Statistica Sinica: Supplement

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]*The University of Texas at San Antonio,* [2]*Meta,* [3]*Fudan University*

[4]*University of Toronto and* [5]*University of California, Irvine*

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.

S1 Derivation of the Approximated Selection Probability

In Section 2 of the main paper, we have obtained the following condition for selecting the covariate $\boldsymbol{X}_j$, $j \in \{1, \ldots, p\}$:

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

It is equivalent to

$$\begin{aligned}
&\left| \frac{\sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big)^2 (\beta_j^{(0)})^2}{n} + \frac{\sum_{k \neq j} \sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sk} \Big) \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big) \beta_j^{(0)} (\beta_k^{(0)} - \beta_{k0} + \beta_{k0})}{n} \right. \\
&\left. - \frac{\sum_{k \neq j} \sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sk} \Big) \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big) \beta_j^{(0)} (\beta_k^{(1)} - \beta_{k0} + \beta_{k0})}{n} \right| \\
=&\left| \frac{\sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big)^2 (\beta_j^{(0)})^2}{n} + \frac{\sum_{k \neq j} \sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sk} \Big) \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big) \beta_j^{(0)} (\beta_k^{(0)} - \beta_{k0})}{n} \right. \\
&\left. - \frac{\sum_{k \neq j} \sum_{i=1}^{n} \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sk} \Big) \Big(\sum_{s=1}^{n} d_{is}^{(0)} x_{sj} \Big) \beta_j^{(0)} (\beta_k^{(1)} - \beta_{k0})}{n} \right| \\
>&\lambda.
\end{aligned} \tag{S1}$$

We consider the following three formulas respectively,

$$\frac{\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)^2 (\beta_j^{(0)})^2}{n},$$

$$\frac{\sum\limits_{k\neq j}\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sk}\Big)\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)\beta_j^{(0)}(\beta_k^{(0)} - \beta_{k0})}{n}, \tag{S2}$$

and

$$\frac{\sum\limits_{k\neq j}\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sk}\Big)\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)\beta_j^{(0)}(\beta_k^{(1)} - \beta_{k0})}{n}. \tag{S3}$$

Since $d_{is}^{(0)}$ is the (i,s)th element of $\mathbf{D}^{\star(0)}$, $\mathbf{D}^{\star(0)} = (\mathbf{D}^{(0)})^{1/2} - (\mathbf{D}^{(0)})^{1/2}\mathbf{1}$ $\times(\mathbf{1}^\top\mathbf{D}^{(0)}\mathbf{1})^{-1}\mathbf{1}^\top\mathbf{D}^{(0)}$ and $\mathbf{D}^{(0)}$ is an $n\times n$ diagonal matrix with the (i,i)th element $D_{ii}^{(0)}$, then by calculation,

$$\frac{\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)^2}{n} = \frac{\sum\limits_{i=1}^{n} D_{ii}^{(0)} x_{ij}^2}{n} - \frac{\left(\frac{\sum_{i=1}^{n} D_{ii}^{(0)} x_{ij}}{n}\right)^2}{\frac{\sum_{i=1}^{n} D_{ii}^{(0)}}{n}}.$$

Since $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are independent and identically distributed random vectors, $D_{ii}(\boldsymbol{\gamma})$ is a continuous function of $\boldsymbol{\gamma}$ and the maximum likelihood estimator $\boldsymbol{\gamma}^{(0)} \xrightarrow{P} \boldsymbol{\gamma}_0$ under some regularity conditions, then by the Law of Large Numbers and Continuous Mapping Theorem, we have $\sum_{i=1}^{n} D_{ii}^{(0)} x_{ij}^2/n \xrightarrow{P} \mathrm{E}(D_{0,ii}x_{ij}^2)$, $\sum_{i=1}^{n} D_{ii}^{(0)} x_{ij}/n \xrightarrow{P} \mathrm{E}(D_{0,ii}x_{ij})$ and $\sum_{i=1}^{n} D_{ii}^{(0)}/n$

$\xrightarrow{P} \mathrm{E}(D_{0,ii})$. Then

$$\frac{\sum\limits_{i=1}^{n}\left(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\right)^2 (\beta_j^{(0)})^2}{n} - \left[\mathrm{E}(D_{0,ii}x_{ij}^2) - \frac{\{\mathrm{E}(D_{0,ii}x_{ij})\}^2}{\mathrm{E}(D_{0,ii})}\right](\beta_j^{(0)})^2 \xrightarrow{P} 0.$$

By calculation, (S2) equals

$$\begin{aligned}
&\sum_{k\neq j}\left(\frac{\sum\limits_{i=1}^{n} x_{ik}D_{ii}^{(0)}x_{ij}}{n} - \frac{\sum\limits_{i=1}^{n}\sum\limits_{s=1}^{n} x_{ik}D_{ii}^{(0)}D_{ss}^{(0)}x_{sj}}{n\sum\limits_{i=1}^{n} D_{ii}^{(0)}}\right)\beta_j^{(0)}(\beta_k^{(0)} - \beta_{k0})\\
=&\sum_{k\neq j}\left(\frac{\sum\limits_{i=1}^{n} x_{ik}D_{ii}^{(0)}x_{ij}}{n} - \frac{\frac{\sum_{i=1}^{n} x_{ik}D_{ii}^{(0)}}{n}\frac{\sum_{s=1}^{n} D_{ss}^{(0)}x_{sj}}{n}}{\frac{\sum_{i=1}^{n} D_{ii}^{(0)}}{n}}\right)\beta_j^{(0)}\frac{\sqrt{n}(\beta_k^{(0)} - \beta_{k0})}{\sqrt{n}}.
\end{aligned}$$

Because of the same reason as before, $\sum_{i=1}^{n} x_{ik}D_{ii}^{(0)}x_{ij}/n \xrightarrow{P} \mathrm{E}(x_{ik}D_{0,ii}x_{ij})$, $\sum_{i=1}^{n} x_{ik}D_{ii}^{(0)}/n \xrightarrow{P} \mathrm{E}(x_{ik}D_{0,ii})$, $\sum_{s=1}^{n} D_{ss}^{(0)}x_{sj}/n \xrightarrow{P} \mathrm{E}(D_{0,ss}x_{sj})$ and $\sum_{i=1}^{n} D_{ii}^{(0)}/n \xrightarrow{P} \mathrm{E}(D_{0,ii})$. By the Central Limit Theorem, $\sqrt{n}(\beta_k^{(0)} - \beta_{k0}) \xrightarrow{D} \mathcal{N}(0, \{\mathbf{I}^{-1}(\boldsymbol{\gamma}_0)\}_{k+1,k+1})$, where $\mathbf{I}(\boldsymbol{\gamma}_0) = \mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})/n$. Then $\sqrt{n}(\beta_k^{(0)} - \beta_{k0}) = O_p(1)$. Furthermore, since $\beta_j^{(0)} \xrightarrow{P} \beta_{j0}$ and the number of covariates p is finite, then according to the Slutsky's Theorem, (S2) is $O_p(1/\sqrt{n})$.

Based on the oracle properties of $\boldsymbol{\beta}^{(1)}$, if $\beta_{k0} = 0$, then $P(\beta_k^{(1)} = 0) \to 1$.

Therefore, similar to the previous proof,

$$\begin{aligned}&\frac{\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sk}\Big)\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)\beta_j^{(0)}(\beta_k^{(1)}-\beta_{k0})}{n}\\ &=\left(\frac{\sum\limits_{i=1}^{n} x_{ik}D_{ii}^{(0)}x_{ij}}{n}-\frac{\frac{\sum_{i=1}^{n} x_{ik}D_{ii}^{(0)}}{n}\frac{\sum_{s=1}^{n} D_{ss}^{(0)}x_{sj}}{n}}{\frac{\sum_{i=1}^{n} D_{ii}^{(0)}}{n}}\right)\beta_j^{(0)}(\beta_k^{(1)}-\beta_{k0})\xrightarrow{P}0.\end{aligned}\tag{S4}$$

If $\beta_{k0}\neq 0$, then $\sqrt{n}(\beta_k^{(1)}-\beta_{k0})\xrightarrow{D}\mathcal{N}(0,[\mathbf{I}^{-1}\{(\boldsymbol{\gamma}_0)_{\mathscr{A}}\}]_{\boldsymbol{X}_k})$, where $\mathbf{I}\{(\boldsymbol{\gamma}_0)_{\mathscr{A}}\}$ is the Fisher information matrix knowing $(\boldsymbol{\gamma}_0)_{\mathscr{A}^c}=\mathbf{0}$ and $[\mathbf{I}^{-1}\{(\boldsymbol{\gamma}_0)_{\mathscr{A}}\}]_{\boldsymbol{X}_k}$ is an element of the matrix $\mathbf{I}^{-1}\{(\boldsymbol{\gamma}_0)_{\mathscr{A}}\}$ corresponding to $\boldsymbol{X}_k$. Therefore, $\sqrt{n}(\beta_k^{(1)}-\beta_{k0})=O_p(1)$. Furthermore,

$$\begin{aligned}&\frac{\sum\limits_{i=1}^{n}\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sk}\Big)\Big(\sum\limits_{s=1}^{n} d_{is}^{(0)} x_{sj}\Big)\beta_j^{(0)}(\beta_k^{(1)}-\beta_{k0})}{n}\\ &=\left(\frac{\sum\limits_{i=1}^{n} x_{ik}D_{ii}^{(0)}x_{ij}}{n}-\frac{\frac{\sum_{i=1}^{n} x_{ik}D_{ii}^{(0)}}{n}\frac{\sum_{s=1}^{n} D_{ss}^{(0)}x_{sj}}{n}}{\frac{\sum_{i=1}^{n} D_{ii}^{(0)}}{n}}\right)\beta_j^{(0)}\frac{\sqrt{n}(\beta_k^{(1)}-\beta_{k0})}{\sqrt{n}}=O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}\tag{S5}$$

According to (S4) and (S5), (S3) is also $O_p(1/\sqrt{n})$.

