theta) = \frac{1 - \frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_2 - \theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta + \nu_1))},$$

$$J_4(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_2 - \theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta - \nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta + \nu_1))}.$$

Therefore,

$$\Delta(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{P_s(\theta, \nu_1)} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_0 - \theta))}{P_s(\theta, \nu_0)}.$$

Further, we have $\Delta(\theta) > \Delta_1(\theta) + \Delta_2(\theta)$, where

$$\Delta_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{P_s(\theta, \nu_1)} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})}, \quad \text{and}$$

$$\Delta_2(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\nu_0 - \theta}{\sigma/\sqrt{n}})}{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}})}.$$

It is straightforward to get a bound for $\Delta_1(\theta)$. In fact,

$$\Delta_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{P_s(\theta, \nu_1)\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} \cdot \left[-\Phi\left(\frac{-\theta - \nu_1}{\sigma/\sqrt{n}}\right)\right],$$

here $\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2} < 1 - \alpha$, $P_s(\theta, \nu_1) > \Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}}) > 1 - \frac{\alpha}{2}$, and $-\theta - \nu_1 < -\nu_3 - \nu_1 < -2z_{\alpha/2}\sigma/\sqrt{n}$. Therefore, $\Delta_1(\theta) < 0$ and $|\Delta_1(\theta)| < \frac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2})$.

It takes a few more steps to bound $\Delta_2(\theta)$. In fact,

$$\begin{aligned}
\Delta_2(\theta) &= \frac{1 - \frac{\alpha}{2} - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} - \frac{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}})} \\
&= \left[\frac{1 - \frac{\alpha}{2}}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} - 1\right] \\
&\quad + \Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\left[\frac{1}{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}})} - \frac{1}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})}\right] \\
&> \left[\frac{1 - \frac{\alpha}{2}}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} - 1\right] + \frac{\alpha}{2}\left[\frac{1}{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}})} - \frac{1}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})}\right],
\end{aligned}$$

the inequality holds since $\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) > \frac{\alpha}{2}$. It can also be shown that both $\frac{1}{\Phi(\frac{\theta - \nu_0}{\sigma/\sqrt{n}})} - \frac{1}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})}$ and $\frac{1 - \frac{\alpha}{2}}{\Phi(\frac{\theta - \nu_1}{\sigma/\sqrt{n}})} - 1$ are decreasing functions of θ, given $\theta >$

$\nu_0 > \nu_1$. Therefore,

$$\Delta_2(\theta) > \left[\frac{1-\frac{\alpha}{2}}{\Phi(\frac{\nu_2-\nu_1}{\sigma/\sqrt{n}})} - 1\right] + \frac{\alpha}{2}\left[\frac{1}{\Phi(\frac{\nu_2-\nu_0}{\sigma/\sqrt{n}})} - \frac{1}{\Phi(\frac{\nu_2-\nu_1}{\sigma/\sqrt{n}})}\right]$$
$$= \frac{1-\alpha}{\Phi(\frac{\nu_2-\nu_1}{\sigma/\sqrt{n}})} - \frac{1-\alpha}{1-\frac{\alpha}{2}} > -\frac{\alpha(1-\alpha)}{2-\alpha}.$$

Combining the lower bounds of $\Delta_1(\theta)$ and $\Delta_2(\theta)$, the lower bound for $\Delta(\theta)$ is provided by

$$\Delta(\theta) > -\frac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2}) - \frac{\alpha(1-\alpha)}{2-\alpha} > -\frac{\alpha}{2}.$$

When $\theta \in [c_2, c_4]$,

$$J_1(\theta) = \frac{1 - 2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))},$$
$$J_2(\theta) = \frac{1-\alpha}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$
$$J_3(\theta) = \frac{1-\frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_2-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$
$$J_4(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_2-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}.$$

Therefore,

$$\Delta(\theta) = J_2(\theta) - J_1(\theta) + J_4(\theta) - J_3(\theta)$$
$$= \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{P_s(\theta,\nu_1)} - \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{P_s(\theta,\nu_0)}.$$

Again, $\Delta(\theta) > \Delta_1(\theta) + \Delta_2(\theta)$, where

$$\Delta_1(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{P_s(\theta,\nu_1)\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))} \cdot \left[-\Phi\left(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1)\right)\right], \quad \text{and}$$
$$\Delta_2(\theta) = \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))} - \frac{2\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - 1}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0))}.$$

