BIOMETRICS 00, 1–12 DOI: xxx
00 0000

Supporting Information for "De-biased Lasso for Generalized Linear Models with A Diverging Number of Covariates" by

Lu Xia[1], Bin Nan[2], and Yi Li[3]

[1] Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.

[2] Department of Statistics, University of California, Irvine, Irvine, CA, U.S.A.

[3] Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.

email: nanb@uci.edu; yili@umich.edu

SUMMARY: We present the technical proofs of Theorem 1 in the main text and the lemmas used for the theorem in Web Appendix A. Additional simulation results and descriptive statistics for the Boston Lung Cancer Survivor Cohort are provided in Web Appendix B and Web Appendix C, respectively. Web Appendix D entails additional discussion on the difference between sparsity assumptions in our work and van de Geer et al. (2014).

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A. Technical proofs

Web Appendix A.1 *Lemmas*

We list three lemmas that are used for proving Theorem 1. Without loss of generality, we denote the dimension of the parameter ξ by p instead of $(p+1)$ to simplify the notation in the proofs. Consequently, the matrices such as Σ_ξ and Θ_ξ are considered as $p \times p$ matrices. This simplification of notation does not affect the following derivations.

Lemma 1 bounds the estimation error and the prediction error of the lasso estimator under our assumptions.

LEMMA 1: *Under Assumptions 1–5, we have* $\|\widehat{\xi} - \xi^0\|_1 = \mathcal{O}_P(s_0\lambda)$ *and* $\|X(\widehat{\xi} - \xi^0)\|_2^2/n = \mathcal{O}_P(s_0\lambda^2)$.

Proof. Because $\lambda_{\min}(\Sigma_{\xi^0}) > 0$ in Assumption 2, the compatibility condition holds for all index sets $S \subset \{1, \ldots, p\}$ by Lemma 6.23 (Bühlmann and van de Geer, 2011) and the fact that the adaptive restricted eigenvalue condition implies the compatibility condition. Exploiting Hoeffding's concentration inequality, we have $\|\widehat{\Sigma}_{\xi^0} - \Sigma_{\xi^0}\|_\infty = \mathcal{O}_P[\{log(p)/n\}^{1/2}]$. Then by Lemma 6.17 of Bühlmann and van de Geer (2011), we have the $\widehat{\Sigma}_{\xi^0}$-compatibility condition. Finally, the first part of Lemma 1 follows from Theorem 6.4 in Bühlmann and van de Geer (2011).

For the second claim, Ning and Liu (2017) showed that

$$(\widehat{\xi} - \xi^0)^T \widehat{\Sigma}_{\xi^0} (\widehat{\xi} - \xi^0) = (\widehat{\xi} - \xi^0)^T (X^T W_{\xi^0}^2 X/n)(\widehat{\xi} - \xi^0) = \mathcal{O}_P(s_0\lambda^2),$$

then under Assumption 4, the variance terms in $W_{\xi^0}^2$ are bounded away from 0, and we obtain the desired result that $\|X(\widehat{\xi} - \xi^0)\|_2^2/n = \mathcal{O}_P(s_0\lambda^2)$.

Lemma 2 depicts the convergence rate of the inverse Hessian matrix $\widehat{\Theta}$ to the true inverse information matrix Θ_{ξ^0}.

LEMMA 2: *Suppose the covariate vectors* $x_i,\ i = 1, \ldots, n$, *are independent and identically distributed sub-Gaussian random vectors. Under Assumptions 1–5, if we further assume that*

$s_0\lambda \to 0$ and $p/n \to 0$, then $\widehat{\Theta}$ converges to Θ_{ξ^0} such that

$$\|\widehat{\Theta} - \Theta_{\xi^0}\| = \mathcal{O}_P\{(p/n)^{1/2} + s_0\lambda\}.$$

Proof. Since $\widehat{\Sigma}_{\widehat{\xi}}^{-1} - \Sigma_{\xi^0}^{-1} = \widehat{\Sigma}_{\widehat{\xi}}^{-1}\left(\Sigma_{\xi^0} - \widehat{\Sigma}_{\widehat{\xi}}\right)\Sigma_{\xi^0}^{-1}$, we have

$$\|\widehat{\Sigma}_{\widehat{\xi}}^{-1} - \Sigma_{\xi^0}^{-1}\| \leqslant \|\widehat{\Sigma}_{\widehat{\xi}}^{-1}\| \cdot \|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| \cdot \|\Sigma_{\xi^0}^{-1}\|. \tag{1}$$