In summary, the condition for selecting the covariate $\boldsymbol{X}_j$ becomes

$$\left|\left[\mathrm{E}(D_{0,ii}x_{ij}^2)-\frac{\{\mathrm{E}(D_{0,ii}x_{ij})\}^2}{\mathrm{E}(D_{0,ii})}\right](\beta_j^{(0)})^2+o_p(1)\right|>\lambda.$$

Furthermore,

$$P(\beta_j^{(1)} \neq 0) \approx P\left(\left[\mathrm{E}(D_{0,ii}x_{ij}^2) - \frac{\{\mathrm{E}(D_{0,ii}x_{ij})\}^2}{\mathrm{E}(D_{0,ii})}\right](\beta_j^{(0)})^2 > \lambda\right). \qquad \text{(S6)}$$

By the Central Limit Theorem, $\sqrt{n}(\beta_j^{(0)} - \beta_{j0}) \xrightarrow{D} \mathcal{N}(0, \{\mathbf{I}^{-1}(\boldsymbol{\gamma}_0)\}_{j+1,j+1})$ and $\mathbf{I}(\boldsymbol{\gamma}_0) = \mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})/n$. Therefore, the right hand side of (S6) can be approximated by

$$P_{d,j}^* = \Phi\left(\frac{-\sqrt{\frac{\lambda \mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} + \beta_{j0}}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})\}_{j+1,j+1}^{-1}}}\right) + \Phi\left(\frac{-\sqrt{\frac{\lambda \mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2}} - \beta_{j0}}{\sqrt{\{\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})\}_{j+1,j+1}^{-1}}}\right). \qquad \text{(S7)}$$

S2 Additional Detailed Analysis of the Approximated Selection Probability in Finite Samples

In this selection, we provide an additional detailed analysis of finite-sample properties of the approximated selection probability $P_{d,j}^*$ and provide some plots to illustrate the finite-sample properties of $P_{d,j}^*$ under three different kinds of likelihood-based models.

S2.1 Symmetry of the approximated selection probability

In order to study given any values in $P^*_{d,j}$ except β_{j0}, whether $P^*_{d,j}$ is a symmetric function of β_{j0} or not, we need to study for any $\beta_{j0} \neq 0$, whether $P^*_{d,j}(\beta_{j0})$ is equal to $P^*_{d,j}(-\beta_{j0})$. According to (S7),

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

and

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

Since $D_{0,ii}(\beta_{j0},\boldsymbol{\gamma}_0^{-j}) = -\partial^2\ell_i\{\mu_i(\beta_{j0},\boldsymbol{\gamma}_0^{-j})\}/\partial\mu_i^2$ with $\mu_i(\beta_{j0},\boldsymbol{\gamma}_0^{-j}) = \alpha_0 + \sum_{k\neq j} x_{ik}\beta_{k0}+x_{ij}\beta_{j0}$, and $D_{0,ii}(-\beta_{j0},\boldsymbol{\gamma}_0^{-j}) = -\partial^2\ell_i\{\mu_i(-\beta_{j0},\boldsymbol{\gamma}_0^{-j})\}/\partial\mu_i^2$ with $\mu_i(-\beta_{j0},\boldsymbol{\gamma}_0^{-j}) = \alpha_0 + \sum_{k\neq j} x_{ik}\beta_{k0} - x_{ij}\beta_{j0}$, then one of the sufficient conditions for $P^*_{d,j}(\beta_{j0}) = P^*_{d,j}(-\beta_{j0})$ is that the distribution of x_{ij} is symmetric about zero and x_{ij} is independent of x_{ik} for any $k \neq j$. Under this condition, we have $\mathrm{E}\{D_{0,ii}(\beta_{j0}, \boldsymbol{\gamma}_0^{-j})\} = \mathrm{E}\{D_{0,ii}(-\beta_{j0}, \boldsymbol{\gamma}_0^{-j})\}$, $\mathrm{E}\{D_{0,ii}(\beta_{j0}, \boldsymbol{\gamma}_0^{-j})x_{ij}^2\} = \mathrm{E}\{D_{0,ii}(-\beta_{j0}, \boldsymbol{\gamma}_0^{-j})x_{ij}^2\}$, $\mathrm{E}\{D_{0,ii}(\beta_{j0}, \boldsymbol{\gamma}_0^{-j})x_{ij}\} = -\mathrm{E}\{D_{0,ii}(-\beta_{j0}, \boldsymbol{\gamma}_0^{-j})x_{ij}\}$ and $\mathrm{E}\{\widetilde{\mathbf{X}}^\top \mathbf{D}_0(\beta_{j0}, \boldsymbol{\gamma}_0^{-j})\widetilde{\mathbf{X}}\} = \mathrm{E}\{\widetilde{\mathbf{X}}^\top \mathbf{D}_0(-\beta_{j0}, \boldsymbol{\gamma}_0^{-j})\widetilde{\mathbf{X}}\}$. Furthermore, $P_{d,j}^*(\beta_{j0}) = P_{d,j}^*(-\beta_{j0})$.

However, this sufficient condition may not be satisfied in practice and it is easy to find a case where $P_{d,j}^*(\beta_{j0}) \neq P_{d,j}^*(-\beta_{j0})$. So given any values in $P_{d,j}^*$ except β_{j0}, $P_{d,j}^*$ is not necessarily a symmetric function of β_{j0}.

S2.2 Monotonicity of the approximated selection probability

In order to study the monotonicity of the approximated selection probability, we need to study the first order derivative of $P_{d,j}^*$ with respect to β_{j0}. By calculation,

$$\frac{\partial P_{d,j}^*}{\partial \beta_{j0}} = \frac{1}{f_{2j}} \phi\left(\frac{-\sqrt{f_{1j}} - \beta_{j0}}{\sqrt{f_{2j}}}\right) \delta(\beta_{j0}),$$

where

$$f_{1j} = \frac{\lambda \mathrm{E}(D_{0,ii})}{\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii}) - \{\mathrm{E}(D_{0,ii}x_{ij})\}^2},$$

$$f_{2j} = \{\mathrm{E}(\widetilde{\mathbf{X}}^\top \mathbf{D}_0 \widetilde{\mathbf{X}})\}_{j+1,j+1}^{-1},$$

and

$$
\begin{aligned}
&\delta(\beta_{j0})\\
&=\left[\left\{-\frac{1}{2}(f_{1j})^{-\frac{1}{2}}\frac{\partial f_{1j}}{\partial\beta_{j0}}+1\right\}\sqrt{f_{2j}}-\frac{1}{2}(f_{2j})^{-\frac{1}{2}}(-\sqrt{f_{1j}}+\beta_{j0})\frac{\partial f_{2j}}{\partial\beta_{j0}}\right]\exp\left(\frac{2\sqrt{f_{1j}}\beta_{j0}}{f_{2j}}\right)\\
&\qquad+\left\{-\frac{1}{2}(f_{1j})^{-\frac{1}{2}}\frac{\partial f_{1j}}{\partial\beta_{j0}}-1\right\}\sqrt{f_{2j}}+\frac{1}{2}(f_{2j})^{-\frac{1}{2}}(\sqrt{f_{1j}}+\beta_{j0})\frac{\partial f_{2j}}{\partial\beta_{j0}},
\end{aligned}
$$

with

$$
\begin{aligned}
&\frac{\partial f_{1j}}{\partial\beta_{j0}}\\
&=\frac{\lambda\frac{\partial \mathrm{E}(D_{0,ii})}{\partial\beta_{j0}}\left[\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2\right]}{\left[\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2\right]^2}\\
&\quad-\frac{\lambda\mathrm{E}(D_{0,ii})\left\{\frac{\partial\mathrm{E}(D_{0,ii}x_{ij}^2)}{\partial\beta_{j0}}\mathrm{E}(D_{0,ii})+\mathrm{E}(D_{0,ii}x_{ij}^2)\frac{\partial\mathrm{E}(D_{0,ii})}{\partial\beta_{j0}}-2\mathrm{E}(D_{0,ii}x_{ij})\frac{\partial\mathrm{E}(D_{0,ii}x_{ij})}{\partial\beta_{j0}}\right\}}{\left[\mathrm{E}(D_{0,ii}x_{ij}^2)\mathrm{E}(D_{0,ii})-\{\mathrm{E}(D_{0,ii}x_{ij})\}^2\right]^2},
\end{aligned}
$$

$$
\frac{\partial f_{2j}}{\partial\beta_{j0}}=\left[\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})\}^{-1}\left\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{M}_0\widetilde{\mathbf{X}})\right\}\{\mathrm{E}(\widetilde{\mathbf{X}}^\top\mathbf{D}_0\widetilde{\mathbf{X}})\}^{-1}\right]_{j+1,j+1},
$$

and

$$
\mathbf{M}_0=\mathrm{diag}\left\{\frac{\partial^3\ell_1\{\mu_1(\boldsymbol{\gamma}_0)\}}{\partial\mu_1^3}x_{1j},\ldots,\frac{\partial^3\ell_n\{\mu_n(\boldsymbol{\gamma}_0)\}}{\partial\mu_n^3}x_{nj}\right\}.
$$

To simplify the proof, we first consider the case where $(\mathbf{x}_i, y_i)$ follows a logistic regression model, that is,

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

By calculation, $D_{0,ii} = p_i(1-p_i)$ and $\mathbf{D}_0 = \operatorname{diag}\{p_1(1-p_1), \ldots, p_n(1-p_n)\}$. Assume $p = 2$, x_{i1} and x_{i2} are independent, $\mathrm{E}(x_{ij}) = 0$ and $\mathrm{Var}(x_{ij}) = 1$, $j = 1, 2$. Denote $\exp(\alpha_0 + x_{ik}\beta_{k0})$ as t_k, $k \neq j$. It is easy to show that $\delta(0) = 0$ and

$$\left.\frac{\partial \delta(\beta_{j0})}{\partial \beta_{j0}}\right|_{\beta_{j0}=0}$$
$$= \sqrt{\frac{\lambda}{n}} \times \frac{2\left[\mathrm{E}\left\{\frac{t_k(1-t_k)x_{ik}}{(1+t_k)^3}\right\}\mathrm{E}\left\{\frac{t_k}{(1+t_k)^2}\right\} - \mathrm{E}\left\{\frac{t_k x_{ik}}{(1+t_k)^2}\right\}\mathrm{E}\left\{\frac{t_k(1-t_k)}{(1+t_k)^3}\right\}\right]^2}{\left[\mathrm{E}\left\{\frac{t_k}{(1+t_k)^2}\right\}\right]^3\left[\mathrm{E}\left\{\frac{t_k x_{ik}^2}{(1+t_k)^2}\right\}\mathrm{E}\left\{\frac{t_k}{(1+t_k)^2}\right\} - \left[\mathrm{E}\left\{\frac{t_k x_{ik}}{(1+t_k)^2}\right\}\right]^2\right]} + 2\sqrt{n\lambda} > 0.$$

Therefore,

$$\left.\frac{\partial P_{d,j}^*}{\partial \beta_{j0}}\right|_{\beta_{j0}=0} = 0 \quad \text{and} \quad \left.\frac{\partial^2 P_{d,j}^*}{\partial \beta_{j0}^2}\right|_{\beta_{j0}=0} > 0.$$

It means that $P_{d,j}^*$ obtains a minimum value at $\beta_{j0} = 0$. Furthermore, there exists two positive constant c_1 and c_2 such that $\delta(\beta_{j0}) \geq 0$ for any $\beta_{j0} \in [0, c_1]$ and $\delta(\beta_{j0}) \leq 0$ for any $\beta_{j0} \in [-c_2, 0]$. Thus, $\partial P_{d,j}^*/\partial \beta_{j0} \geq 0$ for any $\beta_{j0} \in [0, c_1]$ and $\partial P_{d,j}^*/\partial \beta_{j0} \leq 0$ for any $\beta_{j0} \in [-c_2, 0]$. In other words, $P_{d,j}^*$ is an increasing function of β_{j0} if $0 < \beta_{j0} < c_1$ and $P_{d,j}^*$ is a decreasing function of β_{j0} if $-c_2 < \beta_{j0} < 0$.

