Statistica Sinica Preprint No: SS-2020-0449	
Title	Weak Signal Identification and Inference in Penalized Likelihood Models for Categorical Responses
Manuscript ID	SS-2020-0449
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0449
Complete List of Authors	Yuexia Zhang, Peibei Shi, Zhongyi Zhu, Linbo Wang and Annie Qu
Corresponding Author	Annie Qu
E-mail	aqu2@uci.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

WEAK SIGNAL IDENTIFICATION AND INFERENCE IN PENALIZED LIKELIHOOD MODELS FOR CATEGORICAL RESPONSES

[1]Yuexia Zhang, [2]Peibei Shi, [3]Zhongyi Zhu, [4]Linbo Wang and [5]Annie Qu

[1,4]*University of Toronto,* [2]*Facebook,* [3]*Fudan University,* [5]*University of California, Irvine*

Abstract:

Penalized likelihood models are widely used to simultaneously select variables and estimate model parameters. However, the existence of weak signals can lead to inaccurate variable selection, biased parameter estimation, and invalid inference. Thus, identifying weak signals accurately and making valid inferences are crucial in penalized likelihood models. We develop a unified approach to identify weak signals and make inferences in penalized likelihood models, including the special case when the responses are categorical. To identify weak signals, we use the estimated selection probability of each covariate as a measure of the signal strength and formulate a signal identification criterion. To construct confidence intervals, we propose a two-step inference procedure. Extensive simulation studies show that the proposed procedure outperforms several existing methods. We illustrate the proposed method by applying it to the Practice Fusion diabetes data set.

Key words and phrases: adaptive lasso, de-biased method, model selection, post-selection inference

1. Introduction

In the big data era, massive data are collected with large-dimensional covariates. However, only some of the covariates might be important. To select the important variables and estimate their effects on the response variable, various penalized likelihood models have been proposed, such as the penalized least squares regression model (Tibshirani, 1996; Zou and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Hastie, 2005; Tibshirani et al., 2005; Yuan and Lin, 2006; Zou, 2006; Zhang, 2010), penalized logistic regression model (Park and Hastie, 2008; Zhu and Hastie, 2004; Wu et al., 2009), and penalized Poisson regression model (Lambert and Eilers, 2005; Jia et al., 2019).

To achieve model selection consistency or the variable screening property for a high-dimensional problem, a common condition is the "beta-min" condition, which requires the nonzero regression coefficients to be sufficiently large (Zhao and Yu, 2006; Huang and Xie, 2007; Van de Geer et al., 2011; Tibshirani, 2011; Zhang and Jia, 2017). Therefore, classical methods for variable selection often focus on strong signals that satisfy such a condition. However, if the "beta-min" condition is violated, the important variables and unimportant variables may be inseparable, and the true important variables might not be selected, even if the sample size goes to infinity (Zhang, 2013). In finite samples, the estimators shrink the true regression coefficients, owing to the penalty function. When the signal strength is weak, its coefficient is more likely to shrink to zero (Shi and Qu, 2017; Liu et al., 2020). Inaccurate variable selection and biased parameter estimation could lead to a poor post-selection inference, for example, the estimation of the confidence intervals could be inaccurate. Thus, both strong and weak signals need to be considered. Identification and inference for weak signals can also help discover potentially important variables in practice. For example, in genome-wide association studies (GWAS), overlooked risk factors for a disease may be recovered by incorporating weak signals (Liu et al., 2020).

For linear regression models, studies have been done on weak signals. In more extreme cases, Jin et al. (2014) assumed all signals were individually weak and proposed graphlet screening for variable selection. Zhang (2017) proposed the perturbed lasso, where signals were strengthened by adding random perturbations to the design matrix. However, these

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

methods focused only on variable selection consistency, and did not aim to identify weak signals or provide statistical inference. For weak signal identification and inference, Shi and Qu (2017) proposed a weak signal identification procedure in finite samples, and introduced a two-step inference method for constructing confidence intervals after signal identification. However, their derivation relies on a crucial assumption that the design matrix is orthogonal, which may not hold in practice. On the other hand, Li et al. (2019) took advantage of the correlations between covariates, detecting weak signals through the partial correlations between strong and weak signals. However, they did not study weak signal inference. Recently, Liu et al. (2020) proposed a method that combines the bootstrap lasso and a partial ridge regression for constructing confidence intervals when there are weak signals in the covariates. However, as stated in their paper, the confidence intervals of the coefficients, with magnitudes of order $1/\sqrt{n}$, may be invalid.

To the best of our knowledge, there has been little work on weak signals in likelihood-based models for categorical responses. One exception is Reangsephet et al. (2020), who proposed variable selection methods for logistic regression models with weak signals. However, they did not conduct weak signal identification or inference.

We address these gaps by developing a new unified approach to weak signal identification and inference in penalized likelihood models, including the special case when the responses are categorical. Specifically, the estimated probability of each covariate being selected by the one-step adaptive lasso estimator is used to measure the signal strength. After signal identification, a two-step inference procedure is proposed for constructing the confidence intervals for the regression coefficients. The proposed method has several advantages. First, we extend the method of Shi and Qu (2017) from linear regression models to likelihood-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

based models, including generalized linear models. However, our extension is not trivial. For example, in Shi and Qu (2017), the selection probability has an explicit expression. For the proposed likelihood-based method, such an explicit expression does not exist for categorical responses. Thus, we propose a new method to estimate the selection probability. Second, in Shi and Qu (2017), the selection probability for the covariate $\boldsymbol{X}_j$ is an increasing function of $|\beta_{j0}|$, where β_{j0} is the corresponding coefficient of $\boldsymbol{X}_j$. Under our current general framework, such a conclusion is not necessarily true. Thus, our signal identification criterion is based directly on the estimated selection probability, in contrast to Shi and Qu (2017). We also discuss how each signal's selection probability is influenced by other covariates, owing to nonlinear modeling or collinearity among the covariates; in Shi and Qu (2017), the selection probability of one covariate is independent of those of other covariates. Third, Shi and Qu (2017) assumed that the design matrix in a linear regression model is orthogonal, whereas the proposed method relaxes this constraint. Fourth, the proposed inference method differs from that of Shi and Qu (2017). Specifically, we construct confidence intervals for the noise variables as well, whereas their method does not. Simulation results show that our proposed two-step inference method outperforms the two-step inference method based on Shi and Qu (2017). In particular, the proposed confidence intervals achieve accurate coverage probabilities for all signal strength levels.

The remainder of this paper is organized as follows. In Section 2, we introduce the one-step adaptive lasso estimator and derive the variable selection condition. In Section 3, we propose the weak signal identification criterion. In Section 4, we develop a two-step inference procedure for constructing confidence intervals. In Section 5, we conduct simulation studies to assess the finite-sample performance of the proposed method. In Section 6, we apply the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

proposed method to an analysis of diabetes data. In Section 7, we provide brief concluding remarks. We provide the technical proofs, implementation details of several methods, and some additional results in the Supplementary Material.

2. One-step adaptive lasso estimator and variable selection condition

In this section, we introduce the one-step penalized likelihood estimator and derive the condition for variable selection, which we use later for weak signal identification and inference.

Let $(\mathbf{x}_1^\top, y_1)^\top, \ldots, (\mathbf{x}_n^\top, y_n)^\top$ be n independent and identically distributed (i.i.d.) random vectors, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a $p \times 1$ vector of predictors and y_i is a response variable. Assume that y_i depends on $\mathbf{x}_i$ through a linear combination $\mathbf{x}_i^\top \boldsymbol{\beta}_0$, and the conditional log-likelihood of y_i given $\mathbf{x}_i$ is $\ell_i(\boldsymbol{\gamma}_0) = \ell_i(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0, y_i)$, where $\boldsymbol{\gamma}_0 = (\alpha_0, \boldsymbol{\beta}_0^\top)^\top$, α_0 is an unknown true location parameter, and $\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{p0})^\top$ is an unknown $p \times 1$ vector of covariate effects. Note that for a likelihood-based model, it is not always possible to eliminate the location parameter by centering the covariates and the response variable. For simplicity, assume $p < n$ and p is fixed. Let $\ell(\boldsymbol{\gamma}) = \sum_{i=1}^n \ell_i(\boldsymbol{\gamma})$ denote the log-likelihood. Assume $\boldsymbol{\gamma}^{(0)}$ is the maximum likelihood estimator of $\boldsymbol{\gamma}_0$; then, $\boldsymbol{\gamma}^{(0)} = (\alpha^{(0)}, \boldsymbol{\beta}^{(0)\top})^\top = \operatorname{argmax}_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma})$. In matrix notation, we set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, with $\boldsymbol{X}_j = (x_{1j}, \ldots, x_{nj})^\top$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$. Furthermore, denote $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i^\top)^\top$ and $\widetilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X})$, where $\mathbf{1}$ is an $n \times 1$ vector with all elements equal to one. Throughout this paper, we assume that $\mathrm{E}(x_{ij}) = 0$ and $\mathrm{Var}(x_{ij}) = 1$, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, which can be realized by standardizing the covariate matrix $\mathbf{X}$, in practice.

Assume that some components of $\boldsymbol{\beta}_0$ are zero. In order to estimate the model parameters and select important variables simultaneously, we consider the penalized likelihood function

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

$\ell(\boldsymbol{\gamma})/n - \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|)$, where $p_{\lambda_j}(\cdot)$ is a penalty function controlled by the tuning parameter λ_j. One popular penalty function is derived from the adaptive lasso estimator (Zou, 2006), where $p_{\lambda_j}(|\beta_j|) = \lambda|\beta_j|/|\beta_j^{(0)}|$. Maximizing the penalized likelihood function is equivalent to minimizing

$$-\frac{1}{n}\ell(\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \tag{2.1}$$

with respect to $\boldsymbol{\gamma}$. According to Wang and Leng (2007) and Zou and Li (2008), if the log-likelihood function has first and second derivatives, then it can be approximated by a Taylor expansion. Furthermore, the objective function (2.1) can be approximated by

$$Q_1(\boldsymbol{\gamma}) = -\frac{1}{2n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)})^\top \ddot{\ell}(\boldsymbol{\gamma}^{(0)})(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{2.2}$$

where $\ddot{\ell}(\cdot)$ is the second derivative of function $\ell(\cdot)$. The one-step penalized likelihood estimator is $\boldsymbol{\gamma}^{(1)} = (\alpha^{(1)}, \boldsymbol{\beta}^{(1)\top})^\top = \text{argmin}_{\boldsymbol{\gamma}} Q_1(\boldsymbol{\gamma})$.