By Assumption 2, $\|\Sigma_{\xi^0}^{-1}\|$ is bounded. We obtain the convergence rate of $\|\widehat{\Sigma}_{\widehat{\xi}}^{-1} - \Sigma_{\xi^0}^{-1}\|$ by calculating the convergence rate of $\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\|$ and showing that $\|\widehat{\Sigma}_{\widehat{\xi}}^{-1}\|$ is bounded with probability going to 1.

Note that $\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| \leqslant \|\widehat{\Sigma}_{\widehat{\xi}} - \widehat{\Sigma}_{\xi^0}\| + \|\widehat{\Sigma}_{\xi^0} - \Sigma_{\xi^0}\|$. When the rows of X are sub-Gaussian, so are the rows of X_{ξ^0} due to the boundedness of the weights w_i in Assumption 3. It can be shown that $L = \|\Sigma_{\xi^0}^{-1/2} x_1 \omega_1(\xi^0)\|_{\psi_2} = \mathcal{O}(1)$. First, for $\|\widehat{\Sigma}_{\xi^0} - \Sigma_{\xi^0}\|$, Vershynin (2012) shows that for every $t > 0$, it holds with probability at least $1 - 2\exp(-c_L' t^2)$ that

$$\|\widehat{\Sigma}_{\xi^0} - \Sigma_{\xi^0}\| \leqslant \|\Sigma_{\xi^0}\| \max(\delta, \delta^2) \leqslant c_{\max} \max(\delta, \delta^2), \tag{2}$$

where $\delta = C_L(p/n)^{1/2} + t/n^{1/2}$. Here C_L, $c_L' > 0$ depend only on $L = \|\Sigma_{\xi^0}^{-1/2} x_1 \omega_1(\xi^0)\|_{\psi_2}$. In fact $c_L' = c_1/L^4$ and $C_L = L^2(\log 9/c_1)^{1/2}$, where c_1 is an absolute constant. For $s > 0$ and $t = sC_L p^{1/2}$, the probability becomes $1 - 2\exp(-c_2 s^2 p)$, $c_2 > 0$ being some absolute constant, and $\delta = (s+1)C_L(p/n)^{1/2}$. Thus $\|\widehat{\Sigma}_{\xi^0} - \Sigma_{\xi^0}\| = \mathcal{O}_P\{L^2(p/n)^{1/2}\} = \mathcal{O}_P\{(p/n)^{1/2}\}$.

Note that

$$\begin{aligned}
\|\widehat{\Sigma}_{\widehat{\xi}} - \widehat{\Sigma}_{\xi^0}\| &= \|X^T(W_{\widehat{\xi}}^2 - W_{\xi^0}^2)X/n\| \\
&\leqslant \|X^T\| \cdot \|X\|/n \cdot \|W_{\widehat{\xi}}^2 - W_{\xi^0}^2\| \\
&= \lambda_{\max}(X^T X/n) \cdot \|W_{\widehat{\xi}}^2 - W_{\xi^0}^2\|.
\end{aligned}$$

By Assumptions 1 and 3,

$$\begin{aligned}
\|W_{\widehat{\xi}}^2 - W_{\xi^0}^2\| &= \max_i |\ddot{\rho}(y_i, x_i^T\widehat{\xi}) - \ddot{\rho}(y_i, x_i^T\xi^0)| \\
&\leqslant c_{Lip} \cdot \max_i |x_i^T(\widehat{\xi} - \xi^0)| \\
&\leqslant c_{Lip} K \cdot \|\widehat{\xi} - \xi^0\|_1.
\end{aligned} \tag{3}$$

By Lemma 1, we have $\|\widehat{\xi} - \xi^0\|_1 = \mathcal{O}_P(s_0\lambda)$. In this case, $\|W_{\widehat{\xi}}^2 - W_{\xi^0}^2\| = \mathcal{O}_P(s_0\lambda)$. By Assumption

5 and Vershynin (2012), $\lambda_{\max}(X^TX/n) = \mathcal{O}_P(1)$. Thus $\|\widehat{\Sigma}_{\widehat{\xi}} - \widehat{\Sigma}_{\xi^0}\| = \mathcal{O}_P(s_0\lambda)$. Therefore, after combining the two parts, we have $\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| = \mathcal{O}_P\{L^2(p/n)^{1/2} + s_0\lambda\}$. Under $p/n = o(1)$ and $s_0\lambda = o(1)$, we have $\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| = o_P(1)$.