Second, we consider the case where $(\mathbf{x}_i, y_i)$ follows a Poisson regression model, that is,

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

where $\lambda_i = \mathrm{E}(y_i|\mathbf{x}_i) = \exp(\alpha_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_0)$. By calculation, $D_{0,ii} = \lambda_i$ and $\mathbf{D}_0 = \mathrm{diag}\{\lambda_1, \dots, \lambda_n\}$. Assume $p = 2$, x_{i1} and x_{i2} are independent, $\mathrm{E}(x_{ij}) = 0$ and $\mathrm{Var}(x_{ij}) = 1$, $j = 1, 2$. Denote $\exp(\alpha_0 + x_{ik}\beta_{k0})$ as t_k, $k \neq j$. Then

$$\frac{\partial P^*_{d,j}}{\partial \beta_{j0}} = \frac{n\lambda}{f_{1j}} \phi \left(-\sqrt{n\lambda} - \beta_{j0} \sqrt{\frac{n\lambda}{f_{1j}}} \right) \delta(\beta_{j0}),$$

with

$$\delta(\beta_{j0}) = \left(\sqrt{\frac{f_{1j}}{n\lambda}} - \frac{\beta_{j0}}{2\sqrt{n\lambda f_{1j}}} \frac{\partial f_{1j}}{\beta_{j0}} \right) \left\{ \exp\left(\frac{2\beta_{j0} n\lambda}{\sqrt{f_{1j}}} \right) - 1 \right\},$$

$$f_{1j} = \frac{\lambda \mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\}}{\mathrm{E}(t_k) \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}^2\right\} \mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\} - \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}\right\}\right]^2 \right]},$$

and

$$\begin{aligned}\frac{\partial f_{1j}}{\partial \beta_{j0}} =& \frac{2\lambda \mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\} \mathrm{E}\{\exp(x_{ij}\beta_{j0}) x_{ij}\} \mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}^2\right\}}{\mathrm{E}(t_k) \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}^2\right\} \mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\} - \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}\right\}\right]^2 \right]^2} \\ &- \frac{\lambda \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}\right\}\right]^3 + \lambda \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\}\right]^2 \mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}^3\right\}}{\mathrm{E}(t_k) \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}^2\right\} \mathrm{E}\left\{\exp(x_{ij}\beta_{j0})\right\} - \left[\mathrm{E}\left\{\exp(x_{ij}\beta_{j0}) x_{ij}\right\}\right]^2 \right]^2}.\end{aligned}$$

In particular, if x_{ij} follows the standard normal distribution, then

$$\begin{aligned}\frac{\partial P^*_{d,j}}{\partial \beta_{j0}} =& n\mathrm{E}(t_k)\exp(\beta_{j0}^2/2)\phi\left[-\sqrt{n\lambda}-\beta_{j0}\sqrt{n\mathrm{E}(t_k)\exp(\beta_{j0}^2/2)}\right]\\ &\times\left\{\sqrt{\frac{1}{n\mathrm{E}(t_k)\exp(\beta_{j0}^2/2)}+\frac{\beta_{j0}^2}{2\sqrt{n\mathrm{E}(t_k)\exp(\beta_{j0}^2/2)}}}\right\}\\ &\times\left[\exp\left\{2\beta_{j0}n\sqrt{\lambda\mathrm{E}(t_k)\exp(\beta_{j0}^2/2)}\right\}-1\right].\end{aligned}$$

Obviously, $\partial P^*_{d,j}/\partial\beta_{j0} > 0$ if $\beta_{j0} > 0$, $\partial P^*_{d,j}/\partial\beta_{j0} = 0$ if $\beta_{j0} = 0$ and $\partial P^*_{d,j}/\partial\beta_{j0} < 0$ if $\beta_{j0} < 0$. Thus, $P^*_{d,j}$ is an increasing function of β_{j0} if $\beta_{j0} > 0$ and $P^*_{d,j}$ is a decreasing function of β_{j0} if $\beta_{j0} < 0$.

S3 Proof for Theorem 1

According to (2.4) in the main paper, the objective function about $\boldsymbol{\beta}$ for the one-step adaptive lasso estimator is

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)})^\top\mathbf{X}^\top\mathbf{D}^{\dagger(0)}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)}) + \sum_{j=1}^{p}\lambda\frac{|\beta_j|}{|\beta_j^{(0)}|}.$$

For $\beta_j \approx \beta_j^{(1)}$, $Q(\boldsymbol{\beta})$ can be approximated by

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

where $L(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)})^\top \mathbf{X}^\top \mathbf{D}^{\dagger(0)}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)})/(2n)$.

It can be shown easily that there exists a $\boldsymbol{\beta}_{\mathscr{A}}^{(1)}$ that is a $\sqrt{n}$-consistent local minimizer of $Q\{(\boldsymbol{\beta}_{\mathscr{A}}^\top, \mathbf{0}_{\mathscr{A}^c}^\top)^\top\}$ and satisfies the following condition:

$$
\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_{\mathscr{A}}^{(1)} \\ \mathbf{0}_{\mathscr{A}^c}\end{pmatrix}} = 0 \quad \text{for} \quad j = 1, \ldots, q,
$$

where $\mathscr{A} = \{j : \beta_{j0} \neq 0, j = 1, \ldots, p\}$ and $\mathscr{A}^c = \{j : \beta_{j0} = 0, j = 1, \ldots, p\}$. Without loss of generality, assume $\mathscr{A} = \{1, \ldots, q\}$ and $q \leq p$.

Note that $\boldsymbol{\beta}_{\mathscr{A}}^{(1)}$ is a consistent estimator, then

$$
\begin{aligned}
&\left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_{\mathscr{A}}^{(1)} \\ \mathbf{0}_{\mathscr{A}^c}\end{pmatrix}} + \frac{\lambda}{|\beta_j^{(0)}||\beta_j^{(1)}|}\beta_j^{(1)} \\
=&\left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_{\mathscr{A}}^{(1)} \\ \mathbf{0}_{\mathscr{A}^c}\end{pmatrix}} + \frac{\lambda}{|\beta_j^{(0)}|}\mathrm{sgn}(\beta_j^{(1)}) \\
=&\frac{\partial L(\boldsymbol{\beta_0})}{\partial \beta_j} + \sum_{\ell=1}^{q}\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_\ell} + o_p(1)\right\}(\beta_\ell^{(1)} - \beta_{\ell 0}) \\
&+ \frac{\lambda}{|\beta_j^{(0)}|}\mathrm{sgn}(\beta_{j0}) + \frac{\lambda}{|\beta_j^{(0)}||\beta_j^{(1)}|}(\beta_j^{(1)} - \beta_{j0}) = 0.
\end{aligned}
\tag{S8}
$$

Denote $\mathbf{X}^\top \mathbf{D}^{\dagger(0)} \mathbf{X}$ as $\mathbf{Z}^{(0)}$, then according to (S8),

$$
\begin{aligned}
&\sqrt{n}\left\{\frac{1}{n}\mathbf{Z}^{(0)}_{\mathscr{A}} + \Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}})\right\} \\
&\times \left[\boldsymbol{\beta}^{(1)}_{\mathscr{A}} - \boldsymbol{\beta}_{0,\mathscr{A}} + \left\{\frac{1}{n}\mathbf{Z}^{(0)}_{\mathscr{A}} + \Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}})\right\}^{-1} \boldsymbol{b}(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}^{(0)}_{\mathscr{A}})\right] \qquad \text{(S9)} \\
= &-\sqrt{n}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_{\mathscr{A}}} = \frac{1}{\sqrt{n}}\mathbf{Z}^{(0)}_{\mathscr{A}}(\boldsymbol{\beta}^{(0)}_{\mathscr{A}} - \boldsymbol{\beta}_{0,\mathscr{A}}),
\end{aligned}
$$

where $\Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}}) = \mathrm{diag}\{\lambda/(|\beta^{(0)}_1||\beta^{(1)}_1|), \ldots, \lambda/(|\beta^{(0)}_q||\beta^{(1)}_q|)\}$ and $\boldsymbol{b}(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}^{(0)}_{\mathscr{A}}) = (\lambda \times \mathrm{sgn}(\beta_{10})/|\beta^{(0)}_1|, \ldots, \lambda \times \mathrm{sgn}(\beta_{q0})/|\beta^{(0)}_q|)^\top$. According to the Central Limit Theorem, $\sqrt{n}(\boldsymbol{\beta}^{(0)}_{\mathscr{A}} - \boldsymbol{\beta}_{0,\mathscr{A}}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \{(\mathbf{I}_{0,\mathscr{B}})^{-1}\}_{\mathscr{A}})$, where $\mathscr{B} = \{k : \gamma_{k0} \neq 0, k = 1, \ldots, p+1\}$. Furthermore, according to the Slutsky's Theorem, the asymptotic bias of $\boldsymbol{\beta}^{(1)}_{\mathscr{A}}$ is

$$
\mathrm{bias}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}}) = -\left\{\frac{1}{n}\mathbf{Z}_{0,\mathscr{A}} + \Sigma_\lambda(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}_{0,\mathscr{A}})\right\}^{-1} \boldsymbol{b}(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}_{0,\mathscr{A}}),
$$

where $\mathbf{Z}_0 = \mathrm{E}(\mathbf{X}^\top \mathbf{D}^\dagger_0 \mathbf{X})$. The asymptotic covariance matrix of $\boldsymbol{\beta}^{(1)}_{\mathscr{A}}$ is

$$
\begin{aligned}
\mathrm{cov}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}}) = \frac{1}{n^3}\left\{\frac{1}{n}\mathbf{Z}_{0,\mathscr{A}} + \Sigma_\lambda(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}_{0,\mathscr{A}})\right\}^{-1} &\mathbf{Z}_{0,\mathscr{A}}\{(\mathbf{I}_{0,\mathscr{B}})^{-1}\}_{\mathscr{A}}\mathbf{Z}_{0,\mathscr{A}} \\
&\times \left\{\frac{1}{n}\mathbf{Z}_{0,\mathscr{A}} + \Sigma_\lambda(\boldsymbol{\beta}_{0,\mathscr{A}}, \boldsymbol{\beta}_{0,\mathscr{A}})\right\}^{-1}.
\end{aligned}
$$

If $\lambda \to 0$ as n goes to infinity, then $\mathrm{bias}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}}) \to \mathbf{0}$ and $n\mathrm{cov}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}}) \to \{(\mathbf{I}_{0,\mathscr{B}})^{-1}\}_{\mathscr{A}}$.