Denote $\mu_i(\boldsymbol{\gamma}) = \mu_i = \widetilde{\mathbf{x}}_i^\top \boldsymbol{\gamma}$ and $\ell_i\{\mu_i(\boldsymbol{\gamma})\} = \ell_i(\widetilde{\mathbf{x}}_i^\top \boldsymbol{\gamma}, y_i)$. Let $\mathbf{D}(\boldsymbol{\gamma})$ be an $n \times n$ diagonal matrix with the (i,i)th element $D_{ii}(\boldsymbol{\gamma}) = -\partial^2 \ell_i\{\mu_i(\boldsymbol{\gamma})\}/\partial \mu_i^2$, for $i = 1, \ldots, n$. Then, $\ddot{\ell}(\boldsymbol{\gamma}) = -\widetilde{\mathbf{X}}^\top \mathbf{D}(\boldsymbol{\gamma})\widetilde{\mathbf{X}}$. Furthermore, we assume $D_{ii}(\boldsymbol{\gamma})$ is a continuous function of $\boldsymbol{\gamma}$. For simplicity, denote $\mathbf{D}(\boldsymbol{\gamma}^{(0)})$, $\mathbf{D}(\boldsymbol{\gamma}_0)$, $D_{ii}(\boldsymbol{\gamma}^{(0)})$, and $D_{ii}(\boldsymbol{\gamma}_0)$ as $\mathbf{D}^{(0)}$, $\mathbf{D}_0$, $D_{ii}^{(0)}$, and $D_{0,ii}$, respectively. By solving the equation $\partial Q_1(\boldsymbol{\gamma})/\partial \alpha = 0$, we obtain that

$$\alpha - \alpha^{(0)} = (\mathbf{1}^\top \mathbf{D}^{(0)} \mathbf{1})^{-1} \mathbf{1}^\top \mathbf{D}^{(0)} \mathbf{X}(\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}). \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Replacing $\alpha - \alpha^{(0)}$ by (2.3) in (2.2), we obtain the following objective function $Q_2(\boldsymbol{\beta})$:

$$\begin{aligned} Q_2(\boldsymbol{\beta}) &= \frac{1}{2n}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^\top \mathbf{X}^\top \mathbf{D}^{\dagger(0)} \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \\ &= \frac{1}{2n}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^\top \mathbf{X}^\top \mathbf{D}^{\star(0)\top} \mathbf{D}^{\star(0)} \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \end{aligned} \tag{2.4}$$

where $\mathbf{D}^{\dagger(0)} = \mathbf{D}^{(0)} - \mathbf{D}^{(0)}\mathbf{1}(\mathbf{1}^\top \mathbf{D}^{(0)}\mathbf{1})^{-1}\mathbf{1}^\top \mathbf{D}^{(0)}$ and $\mathbf{D}^{\star(0)} = (\mathbf{D}^{(0)})^{1/2} - (\mathbf{D}^{(0)})^{1/2}\mathbf{1}(\mathbf{1}^\top \mathbf{D}^{(0)}\mathbf{1})^{-1} \times \mathbf{1}^\top \mathbf{D}^{(0)}$. Denote $\mathbf{D}_0^\dagger = \mathbf{D}_0 - \mathbf{D}_0\mathbf{1}(\mathbf{1}^\top \mathbf{D}_0\mathbf{1})^{-1}\mathbf{1}^\top \mathbf{D}_0$ and $\mathbf{D}_0^\star = \mathbf{D}_0^{1/2} - \mathbf{D}_0^{1/2}\mathbf{1}(\mathbf{1}^\top \mathbf{D}_0\mathbf{1})^{-1}\mathbf{1}^\top \mathbf{D}_0$, correspondingly.

We focus mainly on weak signal identification using the one-step adaptive lasso estimator. However, our method can be extended to other penalized likelihood estimators. Following the idea of Zou and Li (2008), the algorithm for computing the one-step adaptive lasso estimator $\boldsymbol{\gamma}^{(1)}$ is as follows:

Step 1. Create the working data by $\mathbf{X}^\star = \mathbf{D}^{\star(0)}\mathbf{X}\mathbf{W}$ and $\boldsymbol{Y}^\star = \mathbf{D}^{\star(0)}\mathbf{X}\boldsymbol{\beta}^{(0)}$, where $\mathbf{W} = \text{diag}\{|\beta_1^{(0)}|, \ldots, |\beta_p^{(0)}|\}$.

Step 2. Apply the coordinate descent algorithm to solve

$$\hat{\boldsymbol{\beta}}^\star = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i^\star - \sum_{j=1}^{p} x_{ij}^\star \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.5}$$

where $\hat{\boldsymbol{\beta}}^\star = (\hat{\beta}_1^\star, \ldots, \hat{\beta}_p^\star)^\top$, $y_i^\star$ is the ith element of $\boldsymbol{Y}^\star$ and $x_{ij}^\star$ is the (i,j)th element of $\mathbf{X}^\star$.

Step 3. Obtain the value of $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \ldots, \beta_p^{(1)})^\top$ using $\beta_j^{(1)} = \hat{\beta}_j^\star |\beta_j^{(0)}|$, for $j = 1 \ldots, p$.

Step 4. Obtain the value of $\alpha^{(1)}$ as $\alpha^{(1)} = (\mathbf{1}^\top \mathbf{D}^{(0)}\mathbf{1})^{-1}\mathbf{1}^\top \mathbf{D}^{(0)}\mathbf{X}(\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^{(1)}) + \alpha^{(0)}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

From the above algorithm, if $\hat{\beta}_j^\star \neq 0$, then the covariate $\boldsymbol{X}_j$ will be selected. According to (2.5), by using the coordinate descent algorithm, we obtain that

$$\hat{\beta}_j^\star = s\left\{\frac{\sum_{i=1}^n \Big(y_i^\star - \sum_{k\neq j} x_{ik}^\star \hat{\beta}_k^\star\Big) x_{ij}^\star}{\sum_{i=1}^n (x_{ij}^\star)^2}, \frac{n\lambda}{\sum_{i=1}^n (x_{ij}^\star)^2}\right\},$$

where $s(z,r) = \mathrm{sgn}(z)(|z|-r)_+$. Then, the condition for $\hat{\beta}_j^\star \neq 0$ ($\beta_j^{(1)} \neq 0$) is

$$\left|\frac{\sum_{i=1}^n \Big(y_i^\star - \sum_{k\neq j} x_{ik}^\star \hat{\beta}_k^\star\Big) x_{ij}^\star}{\sum_{i=1}^n (x_{ij}^\star)^2}\right| > \frac{n\lambda}{\sum_{i=1}^n (x_{ij}^\star)^2}. \tag{2.6}$$

For each $i \in \{1,\ldots,n\}$ and $s \in \{1,\ldots,n\}$, let $d_{is}^{(0)}$ be the (i,s)th element of $\mathbf{D}^{\star(0)}$. Then the variable selection condition (2.6) is equivalent to

$$\left|\sum_{i=1}^n \Big(\sum_{s=1}^n d_{is}^{(0)} x_{sj}\Big)^2 (\beta_j^{(0)})^2 + \sum_{k\neq j}\sum_{i=1}^n \Big(\sum_{s=1}^n d_{is}^{(0)} x_{sk}\Big)\Big(\sum_{s=1}^n d_{is}^{(0)} x_{sj}\Big)\beta_j^{(0)}(\beta_k^{(0)} - \beta_k^{(1)})\right| > n\lambda. \tag{2.7}$$

Similarly to the proof in Zou and Li (2008), we obtain that if the tuning parameter λ satisfies the conditions of $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, then the one-step adaptive lasso estimator enjoys model selection consistency, and the nonzero one-step adaptive lasso estimators have the property of asymptotic normality.

3. Weak signal definition and identification

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

3.1 Weak signal definition

Suppose a model contains both strong and weak signals. Without loss of generality, assume the covariate matrix $\mathbf{X}$ consists of three components, that is, $\mathbf{X} = \{\mathbf{X}^{(S)}, \mathbf{X}^{(W)}, \mathbf{X}^{(N)}\}$, where $\mathbf{X}^{(S)}$, $\mathbf{X}^{(W)}$, and $\mathbf{X}^{(N)}$ represent the subsets of strong signals, weak signals, and noise variables, respectively. Following Shi and Qu (2017), we use the selection probability of each covariate to measure the signal strength. Specifically, for any penalized model selection estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$, we define $P_{d,j}$ as the probability of selecting the covariate $\boldsymbol{X}_j$, that is, $P_{d,j} = P(\hat{\beta}_j \neq 0)$, $j \in \{1, \ldots, p\}$. For the one-step adaptive lasso estimator $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \ldots, \beta_p^{(1)})^\top$, based on the variable selection condition (2.7), $P_{d,j}$ does not have an explicit form. However, in the Supplementary Material S1, we show that $P_{d,j}$ can be approximated by $P_{d,j}^*$, where

$$P_{d,j}^* = \Phi\left(\frac{-\sqrt{\frac{\lambda \mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} + \beta_{j0}}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})\}_{j+1,j+1}^{-1}}}\right) + \Phi\left(\frac{-\sqrt{\frac{\lambda \mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} - \beta_{j0}}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})\}_{j+1,j+1}^{-1}}}\right). \quad (3.1)$$

Intuitively, in the derivation of the selection probability, we can omit the terms of (S2) and (S3) in the Supplementary Material S1, and simplify the calculation using asymptotic theory. Then we can relax the orthogonality assumption required in Shi and Qu (2017). We require the following mild assumption to ensure (3.1) is valid.

Assumption 1. *For each* $i \in \{1, \ldots, n\}$ *and* $j \in \{1, \ldots, p\}$, $P(D_{0,ii} > 0) = 1$, $\mathrm{E}(D_{0,ii}) < \infty$, $\mathrm{E}(D_{0,ii}x_{ij}^2) < \infty$, *and* $\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})$ *is positive definite.*

The condition $P(D_{0,ii} > 0) = 1$ implies that the conditional log-likelihood function of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

y_i given $\mathbf{x}_i$, $\ell_i\{\mu_i(\boldsymbol{\gamma})\}$, is a concave function of $\mu_i(\boldsymbol{\gamma})$. This is a necessary condition for the uniqueness of the maximum likelihood estimator $\boldsymbol{\gamma}^{(0)}$. In addition, according to the Cauchy–Schwarz inequality, this also ensures that $\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2 > 0$. The conditions of $\mathrm{E}(D_{0,ii}) < \infty$ and $\mathrm{E}(D_{0,ii}x_{ij}^2) < \infty$ guarantee that all expectations of random variables in (3.1) are bounded for finite n. The positive-definite condition of $\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})$ is a necessary condition for the asymptotic normality of the maximum likelihood estimator $\boldsymbol{\gamma}^{(0)}$, and ensures $\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})\}^{-1}_{j+1,j+1} > 0$.