Now for any vector x with $\|x\|_2 = 1$, we have

$$\inf_{\|y\|_2=1} \|\widehat{\Sigma}_{\widehat{\xi}} y\|_2 \leqslant \|\widehat{\Sigma}_{\widehat{\xi}} x\|_2 \leqslant \|\Sigma_{\xi^0} x\|_2 + \|(\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0})x\|_2 \leqslant \|\Sigma_{\xi^0} x\|_2 + \sup_{\|z\|_2=1} \|(\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0})z\|_2,$$

which indicates that $\lambda_{\min}(\widehat{\Sigma}_{\widehat{\xi}}) \leqslant \lambda_{\min}(\Sigma_{\xi^0}) + \|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\|$. Similarly, we have $\lambda_{\min}(\Sigma_{\xi^0}) \leqslant \lambda_{\min}(\widehat{\Sigma}_{\widehat{\xi}}) + \|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\|$. So $|\lambda_{\min}(\Sigma_{\xi^0}) - \lambda_{\min}(\widehat{\Sigma}_{\widehat{\xi}})| \leqslant \|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\|$. Thus, for any $0 < \epsilon < \min\{\|\Sigma_{\xi^0}\|, \lambda_{\min}(\Sigma_{\xi^0})/2\}$, we have that

$$\begin{aligned} pr\left(\|\widehat{\Sigma}_{\widehat{\xi}}^{-1}\| \geqslant \frac{1}{\lambda_{\min}(\Sigma_{\xi^0}) - \epsilon}\right) &= pr(\lambda_{\min}(\widehat{\Sigma}_{\widehat{\xi}}) \leqslant \lambda_{\min}(\Sigma_{\xi^0}) - \epsilon) \\ &\leqslant pr(|\lambda_{\min}(\widehat{\Sigma}_{\widehat{\xi}}) - \lambda_{\min}(\Sigma_{\xi^0})| \geqslant \epsilon) \\ &\leqslant pr(\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| \geqslant \epsilon). \end{aligned}$$

Since $\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\| = o_P(1)$, we have $\|\widehat{\Sigma}_{\widehat{\xi}}^{-1}\| = \mathcal{O}_P(1)$. Finally, by (1), $\|\widehat{\Sigma}_{\widehat{\xi}}^{-1} - \Sigma_{\xi^0}^{-1}\| = \mathcal{O}_P(\|\widehat{\Sigma}_{\widehat{\xi}} - \Sigma_{\xi^0}\|) = \mathcal{O}_P\{(p/n)^{1/2} + s_0\lambda\}$.

LEMMA 3: *Under Assumptions 1–3, when $p/n \to 0$, it holds that for any vector $\alpha_n \in \mathbb{R}^p$ with $\|\alpha_n\|_2 = 1$,*

$$\frac{n^{1/2}\alpha_n^T\Theta_{\xi^0}P_n\dot{\rho}_{\xi^0}}{(\alpha_n^T\Theta_{\xi^0}\alpha_n)^{1/2}} \to N(0,1)$$

in distribution as $n \to \infty$.

Proof. We invoke the Lindeberg-Feller Central Limit Theorem. For $i = 1, \ldots, n$, let

$$Z_{ni} = n^{-1/2}\alpha_n^T\Theta_{\xi^0}\dot{\rho}_{\xi^0}(y_i, x_i) = n^{-1/2}\alpha_n^T\Theta_{\xi^0}x_i\dot{\rho}(y_i, x_i^T\xi^0),$$

and $s_n^2 = Var\left(\sum_{i=1}^n Z_{ni}\right)$. Note that $E\{\dot{\rho}(y_i, x_i^T\xi^0) \mid x_i\} = 0$ and consequently $E(Z_{ni}) = 0$. Because $\{(y_i, \widetilde{x}_i)\}_{i=1}^n$ are independent and identically distributed, we can show that $s_n^2 = \alpha_n^T\Theta_{\xi^0}\alpha_n$. To show $\sum_{i=1}^n Z_{ni}/s_n \to N(0,1)$ in distribution, we first check the Lindeberg condition and then the conclusion shall follow by the Lindeberg-Feller Central Limit Theorem. Specifically, for any