First,

$$\Delta_1(\theta) < 0 \quad \text{and} \quad |\Delta_1(\theta)| < \frac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2}),$$

which holds true because $\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2} < 1-\alpha$, $P_s(\theta, \nu_1) > \Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) > 1-\frac{\alpha}{2}$, $\frac{1}{2}z_{\alpha/2} < z_{\tau/2}$, and $-\nu_3 - \nu_1 = -(z_{\alpha/2}+2z_{\tau/2})\frac{\sigma}{\sqrt{n}} < -2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Second, when $\theta > c_2$, it holds that $\theta > \nu_0 + z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sqrt{n}}$ according to Lemma 5, and further $\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) > \Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})$. Therefore,

$$\begin{aligned}
\Delta_2(\theta) &> \frac{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \frac{\alpha}{2}}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1))} - \frac{2\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - 1}{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})} \\
&> \Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right) - \frac{\alpha}{2} - \frac{2\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - 1}{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})} \\
&= \Phi\left(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right) + \frac{1}{\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})} \\
&\quad - \frac{\alpha}{2} - 2.
\end{aligned}$$

The function on the right-hand side is a decreasing function of $\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})$. Given that $\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) < 1-\frac{\alpha}{2}$, we have

$$\Delta_2(\theta) > 1 - \frac{\alpha}{2} - \frac{1}{1-\frac{\alpha}{2}} - \frac{\alpha}{2} - 2 = -\frac{\alpha(1-\alpha)}{2-\alpha}.$$

Combining the lower bounds for $\Delta_1(\theta)$ and $\Delta_2(\theta)$, we have

$$\Delta(\theta) > -\frac{4(1-\alpha)}{(2-\alpha)^2}\Phi(-2z_{\alpha/2}) - \frac{\alpha(1-\alpha)}{2-\alpha}. \tag{A.11}$$

This lower bound for $\Delta(\theta)$ is exactly the same with that in the interval $[\nu_3, c_2]$.

When $\theta \in [c_4, \nu_4]$,

$$\begin{aligned}
J_1(\theta) &= \frac{1 - 2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))}, \\
J_2(\theta) &= \frac{1-\alpha}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}, \\
J_3(\theta) &= \frac{1-\frac{\alpha}{2} - \Phi(\frac{\sqrt{n}}{\sigma}(\nu_2-\theta))}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}, \\
J_4(\theta) &= \frac{1 - 2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}.
\end{aligned}$$

Therefore,

$$\Delta(\theta) = \frac{\Phi(\frac{\sqrt{n}}{\sigma}(\nu_2-\theta)) - \frac{\alpha}{2}}{P_s(\theta,\nu_1)} + \left[1-2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\right]\left[\frac{1}{P_s(\theta,\nu_1)} - \frac{1}{P_s(\theta,\nu_0)}\right]$$
$$> \left[1-2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\right]\left[\frac{1}{P_s(\theta,\nu_1)} - \frac{1}{P_s(\theta,\nu_0)}\right],$$

the inequality holds since $\nu_2 - \theta \geq \nu_2 - \nu_4 = -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ when $\theta \leq \nu_4$.

Let $\Delta_1(\theta) = [1-2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})][\frac{1}{P_s(\theta,\nu_1)} - \frac{1}{P_s(\theta,\nu_0)}]$. We show that $\Delta_1(\theta)$ is negative but quite close to zero. When $\theta > c_4$, $P_s(\theta,\nu_1) > P_s(\theta,\nu_0) > \Phi(\frac{3}{2}z_{\alpha/2})$, and further $P_s(\theta,\nu_1) - P_s(\theta,\nu_0) \in (0, \Phi(-\frac{3}{2}z_{\alpha/2}))$. Therefore,

$$0 < \frac{1}{P_s(\theta,\nu_0)} - \frac{1}{P_s(\theta,\nu_1)} = \frac{P_s(\theta,\nu_1) - P_s(\theta,\nu_0)}{P_s(\theta,\nu_1)P_s(\theta,\nu_0)} < \frac{\Phi(-\frac{3}{2}z_{\alpha/2})}{\Phi(\frac{3}{2}z_{\alpha/2})^2},$$

together with $\Phi(z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) - \Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}) < 1-\alpha$, we have

$$\Delta_1(\theta) < 0 \quad \text{and} \quad \left|\Delta_1(\theta)\right| < (1-\alpha)\frac{\Phi(-\frac{3}{2}z_{\alpha/2})}{\Phi(\frac{3}{2}z_{\alpha/2})^2}.$$