For a deeper understanding of $P^*_{d,j}$, we first study the asymptotic properties of $P^*_{d,j}$. When $\beta_{j0} = 0$,

$$P^*_{d,j} = 2\Phi\left(\frac{-\sqrt{\frac{n\lambda}{[\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2]/\mathrm{E}(D_{0,ii})}}}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})/n\}^{-1}_{j+1,j+1}}}\right).$$

Under Assumption 1, $[\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2]/\mathrm{E}(D_{0,ii})$ and $\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})/n\}^{-1}_{j+1,j+1}$ are both positive and bounded. If $n\lambda \to \infty$, then $P^*_{d,j} \to 0$.

When $\beta_{j0} \neq 0$,

$$P^*_{d,j} = \Phi\left(\frac{-\sqrt{n}\left[\sqrt{\frac{\lambda\mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} - \beta_{j0}\right]}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})/n\}^{-1}_{j+1,j+1}}}\right) + \Phi\left(\frac{-\sqrt{n}\left[\sqrt{\frac{\lambda\mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} + \beta_{j0}\right]}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})/n\}^{-1}_{j+1,j+1}}}\right).$$

If $\sqrt{n}\lambda \to 0$, then $P^*_{d,j} \to 1$ under Assumption 1.

These asymptotic properties of $P^*_{d,j}$ are consistent with the conclusion that the one-step adaptive lasso estimator enjoys model selection consistency if λ satisfies the conditions of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

$\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$.

In the following, we study the finite-sample properties of $P^*_{d,j}$. To illustrate, we first consider three special cases, where the likelihood-based model is taken as a linear regression model, a logistic regression model, and a Poisson regression model, respectively.

Case One: Linear regression model

We first illustrate the simplest case under the linear regression model setting. Let $y_i = \alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$, where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$; then, $D_{0,ii} = 1/\sigma^2$. If we assume $\text{corr}(x_{ij}, x_{ik}) = 0$ for any k, $k \neq j$, then

$$P^*_{d,j} = \Phi\left(\frac{\beta_{j0} - \sqrt{\lambda}\sigma}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-\beta_{j0} - \sqrt{\lambda}\sigma}{\sigma/\sqrt{n}}\right).$$

Note that if the tuning parameter λ is replaced by $\lambda_{Shi} = \lambda\sigma^2$, then $P^*_{d,j}$ has the same form as that in Shi and Qu (2017), where the covariate matrix is assumed to be orthogonal. In this case, $P^*_{d,j}$ does not depend on $\boldsymbol{\gamma}_0^{-j}$, where $\boldsymbol{\gamma}_0^{-j}$ stands for the components in $\boldsymbol{\gamma}_0$ other than β_{j0}. In addition, given any values in $P^*_{d,j}$ except β_{j0}, $P^*_{d,j}$ is a symmetric function of β_{j0} and increases with $|\beta_{j0}|$. Thus, both $P^*_{d,j}$ and $|\beta_{j0}|$ can be used to measure the signal strength of $\boldsymbol{X}_j$, as shown in Shi and Qu (2017).

However, if $\text{corr}(x_{ij}, x_{ik}) \neq 0$; for some k, $k \neq j$, then

$$P^*_{d,j} = \Phi\left(\frac{\beta_{j0} - \sqrt{\lambda}\sigma}{\sigma\Big/\left[\sqrt{n}\sqrt{\{\text{corr}(\tilde{\mathbf{X}})\}^{-1}_{j+1,j+1}}\right]}\right) + \Phi\left(\frac{-\beta_{j0} - \sqrt{\lambda}\sigma}{\sigma\Big/\left[\sqrt{n}\sqrt{\{\text{corr}(\tilde{\mathbf{X}})\}^{-1}_{j+1,j+1}}\right]}\right).$$

Thus, $P^*_{d,j}$ also depends on the correlations between covariates. Given any values in $P^*_{d,j}$ except β_{j0}, $P^*_{d,j}$ is still a symmetric function of β_{j0} and an increasing function of $|\beta_{j0}|$. However, under different correlation structures of $\tilde{\mathbf{X}}$, the shape of $P^*_{d,j}$ can vary with the value of β_{j0}. Therefore, both the value of $|\beta_{j0}|$ and the correlation structure of $\tilde{\mathbf{X}}$ influence

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

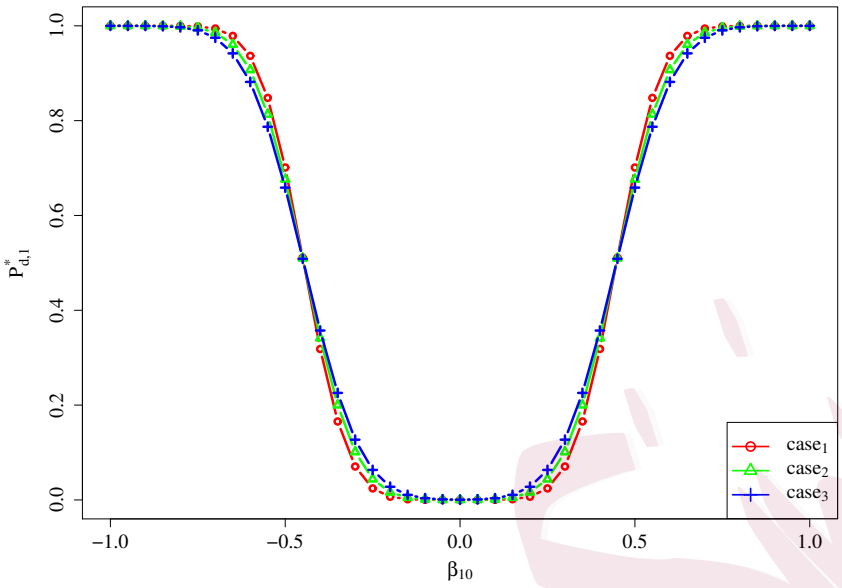

Figure 1: The plots for $P^*_{d,1}$ as β_{10} varies under three different cases in linear regression models. In case 1, the correlation structure of $\mathbf{X}$ is taken to be the independence correlation structure; in case 2, the correlation structure of $\mathbf{X}$ is taken to be the AR(1) correlation structure with $\rho = 0.5$; in case 3, the correlation structure of $\mathbf{X}$ is taken to be the exchangeable correlation structure with $\rho = 0.5$. In all cases, $n = 100$, $p = 5$, $\lambda = 0.2$, $\sigma = 1$, and β_{10} varies between -1 and 1; with a step size of 0.05.

the signal strength of $\boldsymbol{X}_j$, as illustrated in Figure 1.

Case Two: Logistic regression model

Under the logistic regression model setting,

$$\mathrm{E}(y_i|\mathbf{x}_i) = p_i = \frac{\exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)}{1 + \exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)}.$$

We obtain that in (3.1), $D_{0,ii} = p_i(1-p_i)$ and $\mathbf{D}_0 = \mathrm{diag}\{p_1(1-p_1), \ldots, p_n(1-p_n)\}$. Thus, $P^*_{d,j}$ not only depends on β_{j0}, but also depends on $\boldsymbol{\gamma}_0^{-j}$, the coefficients of the other covariates. This is a fundamental difference between logistic regression models and linear regression models in terms of selection probability. In contrast to linear regression models, $\mathbf{x}_i$ influences $P^*_{d,j}$ through the matrix $\mathrm{E}[\widetilde{\mathbf{X}}^\top \mathrm{diag}\{p_1(1-p_1), \ldots, p_n(1-p_n)\}\widetilde{\mathbf{X}}]$, rather than through the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

correlation matrix of $\tilde{\mathbf{X}}$, in logistic regression models. In addition, in the Supplementary Material S2.1, we show that $P^*_{d,j}$ is not necessarily a symmetric function of β_{j0}, given other values in $P^*_{d,j}$. Thus, $|\beta_{j0}|$ cannot be used to measure the signal strength of $\boldsymbol{X}_j$ instead of $P^*_{d,j}$, which differs from Shi and Qu (2017).

In addition, for the logistic regression model, the range of $\boldsymbol{\gamma}_0$ is bounded so that p_i can satisfy the condition $0 < c_1 < p_i < c_2 < 1$, where c_1 and c_2 are some positive constants. We show that, given any values in $P^*_{d,j}$ except β_{j0}, $P^*_{d,j}$ is an increasing function of β_{j0} if $0 < \beta_{j0} < c_3$, and $P^*_{d,j}$ is a decreasing function of β_{j0} if $-c_4 < \beta_{j0} < 0$, where c_3 and c_4 are some bounded positive constants depending on c_1 and c_2. Proofs of the above findings are provided in the Supplementary Material S2.2. We also illustrate these properties in Figure 2. Note that in this case, the response variable has two categories. However, it can be easily extended to the case where there are more than two categories.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

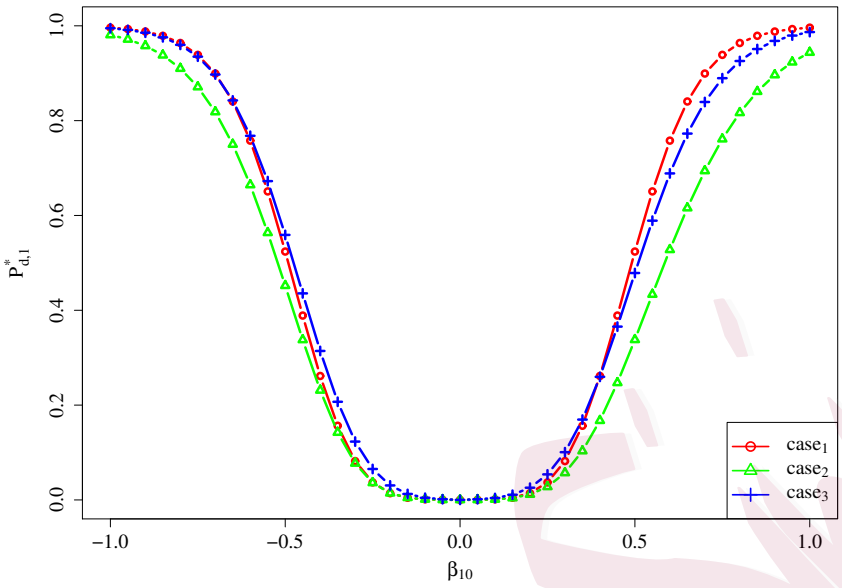

Figure 2: The plots for $P^*_{d,1}$ as β_{10} varies under three different cases in logistic regression models. In case 1, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the standard normal distribution, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent; in case 2, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the centralized exponential distribution with mean zero and variance one, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent; in case 3, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the standard normal distribution, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the correlation of 0.5. In all cases, $n = 300$, $\boldsymbol{\gamma}_0 = (0.3, \beta_{10}, 0.2)'$, $\lambda = 0.05$, and β_{10} varies between -1 and 1, with a step size of 0.05.