$\epsilon > 0$, we show that as $n \to \infty$,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} E\left\{ Z_{ni}^2 \cdot 1_{(|Z_{ni}| > \epsilon s_n)} \right\} \to 0.$$

Due to the boundedness of the eigenvalues of Σ_{ξ^0}, $\alpha_n^T \Theta_{\xi^0} \alpha_n \geqslant \lambda_{\min}(\Theta_{\xi^0}) = 1/\lambda_{\max}(\Sigma_{\xi^0}) \geqslant c_{\max}^{-1}$. On the other hand, by the Cauchy-Schwarz inequality, it holds almost surely that

$$\left(\alpha_n^T \Theta_{\xi^0} x_i\right)^2 \leqslant \|\alpha_n\|_2^2 \cdot \|\Theta_{\xi^0} x_i\|_2^2 \leqslant \left[\|\Theta_{\xi^0}\| \cdot \|x_i\|_2\right]^2 \leqslant c_{\min}^{-2} \cdot \mathcal{O}(pK^2).$$

Inside the indicator, it holds almost surely that

$$\begin{aligned} \frac{Z_{ni}^2}{s_n^2} &= \frac{[\dot{\rho}(y_i, x_i^T \xi_0)]^2 \left(\alpha_n^T \Theta_{\xi^0} x_i\right)^2}{n \alpha_n^T \Theta_{\xi^0} \alpha_n} \\ &\leqslant [\dot{\rho}(y_i, x_i^T \xi_0)]^2 \cdot c_{\min}^{-2} c_{\max} \cdot \mathcal{O}(K^2 \frac{p}{n}) \\ &\leqslant K_1^2 c_{\min}^{-2} c_{\max} \cdot \mathcal{O}(K^2 \frac{p}{n}), \end{aligned}$$

where the last inequality follows from the boundedness of $\dot{\rho}(y_i, x_i^T \xi_0)$ in Assumption 3. Hence, we have $Z_{ni}^2/s_n^2 \to 0$ almost surely as $p/n \to 0$. When n is large enough, $Z_{ni}^2/s_n^2 < \epsilon^2$ and all the indicators become 0. Therefore, the Lindeberg condition holds and the Lindeber-Feller Central Limit Theorem guarantees the asymptotic normality.

Web Appendix A.2 *Proof of Theorem 1*

The invertibility of $\widehat{\Sigma}_{\widehat{\xi}}$ is shown in the proof of Lemma 2. Now with the bias decomposition Eq. (6) in the main text,

$$n^{1/2} \alpha_n^T (\widehat{b} - \xi^0) - n^{1/2} \alpha_n^T \widehat{\Theta} \Delta = -n^{1/2} \alpha_n^T \widehat{\Theta} P_n \dot{\rho}_{\xi^0},$$

we first show that $\alpha_n^T \widehat{\Theta} \alpha_n - \alpha_n^T \Theta_{\xi^0} \alpha_n = o_P(1)$ and that

$$n^{1/2} \alpha_n^T \widehat{\Theta} P_n \dot{\rho}_{\xi^0} / (\alpha_n^T \widehat{\Theta} \alpha_n)^{1/2} = n^{1/2} \alpha_n^T \Theta_{\xi^0} P_n \dot{\rho}_{\xi^0} / (\alpha_n^T \Theta_{\xi^0} \alpha_n)^{1/2} + o_P(1),$$

Then by Slutsky's Theorem, the asymptotic distribution of the target $n^{1/2} \alpha_n^T \widehat{\Theta} P_n \dot{\rho}_{\xi^0} / (\alpha_n^T \widehat{\Theta} \alpha_n)^{1/2}$ can be derived by using the asymptotic distribution of $n^{1/2} \alpha_n^T \Theta_{\xi^0} P_n \dot{\rho}_{\xi^0} / (\alpha_n^T \Theta_{\xi^0} \alpha_n)^{1/2}$, which has been proved in Lemma 3. In the final step, as long as $n^{1/2} \alpha_n^T \widehat{\Theta} \Delta = o_P(1)$, the asymptotic distribution of $n^{1/2} \alpha_n^T (\widehat{b} - \xi^0) / (\alpha_n^T \widehat{\Theta} \alpha_n)^{1/2}$ follows immediately.