Therefore,

$$\Delta(\theta) > -(1-\alpha)\frac{\Phi(-\frac{3}{2}z_{\alpha/2})}{\Phi(\frac{3}{2}z_{\alpha/2})^2}.$$

When $\theta \in [\nu_4, +\infty)$,

$$J_1(\theta) = \frac{1-2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_0)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_0))},$$
$$J_2(\theta) = \frac{1-\alpha}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$
$$J_3(\theta) = \frac{1-\alpha}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))},$$
$$J_4(\theta) = \frac{1-2\Phi(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma})}{\Phi(\frac{\sqrt{n}}{\sigma}(\theta-\nu_1)) + \Phi(-\frac{\sqrt{n}}{\sigma}(\theta+\nu_1))}.$$

Therefore,

$$\Delta(\theta) = \left[1-2\Phi\left(-z_{\alpha/2}\frac{\tilde{\sigma}(\theta)}{\sigma}\right)\right]\left[\frac{1}{P_s(\theta,\nu_1)} - \frac{1}{P_s(\theta,\nu_0)}\right].$$

Here, $\Delta(\theta) < 0$, and

$$\Delta(\theta) > \frac{1}{P_s(\theta, \nu_1)} - \frac{1}{P_s(\theta, \nu_0)} > -\frac{\Phi(-2z_{\alpha/2})}{\Phi(2z_{\alpha/2})},$$

where we use that $\theta - \nu_0 > 2z_\alpha \frac{\sigma}{\sqrt{n}}$ when $\theta > \nu_4$. In fact, $P_s(\theta, \nu_1) \approx P_s(\theta, \nu_0)$ when θ gets quite large, thus $\Delta(\theta) \approx 0$. The proof of case 1 in Theorem 1 is completed.

Acknowledgments. The authors thank the Co-Editor, an Associate Editor and three referees for their constructive comments that have substantially improved an earlier version of this paper. Part of the work has been done during Dr. Shi's postdoc training supported by Professor Yi Li at University of Michigan. The authors would also like to thank Jessica Minnier and Lu Tian for generously sharing their codes on the perturbation methods and the HIV dataset.

SUPPLEMENTARY MATERIAL

Supplement to "Weak signal identification and inference in penalized model selection" (DOI: 10.1214/16-AOS1482SUPP; .pdf). Due to space constraints, we relegate technical details of the remaining proofs to the supplement.

REFERENCES

[1] ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

[2] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

[3] DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

[4] EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

[5] EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. Chapman & Hall, New York. MR1270903

[6] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[7] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[8] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[9] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

[10] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

[11] HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

[12] HUANG, J. and XIE, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. In *Asymptotics*: *Particles*, *Processes and Inverse Problems*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 149–166. IMS, Beachwood, OH. MR2459937

[13] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[14] JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749
[15] JOHNSON, V. A., BRUN-VÉZINET, F., CLOTET, B., KURITZKES, D. R., PILLAY, D., SCHAPIRO, J. M. and RICHMAN, D. D. (2006). Update of the drug resistance mutations in HIV-1: Fall 2006. *Top HIV Med* **14** 125–130.
[16] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787
[17] LEE, J. D., SUN, D. L., SUN, Y.and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948
[18] LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900
[19] MINNIER, J., TIAN, L. and CAI, T. (2011). A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.* **106** 1371–1382. MR2896842
[20] PÖTSCHER, B. M. and LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.* **100** 2065–2082. MR2543087
[21] PÖTSCHER, B. M. and SCHNEIDER, U. (2009). On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference* **139** 2775–2790. MR2523666
[22] PÖTSCHER, B. M. and SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4** 334–360. MR2645488
[23] SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354
[24] SHI, and QU, (2016). Supplement to "Weak signal identification and inference in penalized model selection." DOI:10.1214/16-AOS1482SUPP.
[25] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166
[26] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[27] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285
[28] WANG, H. and LENG, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102** 1039–1048. MR2411663
[29] WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008
[30] WU, M. (2009). A parametric permutation test for regression coefficients in Lasso regularized regression. Ph.D thesis, Dept. Biostatistics, Harvard School of Public Health, Boston, MA.
[31] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701
[32] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940
[33] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469
[34] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967
[35] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: pshi@umich.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: anniequ@illinois.edu