Case Three: Poisson regression model

Under the Poisson regression model setting,

$$P(y_i = y|\mathbf{x}_i) = \frac{\lambda_i^y}{y!}\exp(-\lambda_i),$$

where $\lambda_i = \mathrm{E}(y_i|\mathbf{x}_i) = \exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)$. Then, in (3.1), $D_{0,ii} = \lambda_i$ and $\mathbf{D}_0 = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. We obtain similar conclusions to those for logistic regression models, except that $P^*_{d,j}$ is influenced by $\mathbf{x}_i$ through the matrix $\mathrm{E}[\widetilde{\mathbf{X}}^\top \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}\widetilde{\mathbf{X}}]$. Note that under Assumption 1, the range of $\boldsymbol{\gamma}_0$ is bounded. Given any other values in $P^*_{d,j}$ except β_{j0}, $P^*_{d,j}$ is an increasing function of β_{j0} if $0 < \beta_{j0} < c_5$, and $P^*_{d,j}$ is a decreasing function of β_{j0} if $-c_6 < \beta_{j0} < 0$, where c_5 and c_6 are some bounded positive constants. The proof for this finding is provided

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

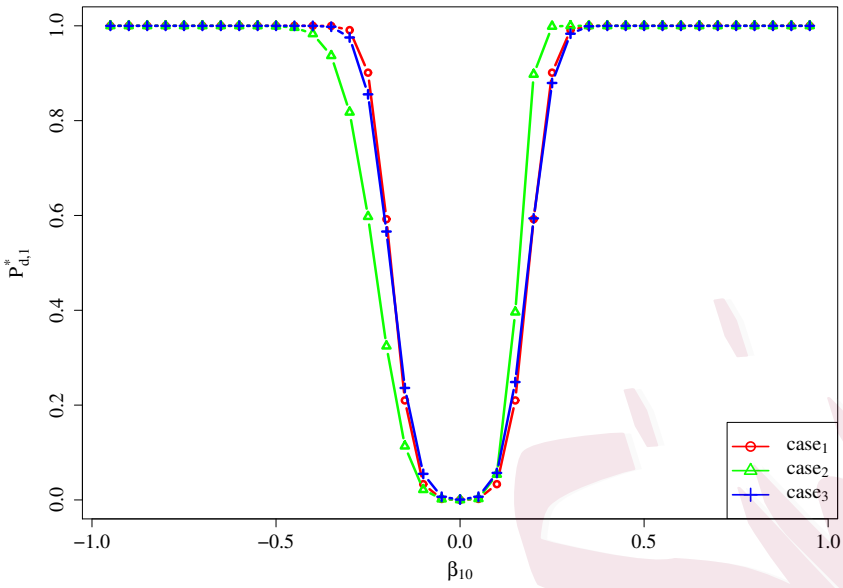

Figure 3: The plots for $P^*_{d,1}$ as β_{10} varies under three different cases in Poisson regression models. In case 1, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the standard normal distribution, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent; in case 2, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the centralized exponential distribution with mean 0 and variance 1, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent; in case 3, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both follow the standard normal distribution, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the correlation of 0.5. In all cases, $n = 300$, $\boldsymbol{\gamma}_0 = (0.3, \beta_{10}, 0.2)'$, $\lambda = 0.05$, β_{10} varies between -0.95 and 0.95, with a step size of 0.05.

in the Supplementary Material S2.2. Figure 3 illustrates $P^*_{d,j}$.

The finite-sample properties of $P^*_{d,j}$ under other likelihood-based models can be analyzed similarly. In general, $P^*_{d,j}$ is an integrated indicator. It shows how the selection probability of $\boldsymbol{X}_j$ is influenced by $\boldsymbol{\gamma}_0$, $\mathbf{x}_i$, n, and λ in finite samples. Given other values in $P^*_{d,j}$ except β_{j0}, $P^*_{d,j}$ is not necessarily a symmetric function of β_{j0} or an increasing function of $|\beta_{j0}|$.

Based on the above analysis, we propose using $P^*_{d,j}$ to measure the signal strength levels directly, rather than using $|\beta_{j0}|$. Intuitively, if $P^*_{d,j}$ is close to one, then the variable $\boldsymbol{X}_j$ is defined to be a strong signal; if $P^*_{d,j}$ is close to zero, then the variable $\boldsymbol{X}_j$ is defined to be a noise variable; if $P^*_{d,j}$ lies between the strong and noise levels, then the variable $\boldsymbol{X}_j$ is defined to be a weak signal. Specifically, we introduce two threshold values, δ^s and δ^w. Then the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

three levels of signal strength can be defined as

$$\begin{cases} \boldsymbol{X}_j \in \mathbf{X}^{(S)}, & \text{if} \quad P^*_{d,j} > \delta^s; \\ \boldsymbol{X}_j \in \mathbf{X}^{(W)}, & \text{if} \quad \delta^w < P^*_{d,j} \leq \delta^s; \\ \boldsymbol{X}_j \in \mathbf{X}^{(N)}, & \text{if} \quad P^*_{d,j} \leq \delta^w, \end{cases} \tag{3.2}$$

where $0 < \tau^w \leq \delta^w < \delta^s \leq \tau^s \leq 1$, $\tau^w = \min_j P^*_{d,j}$, and $\tau^s = \max_j P^*_{d,j}$. Obviously, it is easier to select a stronger signal using the variable selection process than it is to select a weaker signal.

3.2 Weak signal identification

In this section, we show how to identify weak signals. Based on the analysis in Section 3.1, the approximated selection probability $P^*_{d,j}$ depends on the true parameter $\boldsymbol{\gamma}_0$ and the distribution of $\mathbf{x}_i$, but they are always unknown in practice. In the following, we estimate $P^*_{d,j}$ by plugging in the maximum likelihood estimator $\boldsymbol{\gamma}^{(0)}$ and the empirical mean of the random variables in (3.1). That is,

$$\hat{P}^*_{d,j} = \Phi\left(\frac{-\sqrt{\frac{n\lambda \sum_{i=1}^n D^{(0)}_{ii}}{\sum_{i=1}^n D^{(0)}_{ii} x_{ij}^2 \sum_{i=1}^n D^{(0)}_{ii} - (\sum_{i=1}^n D^{(0)}_{ii} x_{ij})^2}} + \beta^{(0)}_j}{\sqrt{(\widetilde{\mathbf{X}}^\top \mathbf{D}^{(0)} \widetilde{\mathbf{X}})^{-1}_{j+1,j+1}}} \right) + \Phi\left(\frac{-\sqrt{\frac{n\lambda \sum_{i=1}^n D^{(0)}_{ii}}{\sum_{i=1}^n D^{(0)}_{ii} x_{ij}^2 \sum_{i=1}^n D^{(0)}_{ii} - (\sum_{i=1}^n D^{(0)}_{ii} x_{ij})^2}} - \beta^{(0)}_j}{\sqrt{(\widetilde{\mathbf{X}}^\top \mathbf{D}^{(0)} \widetilde{\mathbf{X}})^{-1}_{j+1,j+1}}} \right). \tag{3.3}$$

In practice, we identify the signal strength level of $\boldsymbol{X}_j$ based on $\hat{P}^*_{d,j}$, and introduce two threshold values δ_1 and δ_2. We denote the identified subsets of strong signals, weak signals,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

and noise variables as $\hat{\mathbf{S}}^{(S)}$, $\hat{\mathbf{S}}^{(W)}$, and $\hat{\mathbf{S}}^{(N)}$, respectively:

$$\begin{cases} \hat{\mathbf{S}}^{(S)} = \{j : \hat{P}^*_{d,j} > \delta_1\}; \\ \hat{\mathbf{S}}^{(W)} = \{j : \delta_2 < \hat{P}^*_{d,j} \leq \delta_1\}; \\ \hat{\mathbf{S}}^{(N)} = \{j : \hat{P}^*_{d,j} \leq \delta_2\}. \end{cases} \tag{3.4}$$

The selections of δ_1 and δ_2 are crucial to determining the signal type. The threshold value δ_1 is selected to ensure that we can identify strong signals when the selection probabilities of signals are high. Assume α is a significance level, and we choose δ_1 to be larger than $1-\alpha$, so that the identified strong signals are strong. The threshold value δ_2 is selected to control the false positive rate of selecting variable $\boldsymbol{X}_j$. Denote the false positive rate as τ. Then τ can be defined as

$$\tau = P(j \notin \hat{\mathbf{S}}^{(N)} \mid \beta_{j0} = 0, \boldsymbol{\gamma}_0^{-j}) = P(\hat{P}^*_{d,j} > \delta_2 \mid \beta_{j0} = 0, \boldsymbol{\gamma}_0^{-j}). \tag{3.5}$$

Thus, δ_2 can be estimated based on (3.5). Because the value of $\boldsymbol{\gamma}_0$ is unknown in practice, we estimate it using the one-step adaptive lasso estimator $\boldsymbol{\gamma}^{(1)}$. Furthermore, to make the estimated value of the false positive rate equal to τ based on the observed data, we take the value of δ_2 as the $100(1-\tau)\%$ quantile of $\{\hat{P}^*_{d,j} : \beta_j^{(1)} = 0, j = 1, \ldots, p\}$. Because we intend to recover weak signals given finite samples, τ is chosen to be larger than zero. However, the value of τ cannot be too large, because there is a trade-off between recovering weak signals and including noise variables. In practice, if we want to recover more weak signals, we can choose a larger τ; if we want to make the false positive rate lower, we can choose a smaller τ. In the simulation studies, we perform a sensitivity analysis for the choice of δ_1 and τ.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

4. Weak signal inference

In this section, we propose a two-step inference procedure for constructing confidence intervals for the regression coefficients. The procedure consists of two parts: if a covariate is identified as a strong signal, then its confidence interval is constructed based on the asymptotic theory for the nonzero one-step adaptive lasso estimator (Zou and Li, 2008); if a covariate is identified as a weak signal or a noise variable, then we provide a confidence interval based on the following inference theory for the maximum likelihood estimator.

Similarly to the theory in Zou and Li (2008), we can obtain the asymptotic distribution of the one-step adaptive lasso estimator. Without loss of generality, assume $\mathscr{A}_n = \{1, \ldots, s\}$, where s is the number of nonzero elements in $\boldsymbol{\beta}^{(1)}$. Define $\mathscr{B}_n = \{k : \gamma_k^{(1)} \neq 0, k = 1, \ldots, p+1\}$, then $\mathscr{B}_n = \{1, \ldots, s+1\}$. Although the one-step adaptive lasso estimator $\boldsymbol{\beta}_{\mathscr{A}_n}^{(1)}$ is biased, owing to the shrinkage effect in finite samples, we can construct a de-biased confidence interval for the true coefficient based on the estimated bias and covariance matrix for $\boldsymbol{\beta}_{\mathscr{A}_n}^{(1)}$, as shown in Theorem 1. The proof of Theorem 1 is given in the Supplementary Material S3.