According to Lemma 2, it follows that

$$|\alpha_n^T\widehat{\Theta}\alpha_n - \alpha_n^T\Theta_{\xi^0}\alpha_n| = |\alpha_n^T(\widehat{\Theta} - \Theta_{\xi^0})\alpha_n| \leqslant \|\widehat{\Theta} - \Theta_{\xi^0}\| \cdot \|\alpha_n\|_2^2 = o_P(1).$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned} n^{1/2}|\alpha_n^T\widehat{\Theta}P_n\dot{\rho}_{\xi^0} - \alpha_n^T\Theta_{\xi^0}P_n\dot{\rho}_{\xi^0}| &\leqslant n^{1/2}\|\alpha_n\|_2 \cdot \|(\widehat{\Theta} - \Theta_{\xi^0})P_n\dot{\rho}_{\xi^0}\|_2 \\ &\leqslant n^{1/2}\|\widehat{\Theta} - \Theta_{\xi^0}\| \cdot \|P_n\dot{\rho}_{\xi^0}\|_2, \end{aligned}$$

then we have

$$n^{1/2}|\alpha_n^T\widehat{\Theta}P_n\dot{\rho}_{\xi^0} - \alpha_n^T\Theta_{\xi^0}P_n\dot{\rho}_{\xi^0}| \leqslant n^{1/2} \cdot \|P_n\dot{\rho}_{\xi^0}\|_2 \cdot \mathcal{O}_P\{(p/n)^{1/2} + s_0\lambda\}.$$

By definition,

$$\begin{aligned} \|P_n\dot{\rho}_{\xi^0}\|_2^2 &= \sum_{j=1}^{p}\{n^{-1}\sum_{i=1}^{n} x_{ij}\dot{\rho}(y_i, x_i^T\xi^0)\}^2 \\ &= n^{-2}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{n} x_{ij}x_{kj}\dot{\rho}(y_i, x_i^T\xi^0)\dot{\rho}(y_k, x_k^T\xi^0). \end{aligned}$$

With independent observations and $E\{x_{ij}\dot{\rho}(y_i, x_i^T\xi^0)\} = 0$ for any i, it follows that

$$E\|P_n\dot{\rho}_{\xi^0}\|_2^2 = \frac{1}{n^2}\sum_{j=1}^{p}\sum_{i=1}^{n} E\{x_{ij}^2\dot{\rho}^2(y_i, x_i^T\xi^0)\}.$$

By Assumptions 1 and 3, we have $|x_{ij}\dot{\rho}(y_i, x_i^T\xi^0)| \leqslant KK_1$ almost surely holds for all i and j, so $E\|P_n\dot{\rho}_{\xi^0}\|_2^2 = \mathcal{O}(p/n)$. This implies that $\|P_n\dot{\rho}_{\xi^0}\|_2 = \mathcal{O}_P\{(p/n)^{1/2}\}$. Then we have

$$n^{1/2}|\alpha_n^T\widehat{\Theta}P_n\dot{\rho}_{\xi^0} - \alpha_n^T\Theta_{\xi^0}P_n\dot{\rho}_{\xi^0}| \leqslant \mathcal{O}_P\left(p/n^{1/2} + s_0\lambda p^{1/2}\right),$$

which is $o_P(1)$ by our assumption in Theorem 1.