If n is finite, then the bias of $\boldsymbol{\beta}^{(1)}_{\mathscr{A}}$ can not be ignored and $\mathscr{A}_n$ is not necessarily equal to $\mathscr{A}$. Without loss of generality, assume $\mathscr{A}_n = \{j : \beta^{(1)}_j \neq 0, j = 1, \ldots, p\} = \{1, \ldots, s\}$. Then $\mathscr{B}_n = \{k : \gamma^{(1)}_k \neq 0, k = 1, \ldots, p+1\} = \{1, \ldots, s+1\}$. Furthermore, the estimators of bias and covariance matrix of $\boldsymbol{\beta}^{(1)}_{\mathscr{A}_n}$ are given by

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

and

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

where $\Sigma_\lambda(\boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}, \boldsymbol{\beta}^{(1)}_{\mathscr{A}_n}) = \text{diag}\{\lambda/(|\beta^{(0)}_1||\beta^{(1)}_1|), \ldots, \lambda/(|\beta^{(0)}_s||\beta^{(1)}_s|)\}$ and $\boldsymbol{b}(\boldsymbol{\beta}^{(1)}_{\mathscr{A}_n}, \boldsymbol{\beta}^{(0)}_{\mathscr{A}_n}) = (\lambda \times \text{sgn}(\beta^{(1)}_1)/|\beta^{(0)}_1|, \ldots, \lambda \times \text{sgn}(\beta^{(1)}_s)/|\beta^{(0)}_s|)^\top$.

S4 Implementation Details of Several Methods

In this section, we introduce the implementation details of several methods mentioned in the main paper.

S4.1 One-step adaptive lasso estimator

To obtain the one-step adaptive lasso estimator, we use the function `glmnet` in R to solve (2.5). The selection of tuning parameter λ is important. In finite samples, if λ is too large, the bias of the one-step adaptive lasso estimator will be large and the coverage probability of the confidence interval constructed based on the asymptotic theory for the one-step adaptive lasso estimator will be low; if λ is too small, the number of false positives will be large and the width of the confidence interval will also be large. The Bayesian information criterion (BIC) and cross-validation (CV) method are two commonly used tuning parameter selection methods. Based on the simulation results, λ selected based on the Bayesian information criterion proposed by Wang and Leng (2007) is much larger than the value of λ selected by the 5-fold cross-validation method. Denote the values of λ selected by these two methods as λ_{BIC} and λ_{CV}, respectively. We choose λ to be $(\lambda_{\mathrm{BIC}} + \lambda_{\mathrm{CV}})/2$ as a trade-off of these two methods.

S4.2 Estimating equation-based method

In our simulation studies and real-data application, we compare the proposed method with an estimating equation-based method, which is proposed by Neykov et al. (2018) and denoted as "EstEq." We apply their method based on Algorithm 1 in their paper. Using the same notations as in our paper, the implementation details are as follows:

Step 1: Use the R functions `gds` and `cv_gds` to get the generalized Dantzig selector of the regression coefficient $\boldsymbol{\gamma}_0 = (\alpha_0, \boldsymbol{\beta}_0^\top)^\top$ in a logistic regression model and denote the estimator as $\hat{\boldsymbol{\gamma}}$. That is, solve the following optimization problem to obtain an estimate $\hat{\boldsymbol{\gamma}}$:

$$\hat{\boldsymbol{\gamma}} = \arg\min \|\boldsymbol{\gamma}\|_1,$$
$$\text{subject to } \|\boldsymbol{t}(\boldsymbol{\gamma})\| = \left\| -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \right\|_\infty = \left\| -\frac{1}{n}\sum_{i=1}^{n} \{y_i - p_i(\boldsymbol{\gamma})\}\tilde{\mathbf{x}}_i \right\|_\infty \leq \lambda,$$

where $\ell_i(\boldsymbol{\gamma})$ is the conditional log-likelihood function of y_i given $\mathbf{x}_i$ for a logistic regression model and $p_i(\boldsymbol{\gamma}) = \exp(\tilde{\mathbf{x}}_i^\top \boldsymbol{\gamma})/\{1+\exp(\tilde{\mathbf{x}}_i^\top \boldsymbol{\gamma})\}$, $i = 1, \ldots, n$. The tuning parameter of the generalized Dantzig selector, λ, is selected by the 10-fold cross-validation method.

Step 2: Calculate the inverse of $\mathbf{T}(\hat{\boldsymbol{\gamma}}) = \partial \boldsymbol{t}(\hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\gamma}^\top = \tilde{\mathbf{X}}^\top \mathbf{D}(\hat{\boldsymbol{\gamma}})\tilde{\mathbf{X}}/n$, where $\mathbf{D}(\hat{\boldsymbol{\gamma}}) = \mathrm{diag}\{p_1(\hat{\boldsymbol{\gamma}})(1-p_1(\hat{\boldsymbol{\gamma}})), \ldots, p_n(\hat{\boldsymbol{\gamma}})(1-p_n(\hat{\boldsymbol{\gamma}}))\}$. Denote the inverse of $\mathbf{T}(\hat{\boldsymbol{\gamma}})$ as $\boldsymbol{\Omega}$. Define the projection direction for the jth element of $\boldsymbol{\beta}_0$, β_{j0}, as $\hat{\mathbf{v}}_j = \boldsymbol{\Omega}_{(j+1).}$, where $\boldsymbol{\Omega}_{(j+1).}$ is the $(j+1)$th row element of $\boldsymbol{\Omega}$. Note that in Neykov et al. (2018), the authors used the CLIME estimator to estimate the inverse of $\mathbf{T}(\hat{\boldsymbol{\gamma}})$. However, in our problem, we assume $n > p$ and p is fixed, then the inverse of $\mathbf{T}(\hat{\boldsymbol{\gamma}})$ can be calculated directly.

Step 3: Use the R function `uniroot` to solve the sparse projected test function and denote the estimated value of β_{j0} as $\tilde{\beta}_j$.

Step 4: Construct a two-sided $100(1-\alpha)\%$ confidence interval for β_{j0} as

$$\mathrm{CI}_j = \left(\tilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\hat{\sigma}_j/\sqrt{n}, \tilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\hat{\sigma}_j/\sqrt{n}\right),$$

where $\hat{\sigma}_j^2 = \hat{\mathbf{v}}_j^\top \tilde{\mathbf{X}}^\top \mathbf{D}(\hat{\boldsymbol{\gamma}})\tilde{\mathbf{X}}\hat{\mathbf{v}}_j/n$.

S4.3 Two types of bootstrap de-biased lasso methods

Motivated by the idea of Dezeure et al. (2017), we establish two xy-paired bootstrap de-biased lasso methods, which are referred to as "the type-I bootstrap de-biased lasso method" and "the type-II bootstrap de-biased lasso method," respectively. The bootstrap de-biased lasso method is based on the de-biased lasso method proposed by Zhang and Zhang (2014), Van de Geer et al. (2014) and Javanmard and Montanari (2014). Following the idea of Dezeure et al. (2017), the procedure for the type-I bootstrap de-biased lasso method is as follows:

(i) Based on the original data points $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, calculate the lasso estimator and de-biased lasso estimator of the jth element of $\boldsymbol{\beta}_0$, β_{j0}. Denote them as $\hat{b}_j$ and $\hat{\beta}_j$, respectively. Calculate the standard error of the de-biased lasso estimator, $\widehat{\text{s.e.}}_{\cdot j}$.

(ii) Resample $(\mathbf{X}_1^*, Y_1^*), \ldots, (\mathbf{X}_n^*, Y_n^*)$ with replacement from $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ for B times. For the kth bootstrap sample, calculate the de-biased lasso estimator $\hat{b}_{jk}^*$, the standard error for the de-biased lasso estimator $\widehat{\text{s.e.}}_{\cdot jk}^*$ and $T_{jk}^* = (\hat{b}_{jk}^* - \hat{\beta}_j)/\widehat{\text{s.e.}}_{\cdot jk}^*$. Denote the ν-quantile of $\{T_{j1}^*, \ldots, T_{jB}^*\}$ as $q_{j;\nu}^*$.

(iii) Construct a two-sided $100(1-\alpha)\%$ confidence interval for β_{j0} as

$$\text{CI}_j = \left(\hat{b}_j - q_{j;1-\alpha/2}^* \widehat{\text{s.e.}}_{\cdot j}, \hat{b}_j - q_{j;\alpha/2}^* \widehat{\text{s.e.}}_{\cdot j} \right).$$

In addition, the procedure for the type-II bootstrap de-biased lasso method is as follows:

(i) Resample $(\mathbf{X}_1^*, Y_1^*), \ldots, (\mathbf{X}_n^*, Y_n^*)$ with replacement from $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ for B times. For the kth bootstrap sample, calculate the de-biased lasso estimator of the jth element of $\boldsymbol{\beta}_0$, β_{j0}, which is denoted as $\hat{b}_{jk}^*$. Denote the ν-quantile of $\{\hat{b}_{j1}^*, \ldots, \hat{b}_{jB}^*\}$ as $q_{j;\nu}^*$.

(iii) Construct a two-sided $100(1-\alpha)\%$ confidence interval for β_{j0} as

$$\mathrm{CI}_j = \left(q^*_{j;\alpha/2}, q^*_{j;1-\alpha/2}\right).$$

S5 Additional Simulation Results

In this section, we present additional simulation results under the simulation settings in Section 5. Figures S1 and S2 display the results for different types of selection probability for $\boldsymbol{X}_4$ when $\rho = 0.2$ and 0.5, respectively. Figures S3 and S4 present the empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories as the value of θ varies when $\rho = 0.2$ and 0.5, respectively. Tables S1–S4 show the coverage probabilities and average widths of the 95% confidence intervals under all simulation settings. Figures S5–S7 show the simulation results for the proposed method when the threshold value δ_1 varies. Figures S8–S10 show the simulation results for the proposed method when the threshold value τ varies. Figures S11–S13 show the simulation results for the proposed method when the total number of weak signals varies.