Theorem 1. *Denote* $\mathbf{X}^\top \mathbf{D}^{\dagger(0)} \mathbf{X}$ *and* $\widetilde{\mathbf{X}}^\top \mathbf{D}^{(0)} \widetilde{\mathbf{X}}/n$ *as* $\mathbf{Z}^{(0)}$ *and* $\mathbf{I}^{(0)}$*, respectively. The estimators of the bias and the covariance matrix of* $\boldsymbol{\beta}_{\mathscr{A}_n}^{(1)}$ *are given by*

$$\widehat{\text{bias}}(\boldsymbol{\beta}_{\mathscr{A}_n}^{(1)}) = -\left\{ \frac{1}{n} \mathbf{Z}_{\mathscr{A}_n}^{(0)} + \Sigma_\lambda(\boldsymbol{\beta}_{\mathscr{A}_n}^{(0)}, \boldsymbol{\beta}_{\mathscr{A}_n}^{(1)}) \right\}^{-1} \left(\frac{\lambda}{|\beta_1^{(0)}|} \text{sgn}(\beta_1^{(1)}), \ldots, \frac{\lambda}{|\beta_s^{(0)}|} \text{sgn}(\beta_s^{(1)}) \right)^\top$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

and

$$\widehat{\text{cov}}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}_n}) = \frac{1}{n^3}\left\{\frac{1}{n}\mathbf{Z}^{(0)}_{\mathscr{A}_n} + \Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}_n})\right\}^{-1} \mathbf{Z}^{(0)}_{\mathscr{A}_n}\{(\mathbf{I}^{(0)}_{\mathscr{B}_n})^{-1}\}_{\mathscr{A}_n}\mathbf{Z}^{(0)}_{\mathscr{A}_n} \times \left\{\frac{1}{n}\mathbf{Z}^{(0)}_{\mathscr{A}_n} + \Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}_n})\right\}^{-1},$$

respectively, where $\Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}_n}) = \text{diag}\{\lambda/(|\beta^{(0)}_1||\beta^{(1)}_1|), \ldots, \lambda/(|\beta^{(0)}_s||\beta^{(1)}_s|)\}$, $\mathbf{Z}^{(0)}_{\mathscr{A}_n}$ *is the sub-matrix of* $\mathbf{Z}^{(0)}$ *corresponding to* $\boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}$, *and* $\mathbf{I}^{(0)}_{\mathscr{B}_n}$ *is the sub-matrix of* $\mathbf{I}^{(0)}$ *corresponding to* $\boldsymbol{\gamma}^{(0)}_{\mathscr{B}_n}$.

Based on Theorem 1, if the covariate $\boldsymbol{X}_j$ is identified as a strong signal, then the $100(1-\alpha)\%$ confidence interval for β_{j0} can be constructed as

$$(\beta^{(1)}_j - \hat{b}_j - z_{\alpha/2}\hat{\sigma}_j, \beta^{(1)}_j - \hat{b}_j + z_{\alpha/2}\hat{\sigma}_j), \tag{4.1}$$

where $\hat{b}_j$ is the corresponding component of $\widehat{\text{bias}}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}_n})$ and $\hat{\sigma}_j$ is the square root of the corresponding diagonal component of $\widehat{\text{cov}}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}_n})$.

If the covariate $\boldsymbol{X}_j$ is identified as a weak signal or a noise variable, then the $100(1-\alpha)\%$ confidence interval for β_{j0} can be constructed as

$$(\beta^{(0)}_j - z_{\alpha/2}\sigma^{(0)}_j, \beta^{(0)}_j + z_{\alpha/2}\sigma^{(0)}_j), \tag{4.2}$$

where $\sigma^{(0)}_j$ is the square root of the corresponding diagonal component of $\widehat{\text{cov}}(\boldsymbol{\gamma}^{(0)}) = (\widetilde{\mathbf{X}}^\top\mathbf{D}^{(0)}\widetilde{\mathbf{X}})^{-1}$.

Remark 1. Note that Shi and Qu (2017) did not construct confidence intervals for the noise variables, whereas we do. As shown in Figure 6 in the simulation studies, this improves the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

coverage probabilities for the noise variables and weak signals. Using the two-step inference method based on Shi and Qu (2017), the coverage probabilities for the noise variables tend to be lower than $1-\alpha$, and the coverage probabilities for weak signals tend to be higher than $1-\alpha$. This is because we construct confidence intervals for the noise variables only when they are misidentified as weak signals or strong signals, in which case the estimated values of the coefficients tend to be far from the true values, leading to lower coverage probabilities. We do not construct confidence intervals for the weak signals when they are misidentified as noise variables, making the coverage probabilities of the confidence intervals higher. To solve these problems, we propose constructing confidence intervals for the identified noise variables as well. As a result, the coverage probabilities of the confidence intervals become closer to $1-\alpha$.

In summary, our proposed confidence interval for β_{j0} can be written as

$$(\beta_j^{(1)} - \hat{b}_j - z_{\alpha/2}\hat{\sigma}_j, \beta_j^{(1)} - \hat{b}_j + z_{\alpha/2}\hat{\sigma}_j)\mathrm{I}\{j \in \hat{\mathbf{S}}^{(S)}\} + (\beta_j^{(0)} - z_{\alpha/2}\sigma_j^{(0)}, \beta_j^{(0)} + z_{\alpha/2}\sigma_j^{(0)})\mathrm{I}\{j \in \hat{\mathbf{S}}^{(W)} \cup \hat{\mathbf{S}}^{(N)}\}, \quad (4.3)$$

which combines both (4.1) and (4.2).

5. Simulation studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed signal identification criterion and two-step inference procedure. Consider the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

following logistic regression model:

$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)}{1 + \exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)}, \quad i = 1, \ldots, n.$$

We generate the covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{R}(\rho)\sigma^2$, where $\mathbf{R}(\rho)$ is a correlation matrix with the AR(1) correlation structure and $\sigma^2 = 1$. All the generated covariates are standardized by subtracting their sample means and dividing by their sample standard deviations. For each setting, we choose $n = 350$ or 550, $p = 25$ or 35, $\rho = 0$, 0.2, or 0.5, and $\alpha_0 = 0.5$. The regression coefficient vector $\boldsymbol{\beta}_0$ is set to $(1, 1, 0.5, \theta, \underbrace{0, \ldots, 0}_{p-4})^\top$, which consists of two large coefficients 1, one moderate size coefficient 0.5, one varying coefficient θ, and $(p-4)$ zero coefficients. The coefficient θ ranges from zero to one, with a step size of 0.05. In each simulation setting, we repeat the simulations 500 times. The implementation details of the one-step adaptive lasso estimators are given in the Supplementary Material S4.

Figure 4 displays the results for different types of selection probability for $\boldsymbol{X}_4$ when $\rho = 0$. In Figure 4, the approximated selection probability based on (3.1) is close to the empirical selection probability, indicating a small approximation error from the approximated selection probability. In addition, both the empirical selection probability and the approximated selection probability increase with θ, implying that a larger value of θ leads to a stronger signal strength. This observation supports the result in Section 3.1. Although the median of the estimated selection probabilities is not too close to the empirical selection probability when θ is small, the estimated selection probability still increases with the signal strength. We can still use the estimated selection probability to identify the signal strength level. The simulation results for the correlated covariates are provided in Figures S1 and S2 of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

the Supplementary Material S5, and the approximated selection probability is similar to the empirical selection probability. In addition, the empirical selection probability, approximated selection probability, and estimated selection probability, in general, increase with the value of θ. Thus, we can also identify the signal strength level based on the value of θ.

We then identify whether a covariate is a strong signal, weak signal, or noise variable based on the criterion in (3.4). For illustration, we choose δ_1 to be 0.99 and τ to be 0.1. Figure 5 represents the empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories as θ varies and $\rho = 0$. Figure 5 shows that when θ is close to zero, $\boldsymbol{X}_4$ is more likely to be identified as a noise variable; when θ is far away from zero and one, the empirical probability of $\boldsymbol{X}_4$ being identified as a weak signal is highest; as θ becomes larger, the empirical probability of $\boldsymbol{X}_4$ being identified as a strong signal becomes more dominant, and gradually increases to one. The results for the correlated covariates are given in Figures S3 and S4 of the Supplementary Material S5, and we have similar findings. Therefore, our proposed signal identification criterion (3.4) performs well in practice.

After identifying the signal strength levels, we construct the 95% confidence intervals based on the proposed two-step inference procedure. We also compare our method with the two-step inference method based on Shi and Qu (2017), which does not construct confidence intervals for the identified noise variables. In addition, we construct confidence intervals based on the asymptotic theory for the one-step adaptive lasso estimator, as shown in (4.1), the maximum likelihood estimation method, as shown in (4.2), the perturbation method (Minnier et al., 2011), the estimating equation-based method (Neykov et al., 2018), the standard bootstrap method (Efron and Tibshirani, 1994), the smoothed bootstrap method (Efron, 2014), the de-biased lasso method (Javanmard and Montanari, 2014; Van de Geer