S6 Additional Information in Real-data Application

Table S5 shows the candidate predictors used in the real-data application.

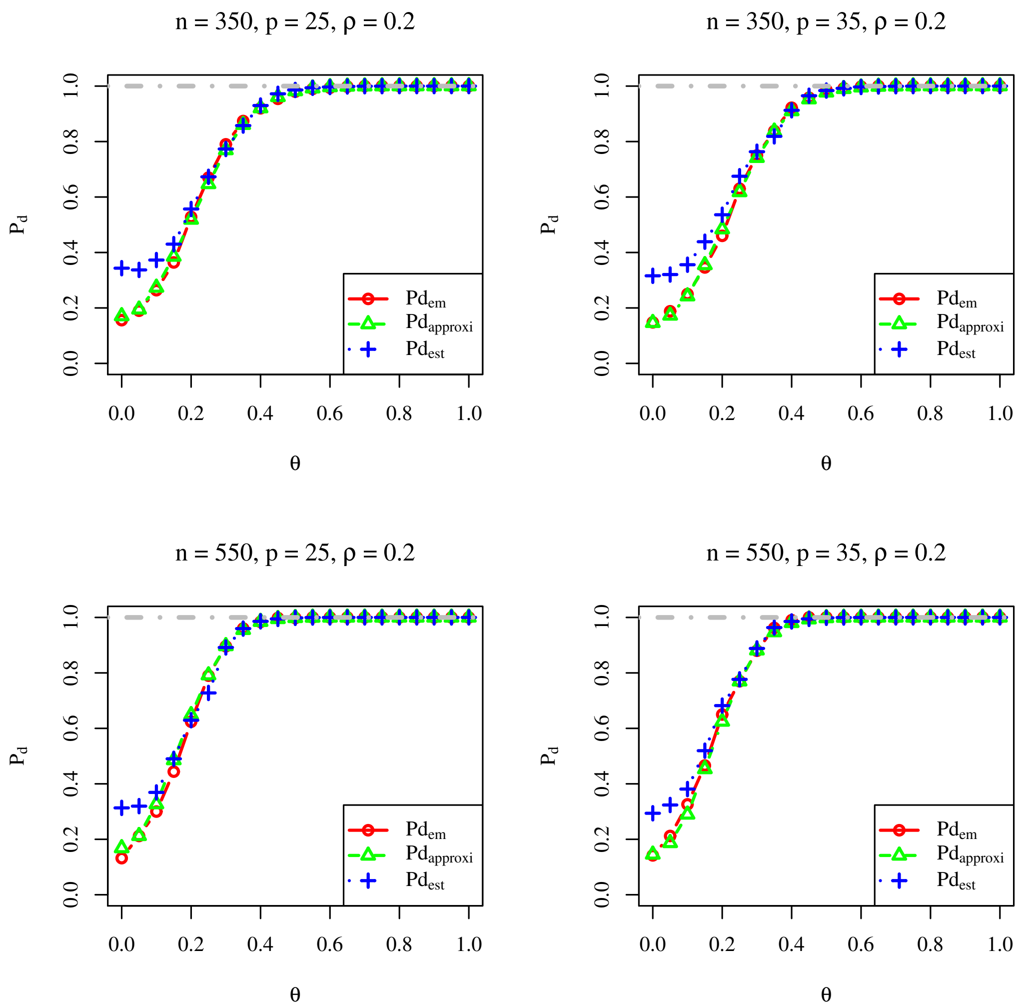

Figure S1: Different types of selection probability for $\boldsymbol{X}_4$ when $\rho = 0.2$. $\mathrm{Pd}_{\mathrm{em}}$: empirical selection probability, which equals the empirical probability of $\{\theta^{(1)} \neq 0\}$ based on 500 Monte Carlo samples; $\mathrm{Pd}_{\mathrm{approxi}}$: approximated selection probability based on (3.1), where the expectations in (3.1) are calculated by using the function `cubintegrate` in R; $\mathrm{Pd}_{\mathrm{est}}$: median of estimated selection probabilities based on (3.3) for 500 Monte Carlo samples.

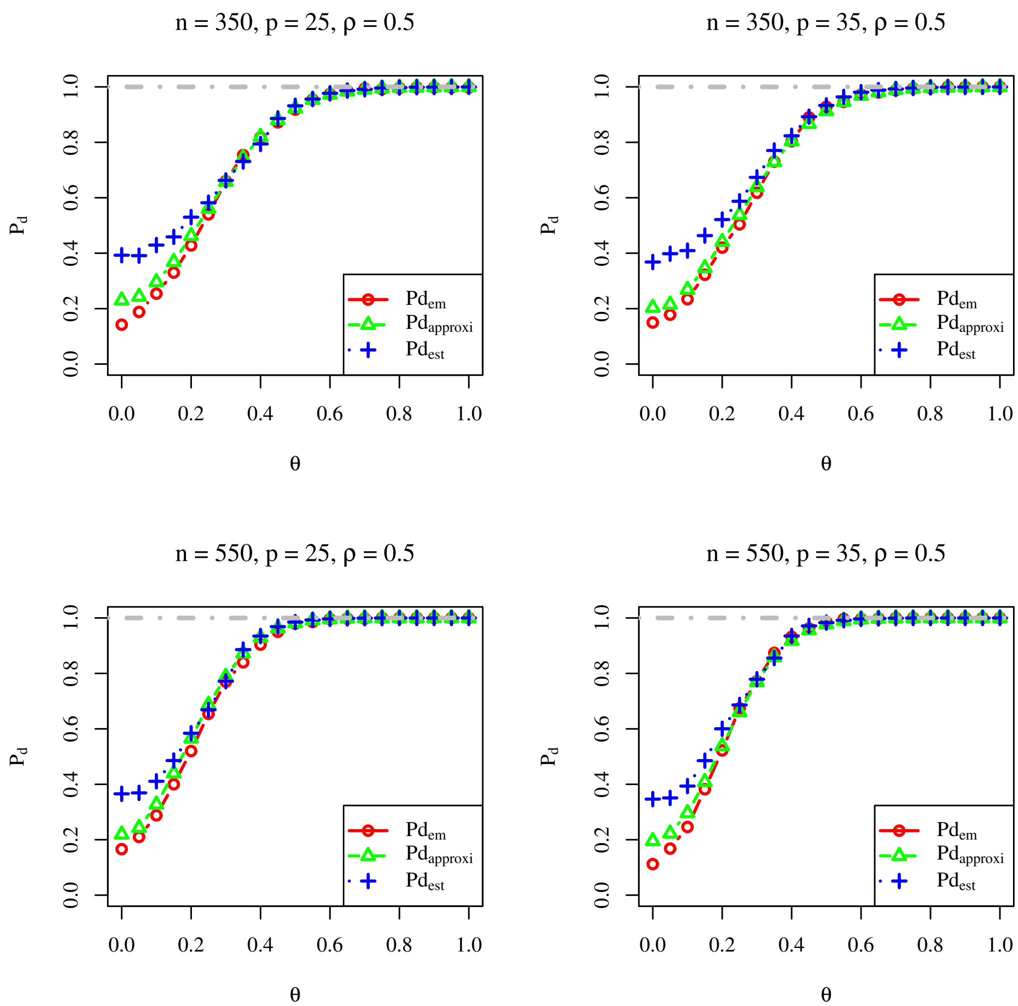

Figure S2: Different types of selection probability for $\boldsymbol{X}_4$ when $\rho = 0.5$. The meanings of notations: see Figure S1.

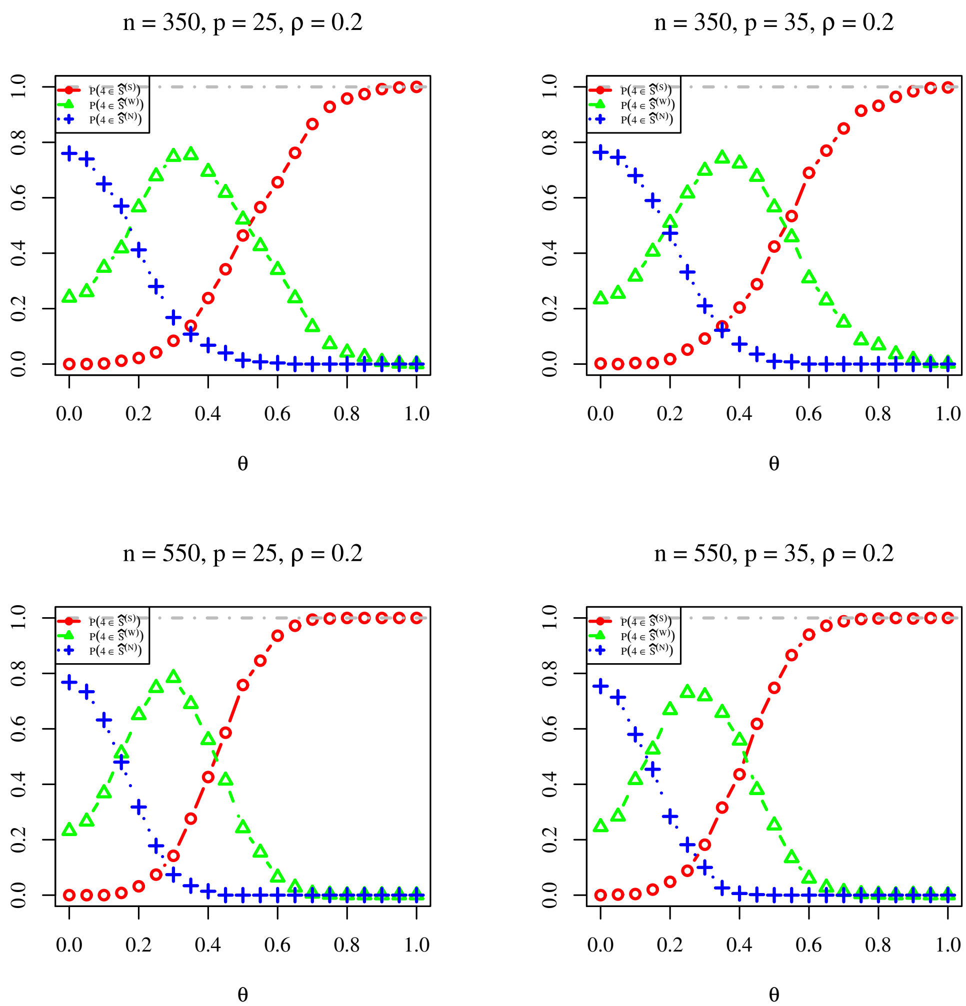

Figure S3: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $\rho = 0.2$.

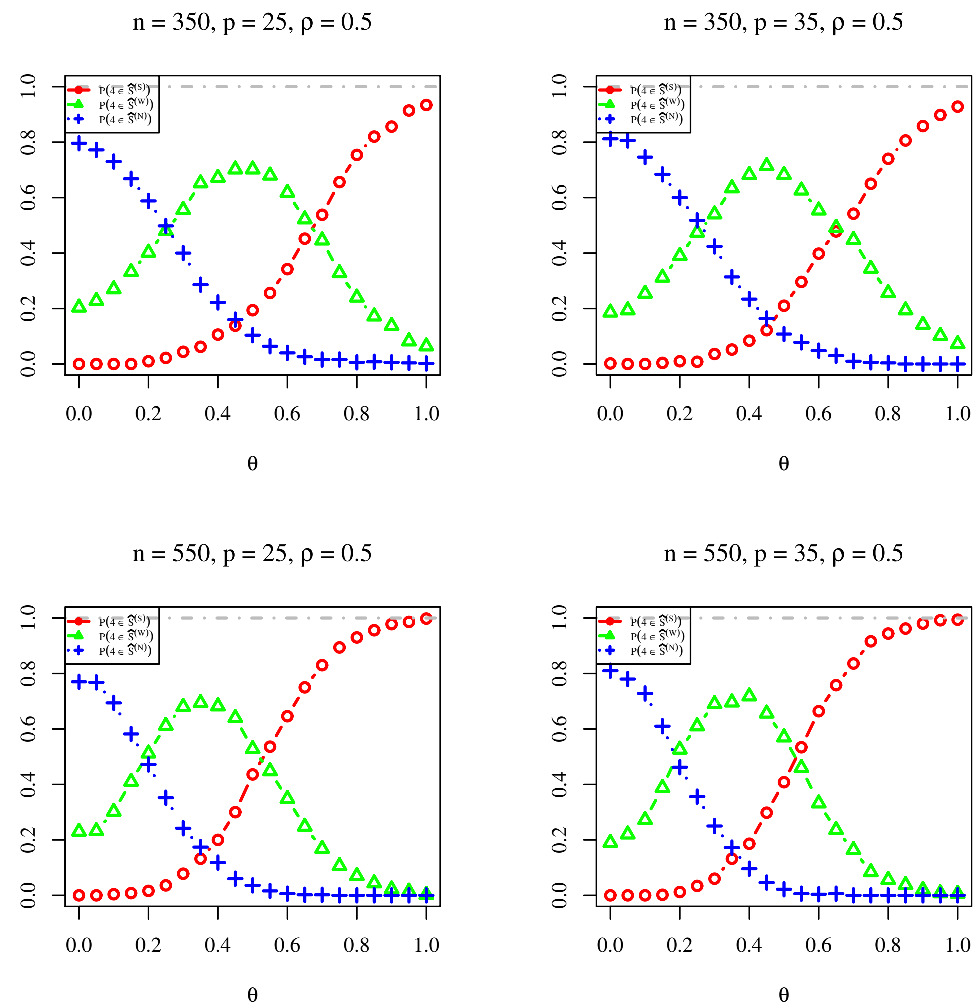

Figure S4: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $\rho = 0.5$.

Table S1: The coverage probabilities (%) of the 95% confidence intervals when the sample size is $n = 350$.

		$p = 25$			$p = 35$		
θ	Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
	Proposed	93.8	94.4	96.2	94.6	92.2	94.8
	OldTwostep	75.8	76.7	81.4	77.1	66.9	72.3
	Asym	3.6	3.8	12.7	4.3	1.4	4.0
	MLE	93.8	94.4	96.2	94.6	92.2	94.8
	Perturb	100.0	100.0	100.0	100.0	100.0	100.0
0	EstEq	94.0	94.2	96.6	95.6	92.8	94.8
	SdBS	99.8	100.0	99.8	99.8	99.8	99.0
	SmBS	100.0	100.0	99.8	100.0	100.0	100.0
	DeLasso	95.8	96.0	98.2	96.4	95.2	96.4
	BSDe1	99.8	100.0	99.8	100.0	100.0	100.0
	BSDe2	94.8	94.4	96.2	95.4	91.8	94.4
	Proposed	94.6	95.2	92.8	95.2	96.4	94.6
	OldTwostep	96.9	96.6	92.0	98.0	96.7	92.4
	Asym	75.5	71.6	61.5	65.8	69.6	69.3
	MLE	92.2	93.4	92.6	92.4	92.0	93.6
	Perturb	57.0	55.0	52.0	38.8	49.0	44.0
0.3	EstEq	92.2	92.6	93.8	92.6	91.6	94.2
	SdBS	72.0	69.6	62.8	53.0	61.0	53.4
	SmBS	65.2	64.6	59.8	39.8	49.4	47.8
	DeLasso	93.8	94.0	92.8	93.0	93.4	95.0
	BSDe1	52.0	58.0	85.6	48.6	60.6	86.4
	BSDe2	94.2	94.6	95.0	96.2	95.0	95.2
	Proposed	95.0	93.6	95.0	96.0	93.8	97.2
	OldTwostep	95.0	93.6	95.4	96.0	93.8	97.2
	Asym	95.0	93.6	91.6	96.0	93.8	92.2
	MLE	90.0	91.6	91.2	87.8	87.8	86.8
	Perturb	93.2	93.0	97.0	95.4	94.2	96.4
0.95	EstEq	90.6	87.4	92.8	89.8	89.4	89.4
	SdBS	93.8	93.8	95.6	93.4	93.4	95.6
	SmBS	87.2	87.8	90.2	68.6	69.6	74.8
	DeLasso	87.6	87.6	90.4	90.4	84.2	89.6
	BSDe1	23.0	26.0	34.8	17.8	15.4	26.4
	BSDe2	94.8	95.6	97.4	94.4	95.0	95.6

Note: 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.

Table S2: The coverage probabilities (%) of the 95% confidence intervals when the sample size is $n = 550$.

		$p = 25$			$p = 35$		
θ	Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
0	Proposed	95.4	94.8	95.4	94.6	94.2	95.8
	OldTwostep	81.7	77.6	80.0	75.9	76.4	78.9
	Asym	4.2	7.6	7.2	1.4	4.2	7.1
	MLE	95.4	94.8	95.4	94.6	94.2	95.8
	Perturb	99.8	100.0	100.0	100.0	100.0	100.0
	EstEq	95.6	93.8	95.6	95.2	95.0	96.8
	SdBS	99.8	99.6	99.6	100.0	100.0	100.0
	SmBS	99.8	100.0	100.0	100.0	100.0	100.0
	DeLasso	96.6	95.4	97.0	96.4	95.8	97.4
	BSDe1	99.8	100.0	99.8	100.0	100.0	100.0
	BSDe2	95.4	94.6	95.6	95.8	94.2	95.6
0.25	Proposed	94.4	95.6	95.0	95.4	93.8	95.6
	OldTwostep	95.8	96.6	94.8	97.0	95.1	94.7
	Asym	69.4	63.8	68.2	72.3	69.9	68.5
	MLE	94.4	95.6	94.4	93.8	92.0	95.2
	Perturb	57.4	52.8	56.2	54.8	55.2	54.6
	EstEq	93.6	95.0	93.8	93.4	91.4	94.8
	SdBS	68.8	65.2	62.8	65.0	66.8	62.0
	SmBS	67.8	66.0	63.6	61.6	62.8	64.4
	DeLasso	93.0	94.8	94.4	94.0	93.0	95.8
	BSDe1	52.8	57.2	79.2	49.2	57.4	79.6
	BSDe2	94.2	96.4	94.8	95.2	96.0	96.0
0.8	Proposed	94.2	94.4	93.8	95.0	95.0	92.2
	OldTwostep	94.2	94.4	93.8	95.0	95.0	92.2
	Asym	94.2	94.4	90.6	95.0	95.0	89.0
	MLE	93.6	94.4	92.6	90.4	89.4	91.2
	Perturb	90.2	93.0	97.0	93.8	94.2	95.8
	EstEq	92.4	93.0	90.6	90.4	92.4	91.6
	SdBS	91.2	93.8	96.2	91.8	91.6	94.2
	SmBS	88.4	93.8	94.4	87.0	86.0	91.2
	DeLasso	87.0	90.2	89.4	89.0	87.2	90.4
	BSDe1	23.0	26.0	41.2	15.8	18.8	33.8
	BSDe2	96.4	97.2	94.4	93.8	95.8	95.2

Note: 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.

Table S3: The widths (×100) of the 95% confidence intervals when the sample size is $n = 350$

		$p = 25$			$p = 35$		
θ	Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
0	Proposed	55.7	60.0	78.4	58.2	62.7	82.1
	OldTwostep	55.9	60.8	79.7	58.6	63.2	82.8
	Asym	19.6	21.6	22.3	19.7	18.9	23.3
	MLE	55.7	60.0	78.4	58.2	62.8	82.1
	Perturb	14.5	14.7	17.6	10.3	11.1	13.9
	EstEq	50.4	53.9	70.1	51.1	54.7	71.0
	SdBS	22.9	23.6	27.9	17.2	17.9	21.7
	SmBS	16.6	16.8	19.4	11.4	11.9	14.3
	DeLasso	48.7	51.9	66.8	49.4	52.6	67.5
	BSDe1	49.6	52.8	67.7	50.6	54.0	68.9
	BSDe2	58.7	63.2	82.8	63.6	69.0	90.6
0.3	Proposed	56.2	60.5	79.5	58.6	63.1	83.8
	OldTwostep	56.2	60.6	79.1	58.6	63.0	83.9
	Asym	33.5	34.0	35.0	30.2	32.8	35.8
	MLE	57.0	61.6	80.7	59.5	64.5	84.9
	Perturb	49.6	51.7	55.9	40.5	47.1	50.3
	EstEq	51.0	54.8	71.6	51.6	55.4	72.5
	SdBS	51.6	53.4	58.4	41.1	45.8	49.5
	SmBS	46.0	47.4	50.1	34.6	39.1	40.8
	DeLasso	49.4	52.9	68.3	49.7	53.2	68.6
	BSDe1	51.2	54.9	70.3	52.7	56.4	72.5
	BSDe2	62.8	67.6	88.0	68.8	74.8	98.5
0.95	Proposed	60.9	63.9	73.4	62.0	64.9	75.1
	OldTwostep	60.9	63.9	73.3	62.0	64.9	75.1
	Asym	60.9	63.8	71.0	62.0	64.8	71.8
	MLE	68.6	73.7	93.7	72.9	78.1	100.5
	Perturb	67.4	70.4	91.6	71.1	76.1	103.2
	EstEq	57.4	61.6	78.9	57.9	61.6	79.3
	SdBS	67.6	70.4	87.4	67.2	70.4	86.4
	SmBS	60.8	63.6	79.7	57.9	61.0	75.9
	DeLasso	53.5	56.8	72.9	53.6	57.0	73.2
	BSDe1	56.0	60.2	77.7	58.0	61.8	80.6
	BSDe2	84.5	91.8	115.8	100.8	108.1	137.6

Note: 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.