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

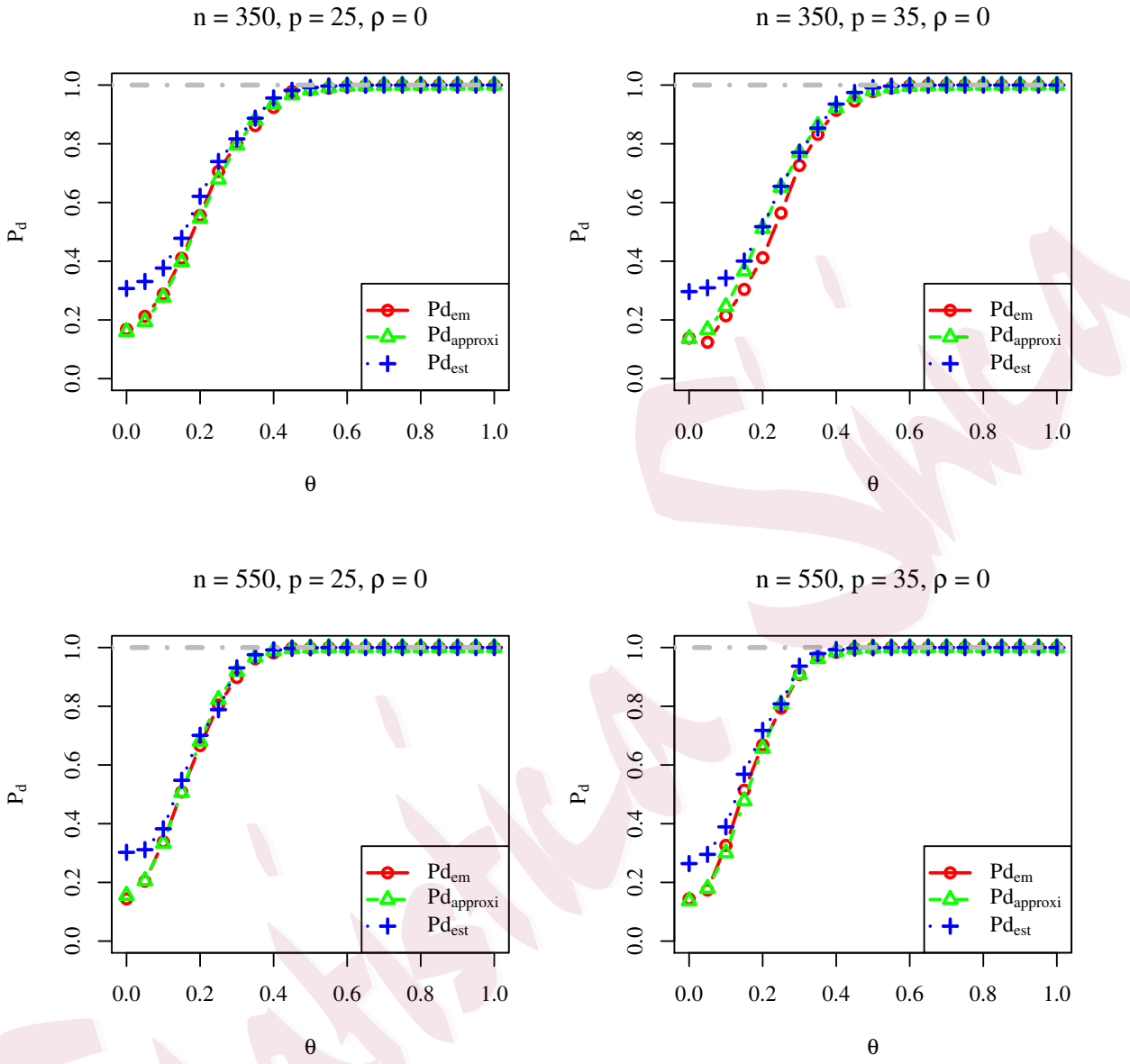

Figure 4: Different types of selection probability for $\boldsymbol{X}_4$ when $\rho = 0$. Pd_{em}: empirical selection probability, which is equal to the empirical probability of $\{\theta^{(1)} \neq 0\}$ based on 500 Monte Carlo samples; $\text{Pd}_{\text{approxi}}$: approximated selection probability based on (3.1), where the expectations in (3.1) are calculated by using the function `cubintegrate` in R; Pd_{est}: median of estimated selection probabilities based on (3.3) for 500 Monte Carlo samples.

et al., 2014; Zhang and Zhang, 2014), and two different types of bootstrap de-biased lasso methods (Dezeure et al., 2017). The number of bootstrap resampling is set to 4000 for all bootstrap methods, and the resampling number is set to 500 for the perturbation method.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

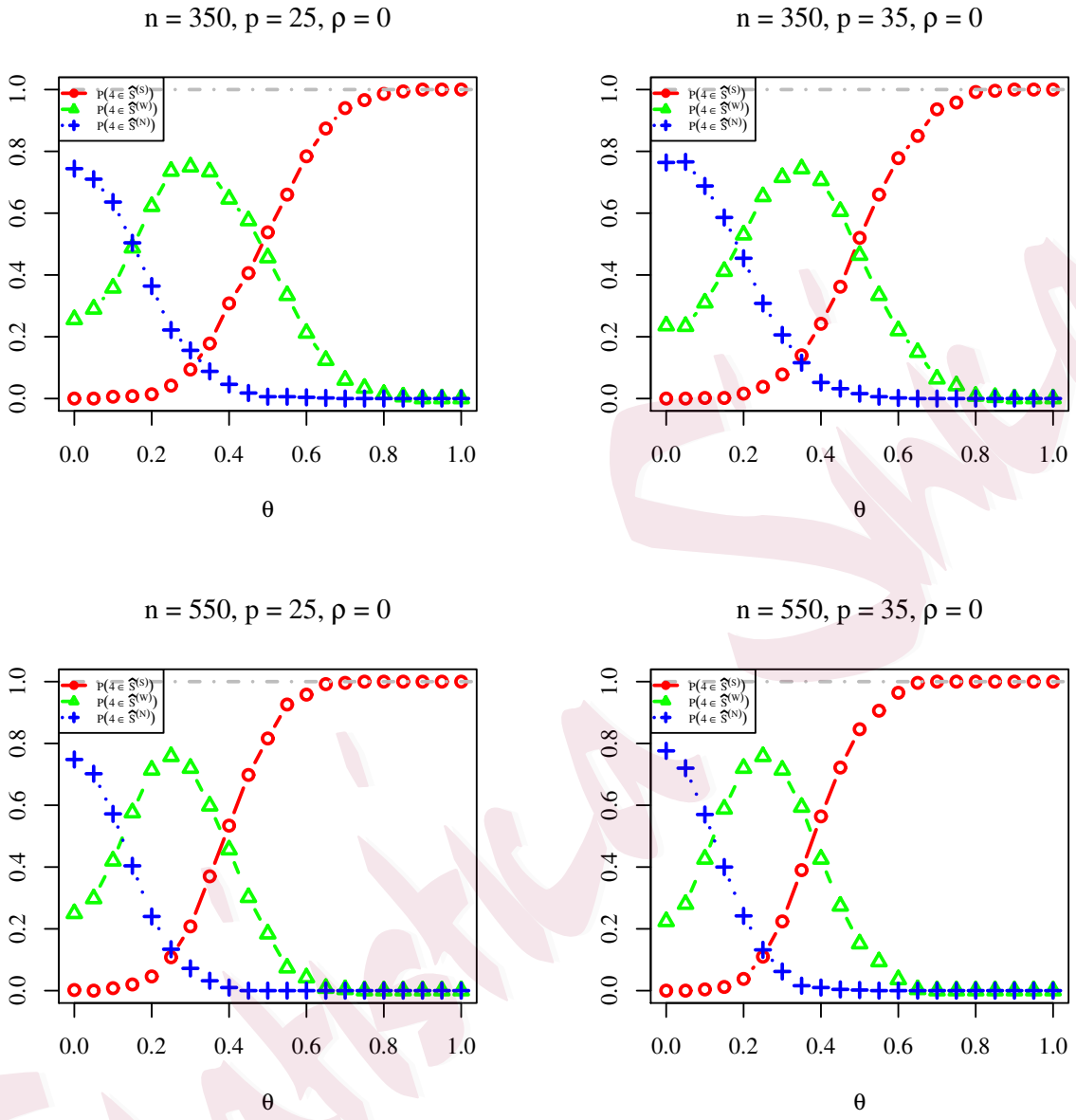

Figure 5: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $\rho = 0$.

The implementation details of the estimating equation-based method and the two types of bootstrap de-biased lasso methods can be found in the Supplementary Material S4. For the method based on the asymptotic theory for the one-step adaptive lasso estimator, if a variable is not selected, then we do not construct a confidence interval for it, because the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

asymptotic normality is established only for the selected variables.

Figures 6 and 7 provide coverage probabilities of the 95% confidence intervals as θ varies and $(n, p, \rho) = (350, 25, 0)$. In Figures 6 and 7, the vertical line on the left shows whether $\boldsymbol{X}_4$ is more likely to be identified as a noise variable or a weak signal, and the vertical line on the right distinguishes whether $\boldsymbol{X}_4$ is more likely to be identified as a weak signal or a strong signal. The threshold values are obtained from Figure 5. Comparing the proposed two-step inference method with the two-step inference method based on Shi and Qu (2017), when θ is small, the former outperforms the latter. When θ is close to zero, the coverage probability of the asymptotic method is too low and close to zero, while the perturbation method, standard bootstrap method, smoothed bootstrap method, and type-I bootstrap de-biased lasso method provide over-coverage confidence intervals, with coverage probabilities approximating to one. When the signal is weak, the asymptotic method, perturbation method, standard bootstrap method, smoothed bootstrap method, and type-I bootstrap de-biased lasso method all perform poorly, and their coverage probabilities are much lower than 95%. In addition, the coverage probability of the estimating equation-based method is slightly lower than 95%. When the signal is stronger, the performance of the maximum likelihood estimation method, estimating equation-based method, de-biased lasso method, and type-I bootstrap de-biased lasso method also become worse. However, the coverage probabilities of the 95% confidence intervals for the proposed method and the type-II bootstrap de-biased lasso method are close to 95% under all signal strength levels of θ.

Figure 8 provides the average widths of the 95% confidence intervals as θ varies and $(n, p, \rho) = (350, 25, 0)$. Note that the widths of the confidence intervals for the two types of two-step inference methods are both very close, while their coverage probabilities are not

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

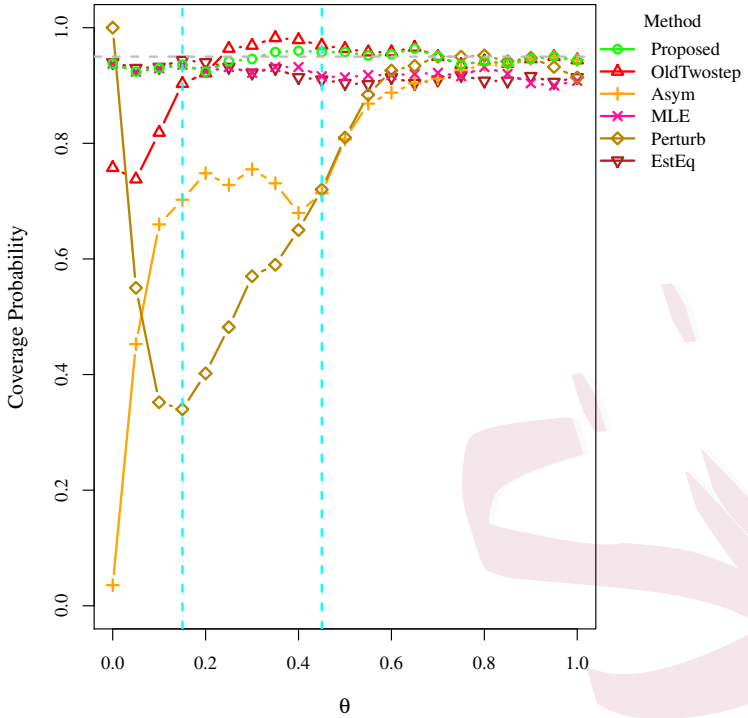

Figure 6: Coverage probabilities of the 95% confidence intervals when $(n, p, \rho) = (350, 25, 0)$. Proposed: the proposed two-step inference method; OldTwostep: the two-step inference method based on Shi and Qu (2017), which does not construct confidence intervals for identified noise variables; Asym: the method based on the asymptotic theory using the one-step adaptive lasso estimator; MLE: the maximum likelihood estimation method; Perturb: the perturbation method; EstEq: the estimating equation-based method.