Table S4: The widths ($\times 100$) of the 95% confidence intervals when the sample size is $n = 550$

		$p = 25$			$p = 35$		
θ	Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
0	Proposed	42.7	45.9	59.9	43.7	47.0	61.5
	OldTwostep	42.8	46.2	60.3	44.0	47.2	61.6
	Asym	14.8	15.3	17.0	13.7	14.7	17.0
	MLE	42.7	45.9	59.9	43.7	47.0	61.5
	Perturb	12.6	13.3	17.1	9.7	10.8	12.7
	EstEq	39.8	42.7	55.5	40.1	42.8	55.8
	SdBS	19.4	19.9	25.2	16.0	17.3	20.3
	SmBS	15.1	15.3	19.4	11.8	12.9	14.7
	DeLasso	38.6	41.2	53.2	38.8	41.4	53.5
	BSDe1	38.8	41.4	53.1	39.1	41.6	53.6
	BSDe2	43.0	46.1	60.3	44.6	48.0	62.9
0.25	Proposed	42.7	46.2	60.6	43.7	47.2	62.3
	OldTwostep	42.7	46.2	60.6	43.7	47.2	62.0
	Asym	25.7	25.2	28.9	25.8	26.4	27.5
	MLE	43.4	46.7	61.2	44.5	48.0	62.9
	Perturb	40.7	41.7	47.7	39.1	41.2	46.2
	EstEq	40.2	43.1	56.3	40.4	43.3	56.7
	SdBS	42.4	43.8	49.8	40.0	41.7	47.8
	SmBS	40.2	41.4	46.0	37.3	39.0	43.6
	DeLasso	39.0	41.7	54.0	39.2	41.7	54.2
	BSDe1	39.9	42.7	54.8	40.4	43.5	55.7
	BSDe2	45.1	48.3	62.8	47.3	51.0	66.5
0.8	Proposed	45.5	47.8	54.9	46.1	48.1	54.8
	OldTwostep	45.5	47.8	54.9	46.1	48.1	54.8
	Asym	45.5	47.8	53.6	46.1	48.1	53.6
	MLE	49.4	53.1	68.0	51.1	54.7	70.2
	Perturb	50.5	53.3	69.3	51.5	53.5	70.2
	EstEq	43.9	47.1	60.8	44.2	47.2	60.9
	SdBS	49.3	52.0	66.2	48.9	50.9	64.4
	SmBS	48.9	51.6	65.8	47.3	49.6	63.2
	DeLasso	41.4	44.2	56.8	41.6	43.9	57.2
	BSDe1	42.9	45.8	59.2	43.3	46.7	60.4
	BSDe2	54.6	58.6	74.9	59.0	63.2	81.5

Note: 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.

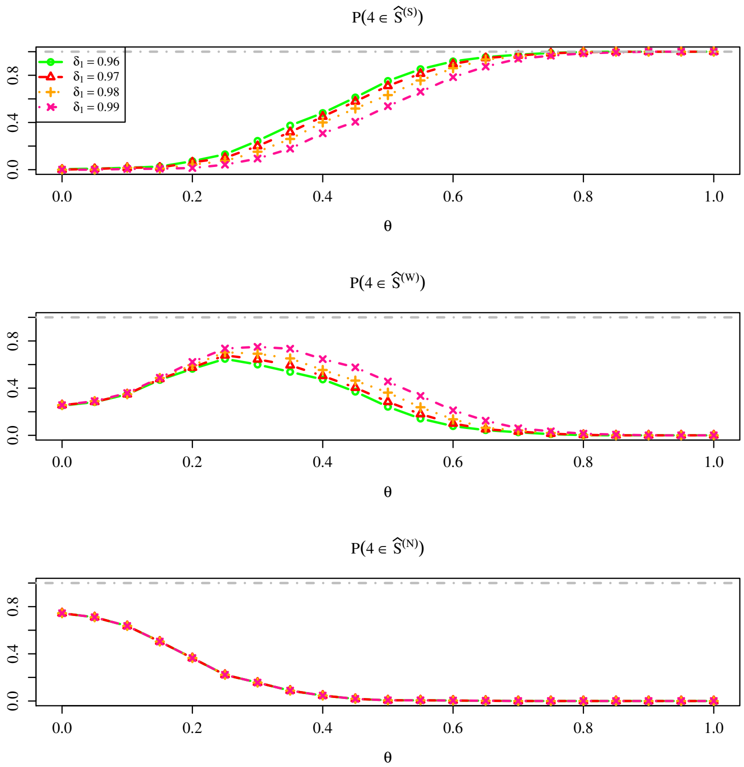

Figure S5: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $(n, p, \rho) = (350, 25, 0)$, $\tau = 0.1$ and the threshold value δ_1 varies.

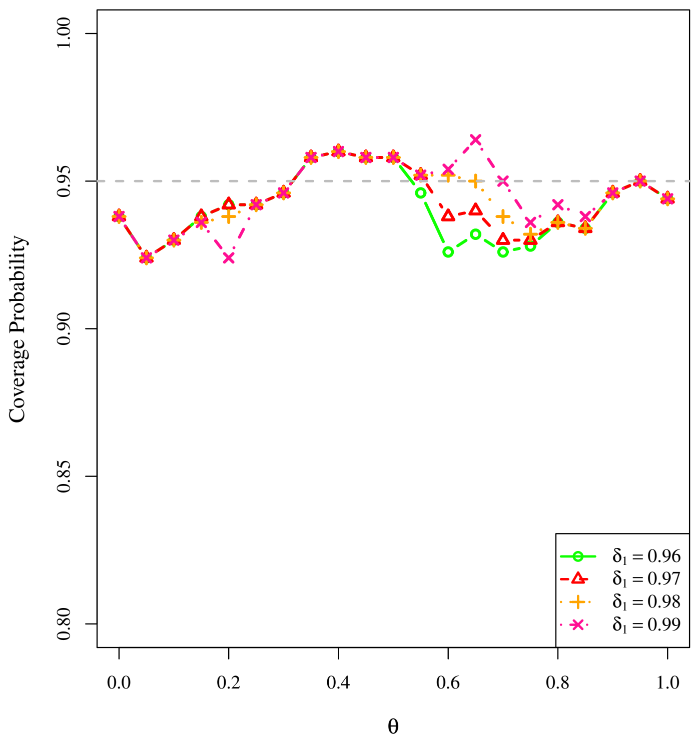

Figure S6: Coverage probabilities of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\tau = 0.1$ and the threshold value δ_1 varies.

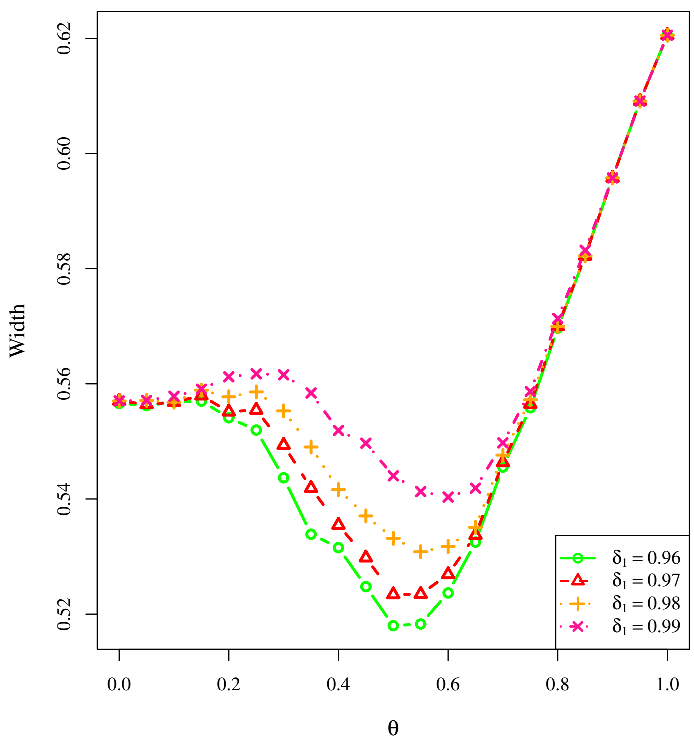

Figure S7: Average widths of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\tau = 0.1$ and the threshold value δ_1 varies.

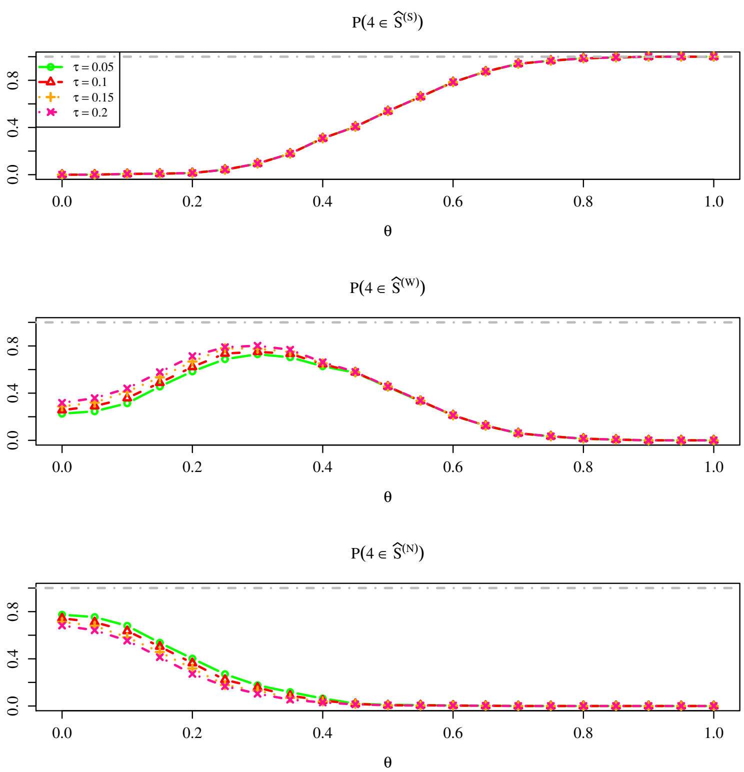

Figure S8: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$ and the threshold value τ varies.