similar when θ is small. The width of the confidence interval using the proposed method is between those of the maximum likelihood estimation method and the asymptotic method. This is not surprising, because the proposed method combines the strengths of these two methods. Although the confidence intervals based on the asymptotic method, perturbation method, standard bootstrap method, and smoothed bootstrap method are narrow when θ is close to zero, the coverage probabilities are not accurate, because they are either too small or too large. When the signal is strong, the widths of the confidence intervals for the perturbation method, standard bootstrap method, and smoothed bootstrap method are, in general, larger than that for the proposed method. Although the estimating equation-based

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

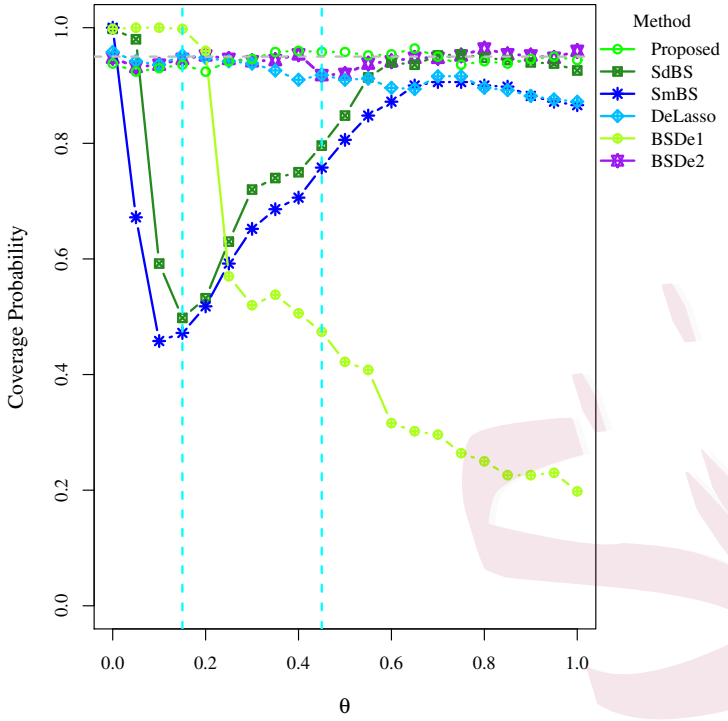

Figure 7: Coverage probabilities of the 95% confidence intervals when $(n, p, \rho) = (350, 25, 0)$. Proposed: the proposed two-step inference method; SdBS: the standard bootstrap method; SmBS: the smoothed bootstrap method; DeLasso: the de-biased lasso method; BSDe1: the type-I bootstrap de-biased lasso method; BSDe2: the type-II bootstrap de-biased lasso method.

method, de-biased lasso method, and type-I bootstrap de-biased lasso method have shorter confidence intervals than that of the proposed method, their coverage probabilities of the confidence intervals decrease as the signal becomes stronger. Overall, the confidence interval for the type-II bootstrap de-biased lasso method is wider than that of the proposed method.

The coverage probabilities and average widths of the 95% confidence intervals under all simulation settings are summarized in Tables S1–S4 of the Supplementary Material S5. For each simulation setting, we select three different values of θ, under which $\boldsymbol{X}_4$ is identified as a noise variable, weak signal, and strong signal, respectively. In summary, the findings from the simulation setting of $(n, p, \rho) = (350, 25, 0)$ still hold under other simulation settings when $\rho = 0$. By comparison, the average widths of the confidence intervals for all methods decrease

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

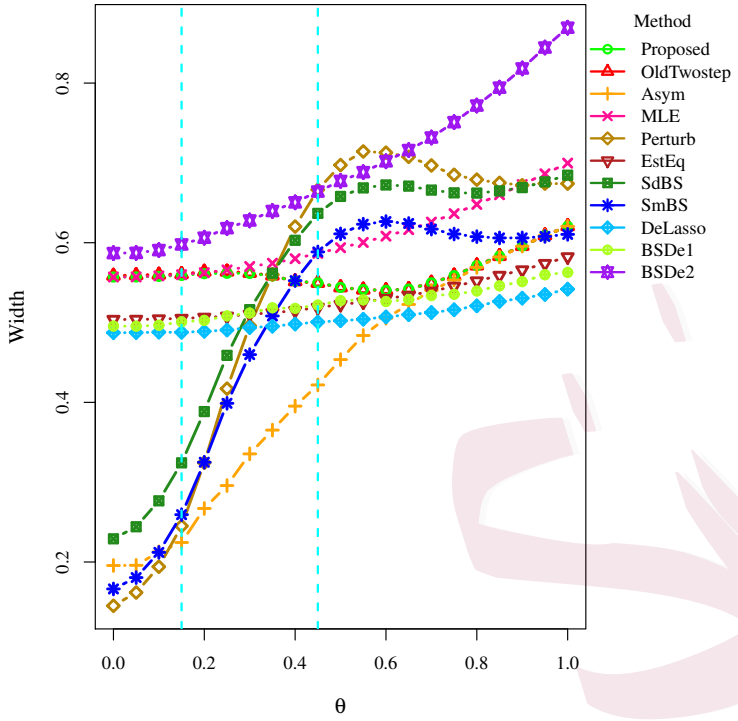

Figure 8: Average widths of the 95% confidence intervals when $(n, p, \rho) = (350, 25, 0)$. Proposed: the proposed two-step inference method; OldTwostep: the two-step inference method based on Shi and Qu (2017), which does not construct confidence intervals for identified noise variables; Asym: the method based on the asymptotic theory using the one-step adaptive lasso estimator; MLE: the maximum likelihood estimation method; Perturb: the perturbation method; EstEq: the estimating equation-based method; SdBS: the standard bootstrap method; SmBS: the smoothed bootstrap method; DeLasso: the de-biased lasso method; BSDe1: the type-I bootstrap de-biased lasso method; BSDe2: the type-II bootstrap de-biased lasso method.

with the sample size and increase with the correlations between the covariates. When $\boldsymbol{X}_4$ is not a strong signal, regardless of the correlations among covariates, the confidence intervals for the asymptotic method have relatively low coverage probabilities. When $\boldsymbol{X}_4$ is a strong signal, if ρ is 0 or 0.2, the asymptotic method provides accurate confidence intervals, but if ρ increases to 0.5, the performance of the asymptotic method deteriorates. However, the coverage probabilities of the confidence intervals for the proposed method are still close to 95% under all simulation settings.

In order to see whether the performance of the proposed method is sensitive to the choice

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

of the threshold values δ_1 and τ, we also consider other combinations of threshold values. For example, when $(n, p, \rho) = (350, 25, 0)$, we set τ as 0.1 and choose δ_1 to be $0.96, 0.97, 0.98$, or 0.99, which is larger than $1-\alpha = 0.95$. The empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories are shown in Figure S5 of the Supplementary Material S5. As the value of δ_1 becomes larger and the value of θ is fixed, the empirical probability of identifying $\boldsymbol{X}_4$ as a weak signal becomes larger, and that of identifying $\boldsymbol{X}_4$ as a strong signal becomes smaller if θ is not sufficiently large. Furthermore, the empirical probability of identifying $\boldsymbol{X}_4$ as a noise variable does not change. This is because of the proposed signal identification criterion. Figures S6–S7 in the Supplementary Material S5 show the corresponding coverage probabilities and average widths of the 95% confidence intervals for the proposed two-step inference method. As shown, the coverage probability becomes larger as δ_1 increases and θ is between 0.6 and 0.75, and the average width becomes larger as δ_1 increases and θ is between 0.15 and 0.75. This is not surprising because when δ_1 increases, the probability of using the maximum likelihood method to construct the confidence intervals becomes larger. As shown in Figures 7 and 8, when θ is not too large, the coverage probability and average width of the confidence interval based on the maximum likelihood method is higher than that based on the asymptotic method. However, as δ_1 varies, the changes of the coverage probability and average width are not large.

We also consider another situation where δ_1 is set to 0.99 and τ is chosen to be $0.05, 0.1, 0.15$, or 0.2. Figure S8 in the Supplementary Material S5 shows the empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories in this situation. Here, we find that as τ increases, the empirical probability of identifying $\boldsymbol{X}_4$ as a weak signal is larger, and that of identifying $\boldsymbol{X}_4$ as a noise variable is smaller if θ is not too large. The empirical probability

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

of identifying $\boldsymbol{X}_4$ as a strong signal remains the same. This is consistent with the proposed signal selection criterion. However, because the proposed two-step inference method uses the same confidence interval construction method for the identified noise variables and weak signals, the confidence interval does not change with the value of τ, as shown in Figures S9–S10 of the Supplementary Material S5.

We also examine whether the performance of the proposed method is sensitive to the total number of weak signals. We reset the regression coefficient vector $\boldsymbol{\beta}_0$ to be $(1, 1, 0.5, \theta, \underbrace{0.3, \ldots, 0.3}_{q}, \underbrace{0, \ldots, 0}_{p-q-4})^\top$, where q is taken to be $0, 1, 2, 3$. For illustration, let $(n, p, \rho) = (350, 25, 0)$, δ_1 be 0.99, and τ be 0.1. Based on the signal identification criterion, all the q covariates corresponding to the coefficient 0.3 are weak signals if θ ranges from zero to one. If the covariate $\boldsymbol{X}_4$ is identified as a weak signal, then the total number of weak signals is $q+1$; otherwise it is q. The empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories are shown in Figure S11 of the Supplementary Material S5, which are not sensitive to the value of q. Figures S12–S13 in the Supplementary Material S5 respectively show the coverage probabilities and average widths of the 95% confidence intervals for the proposed two-step inference method, showing that when θ is small, the average width increases with the value of q, while the coverage probability does not change monotonously with the value of q. In addition, as q varies, the variations of average width and coverage probability are not large. Thus, the performance of the proposed method is quite robust to the total number of weak signals.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

6. Real-data application

To illustrate the performance of the proposed method, we apply it to a data set in the Practice Fusion diabetes study, which was provided by Kaggle as part of the "Practice Fusion Diabetes Classification" challenge (Kaggle, 2012). The data set consists of de-identified electronic medical records for over $10,000$ patients. There are a total of 9948 patients in the training data, including a binary variable indicating whether a patient is diagnosed with Type 2 diabetes mellitus (T2DM), or not. In this analysis, we aim to determine the most important risk factors for the incidence of T2DM, which can be used to identify patients with a high risk of T2DM.

We first extract 119 predictors from the predictors selected by the first-place winner in the Kaggle competition by removing some highly correlated predictors (details can be found in `https://www.kaggle.com/c/pf2012-diabetes/overview/winners`). These predictors can be divided into six categories: basic information, transcript records, diagnosis information, medication information, lab result, and smoking status. Detailed information about these predictors can be found in Table S5 in the Supplementary Material S6. One outlying patient is also removed owing to inaccurate information on the predictors. All the predictors are standardized beforehand. We adopt the following logistic regression model to fit the data set:

$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp\left(\alpha + \sum_{j=1}^{p} x_{ij}\beta_j\right)}{1 + \exp\left(\alpha + \sum_{j=1}^{p} x_{ij}\beta_j\right)}, \quad i = 1, \ldots, n,$$

where $p = 119$ and $n = 9947$.

We first obtain the one-step adaptive lasso estimates of the regression coefficients following the tuning parameter selection procedure given in the Supplementary Material S4. We

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

then identify whether a predictor is a strong signal, weak signal, or noise variable based on criterion (3.4). Here, we choose δ_1 to be 0.99 and τ to be 0.1. From all the predictors, we identify 18 strong signals, 32 weak signals, and 69 noise variables. The 18 strong signals are all selected by the one-step adaptive lasso estimator, indicating consistency between it and our method for strong signal selection. Among the 32 weak signals, 24 are also selected by the one-step adaptive lasso estimator, while the other eight predictors are only identified by our method. These eight additional predictors include the numbers of times being diagnosed with herpes zoster, hypercholesterolemia, hypertensive heart disease, respiratory infection, sleep apnea, and joint pain, and the number of transcripts for cardiovascular disease and the number of diagnoses per weighted year. The relationships between these eight predictors and diabetes have also been studied by other researchers. For example, Papagianni et al. (2018) reviewed studies on associations between herpes zoster and diabetes mellitus, and found that herpes zoster and T2DM were likely to coexist for the same patient.

Next, we construct the 95% confidence intervals using our two-step inference method, together with all other comparison methods in Section 5. Figure 9 shows the average widths of the confidence intervals for the strong and weak signals. For both, the widths of the confidence intervals for the two types of two-step inference methods are the same. For strong signals, the proposed method and the asymptotic method provide the shortest confidence intervals. For weak signals, the widths of the confidence intervals based on the proposed method are smaller than those based on the perturbation method, standard bootstrap method, smoothed bootstrap method, and two types of bootstrap de-biased lasso methods.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

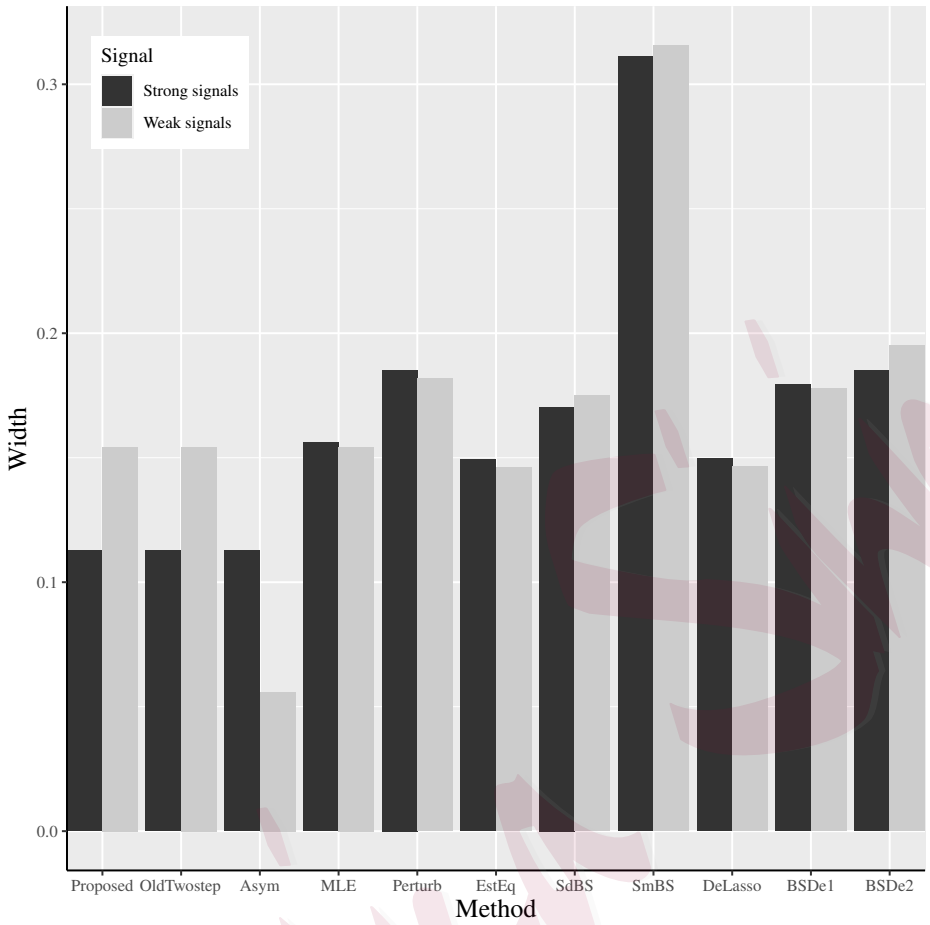

Figure 9: The average widths of the 95% confidence intervals for the diabetes data set. Note that the asymptotic method does not construct confidence intervals for all the weak signals, the result for the weak signals is the average width of the confidence intervals for the weak signals, which are also selected by the asymptotic method. For the meanings of the notation, see Figures 6 and 7.

7. Conclusion

We have proposed a new unified approach for weak signal identification and inference in penalized likelihood models, including the special case when the responses are categorical. To identify weak signals, we propose using the estimated selection probability of each covariate as a measure of the signal strength, and develop a signal identification criterion based directly on the estimated selection probability. To construct confidence intervals for the regression coefficients, we propose a two-step inference procedure. Extensive simulation studies and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

a real-data application show that the proposed signal identification method and two-step inference procedure outperform several existing methods in finite samples.

The proposed method can be extended to a high-dimensional setting where p is not fixed. One possible way is to use the de-biased lasso estimator as an initial estimator for the one-step adaptive lasso estimator, and then leverage the asymptotic properties of the de-biased lasso estimator to derive the selection probability. We can also use a penalized method to estimate the inverse of the information matrix, such as the CLIME estimator (Cai et al., 2011). In addition, our signal identification and inference framework can be extended to longitudinal data. For longitudinal data, we can replace the negative log-likelihood function with the generalized estimating function in the estimation. Finally, in the fields of causal inference and econometrics, there is a popular "weak instrument" problem (Chao and Swanson, 2005; Burgess and Thompson, 2011; Choi et al., 2018), which can be considered a weak signal problem. This is worth further development using our approach.

Supplementary Material

The online Supplementary Material contains six sections. Section S1 derives the approximated selection probability. Section S2 provide an additional detailed analysis of the approximated selection probability in finite samples. Section S3 contains a proof for Theorem 1. Section S4 presents the implementation details of several methods. Sections S5 and S6 provide additional simulation results and information related to the real-data application, respectively.

Acknowledgments

This work was partially supported by the National Science Foundation of the United

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

States (DMS-1821198, DMS-1952406), National Natural Science Foundation of China (11671096, 11731011, 12071087), and Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-07052, DGECR-2019-00453, RGPAS-2019-00093).

References

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in Medicine*, 30(11):1312–1323.

Cai, T., Liu, W., and Luo, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Choi, J., Gu, J., and Shen, S. (2018). Weak-instrument robust inference for two-sample instrumental variables regression. *Journal of Applied Econometrics*, 33(1):109–125.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Huang, J. and Xie, H. (2007). Asymptotic oracle properties of scad-penalized least squares estimators. In *Asymptotics: Particles, processes and inverse problems*, pages 149–166. Institute of Mathematical Statistics.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Jia, J., Xie, F., and Xu, L. (2019). Sparse poisson regression with penalized weighted score function. *Electronic Journal of Statistics*, 13(2):2898–2920.

Jin, J., Zhang, C.-H., and Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *The Journal of Machine Learning Research*, 15(1):2723–2772.

Kaggle (2012). Practice fusion diabetes classification. data retrieved from Kaggle competition dataset, `http://www.kaggle.com/c/pf2012-diabetes`.

Lambert, P. and Eilers, P. H. (2005). Bayesian proportional hazards model with time-varying regression coefficients: A penalized poisson regression approach. *Statistics in Medicine*, 24(24):3977–3989.

Li, Y., Hong, H. G., Ahmed, S. E., and Li, Y. (2019). Weak signals in high-dimensional regression: Detection, estimation and prediction. *Applied Stochastic Models in Business and Industry*, 35(2):283–298.

Liu, H., Xu, X., and Li, J. J. (2020). A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *arXiv preprint arXiv:1706.02150*.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382.

Neykov, M., Ning, Y., Liu, J. S., Liu, H., et al. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443.

Papagianni, M., Metallidis, S., and Tziomalos, K. (2018). Herpes zoster and diabetes mellitus: a review. *Diabetes Therapy*, 9(2):545–550.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Reangsephet, O., Lisawadi, S., and Ahmed, S. E. (2020). Weak signals in high-dimensional logistic regression models. In *International Conference on Management Science and Engineering Management*, pages 121–133. Springer.

Shi, P. and Qu, A. (2017). Weak signal identification and inference in penalized model selection. *The Annals of Statistics*, 45(3):1214–1253.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

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

Van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

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

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

Zhang, H. and Jia, J. (2017). Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signals detection. *arXiv preprint arXiv:1712.03412*.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293.

Zhang, Y. (2017). Recovery of weak signal in high dimensional linear regression by data perturbation. *Electronic Journal of Statistics*, 11(2):3226–3250.

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

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0449

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533.

Department of Computer and Mathematical Sciences, University of Toronto, ON M1C 1A4, Canada.

E-mail: yuexia.zhang@utoronto.ca

Facebook, Menlo Park, CA 94025, United States.

E-mail: pshi@fb.com

Department of Statistics, Fudan University, Shanghai 200433, China.

E-mail: zhuzy@fudan.edu.cn

Department of Statistical Sciences, University of Toronto, ON M5S 3G3, Canada.

E-mail: linbo.wang@utoronto.ca

Department of Statistics, University of California, Irvine, CA 92697, United States.

E-mail: aqu2@uci.edu