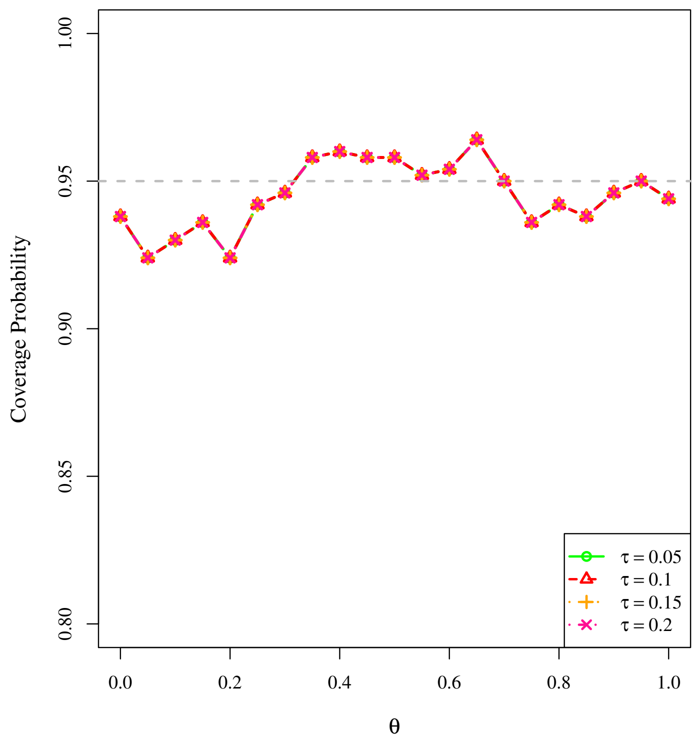

Figure S9: Coverage probabilities of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$ and the threshold value τ varies.

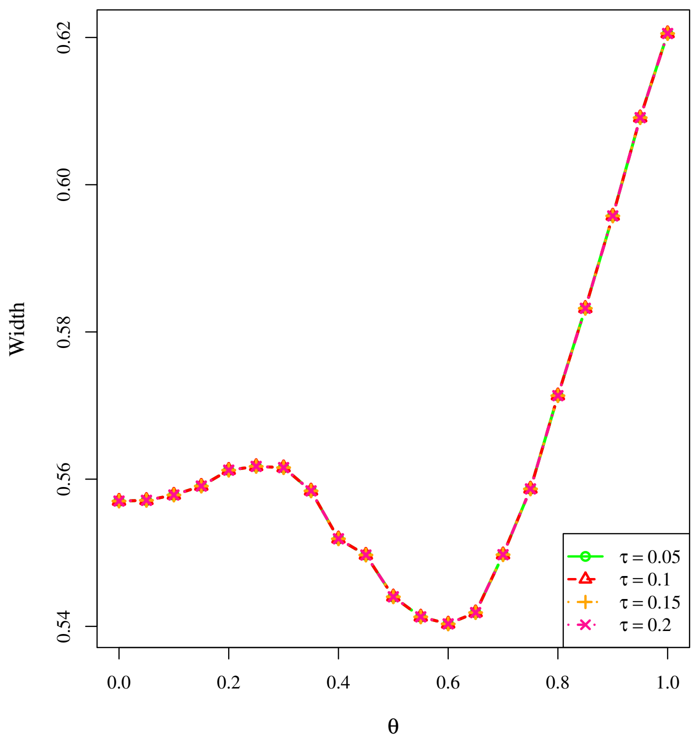

Figure S10: Average widths of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$ and the threshold value τ varies.

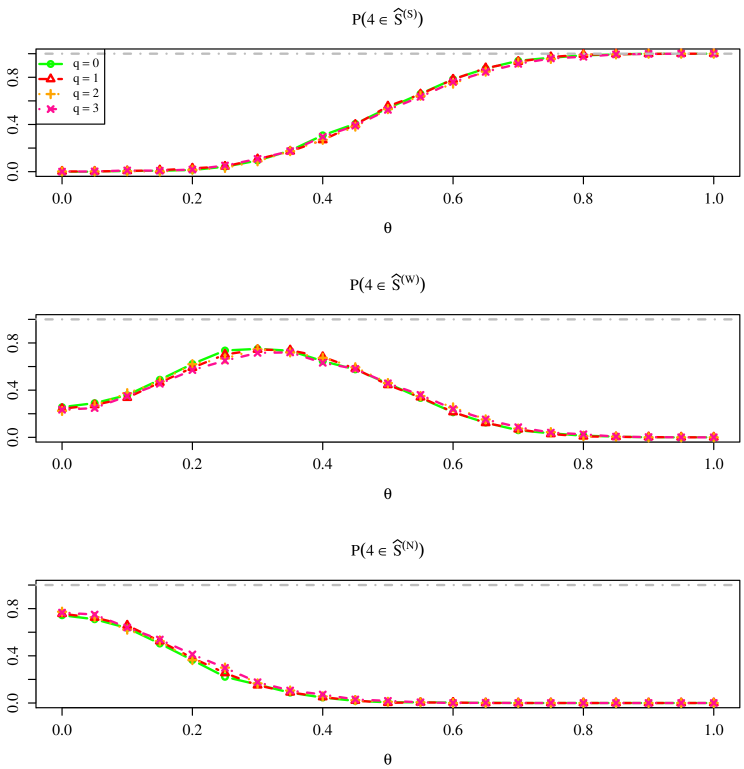

Figure S11: Empirical probabilities of assigning the covariate $\boldsymbol{X}_4$ to different signal categories when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$, $\tau = 0.1$ and the total number of weak signals varies.

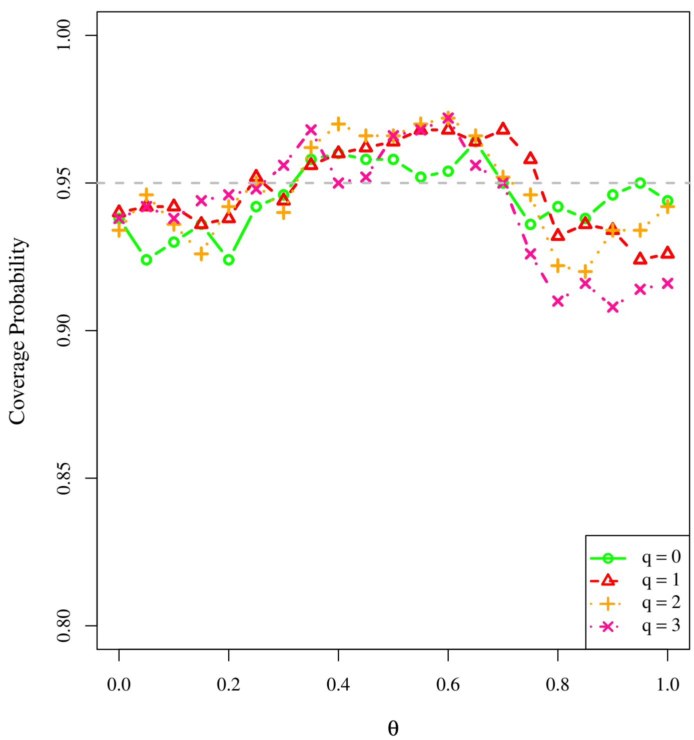

Figure S12: Coverage probabilities of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$, $\tau = 0.1$ and the total number of weak signals varies.

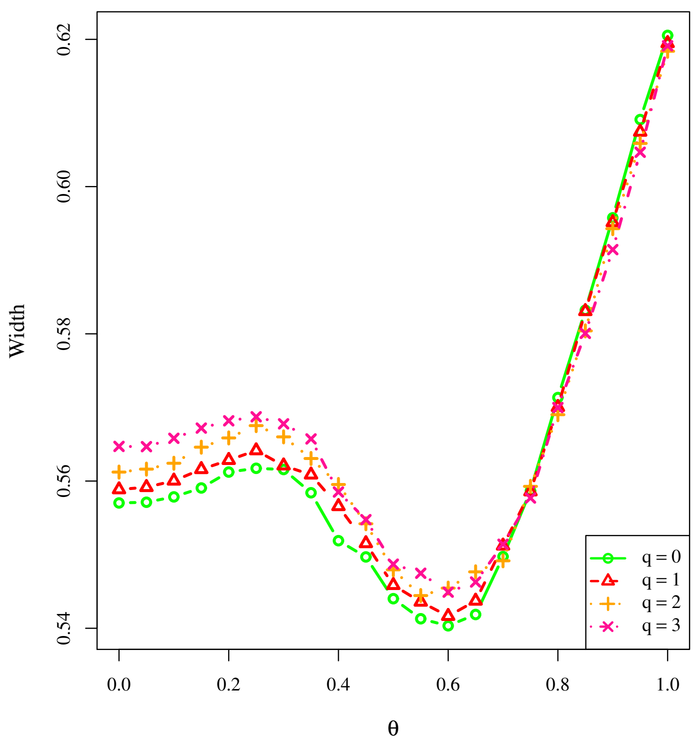

Figure S13: Average widths of the 95% confidence intervals for the proposed two-step inference method when $(n, p, \rho) = (350, 25, 0)$, $\delta_1 = 0.99$, $\tau = 0.1$ and the total number of weak signals varies.

Table S5: The candidate predictors used in the real-data analysis

Category	Predictor
Basic information	year of birth
	gender
	3 predictors indicating whether a patient is from California, Texas, New York or other states
Transcript records	range of BMI
	the median of weights
	the median of heights
	the median of systolic blood pressures
	the medians of Diastolic blood pressures
	the median of respiratory rates
	the median of temperatures
	4 predictors corresponding to the numbers of transcripts for different physician specialties
	number of physicians
	number of transcripts with blank visit year
	number of visits per weighted year
Diagnosis information	69 predictors corresponding to the numbers of times being diagnosed with different diagnoses
	number of diagnoses per weighted year
	number of different 3 digits diagnostics groups in the icd9 table
	number of different 3 digits diagnostics groups with medication
Medication information	23 predictors indicating the dose of active principle
	number of prescriptions or the use of different medications
	number of medications without prescript
	number of active principles
Lab result	1 binary variable indicating whether a patient has any lab test or not
Smoking status	1 binary variable indicating whether a patient smoked in the past

References

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

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.

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.

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

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.

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

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.