Finally, we prove $|n^{1/2}\alpha_n^T\widehat{\Theta}\Delta| = o_P(1)$. By the Cauchy-Schwartz inequality, $|n^{1/2}\alpha_n^T\widehat{\Theta}\Delta| \leqslant n^{1/2}\|\widehat{\Theta}\Delta\|_2$, we only need that $n^{1/2}\|\widehat{\Theta}\Delta\|_2 = o_P(1)$. Note that

$$\Delta_j = \frac{1}{n}\sum_{i=1}^{n}\left\{\ddot{\rho}(y_i, a_i^*) - \ddot{\rho}(y_i, x_i^T\widehat{\xi}\,)\right\}x_{ij}x_i^T(\xi^0 - \widehat{\xi}\,),$$

where a_i^* lies between $x_i^T\widehat{\xi}$ and $x_i^T\xi^0$, i.e. $|a_i^* - x_i^T\widehat{\xi}\,| \leqslant |x_i^T(\widehat{\xi} - \xi^0)|$. Then uniformly for all j,

$$\begin{aligned}
|\Delta_j| &\leqslant \frac{1}{n}\sum_{i=1}^{n} |\ddot{\rho}(y_i, a_i^*) - \ddot{\rho}(y_i, x_i^T\widehat{\xi})| \cdot |x_{ij}| \cdot |x_i^T(\xi^0 - \widehat{\xi})| \\
&\leqslant \frac{1}{n}\sum_{i=1}^{n} c_{Lip}|a_i^* - x_i^T\widehat{\xi}| \cdot K \cdot |x_i^T(\xi^0 - \widehat{\xi})| \\
&\leqslant c_{Lip}K \cdot \frac{1}{n}\sum_{i=1}^{n} |x_i^T(\xi^0 - \widehat{\xi})|^2 \\
&= c_{Lip}K \cdot \mathcal{O}_P(s_0\lambda^2) \\
&= \mathcal{O}_P(s_0\lambda^2),
\end{aligned}$$

where the last equality holds by Lemma 1. Since $\|\Theta_{\xi^0}\| = \mathcal{O}(1)$ and $\|\widehat{\Theta} - \Theta_{\xi^0}\| = o_P(1)$, it follows that $\|\widehat{\Theta}\| = \mathcal{O}_P(1)$, and

$$\begin{aligned}
n^{1/2}\|\widehat{\Theta}\Delta\|_2 &\leqslant n^{1/2}\|\widehat{\Theta}\| \cdot \|\Delta\|_2 \\
&\leqslant n^{1/2}\mathcal{O}_P(1) \cdot p^{1/2}\|\Delta\|_\infty \\
&\leqslant \mathcal{O}_P((np)^{1/2}s_0\lambda^2).
\end{aligned}$$

By the assumption of $(np)^{1/2}s_0\lambda^2 = o(1)$ in Theorem 1, $n^{1/2}\|\widehat{\Theta}\Delta\|_2 = o_P(1)$. Applying Slutsky's Theorem and Lemma 3 gives the result.

Part (ii) in Theorem 1 can be proved using Cramér-Wold device. For any $\widetilde{a} \in \mathbb{R}^m$, let $\alpha_n = A_n^T\widetilde{a}$ in Theorem 1(i), which would still hold when $\|\alpha_n\|_2 \leqslant c'$ for some constant $c' > 0$. In this case, $\|\alpha_n\|_2 = \|A_n^T\widetilde{a}\|_2 \leqslant \|A_n^T\|\|\widetilde{a}\|_2 \leqslant c_*\|\widetilde{a}\|_2$ is upper bounded by a constant since $\widetilde{a}$ has a fixed dimension. Then, as $n \to \infty$,

$$\frac{n^{1/2}\widetilde{a}^T A_n(\widehat{b} - \xi^0)}{(\widetilde{a}^T A_n\widehat{\Theta}A_n^T\widetilde{a})^{1/2}} \xrightarrow{\mathcal{D}} N(0,1).$$

The variance $|\widetilde{a}^T A_n\widehat{\Theta}A_n^T\widetilde{a} - \widetilde{a}^T A_n\Theta_{\xi^0}A_n^T\widetilde{a}| \leqslant \|\widehat{\Theta} - \Theta_{\xi^0}\|\|A_n^T\widetilde{a}\|_2^2 = o_P(1)$. Hence, by Slutsky's Theorem,

$$n^{1/2}\widetilde{a}^T A_n(\widehat{b} - \xi^0) \xrightarrow{\mathcal{D}} N(0, \widetilde{a}^T F\widetilde{a}).$$

Web Appendix B. Additional Simulations

Web Appendix B.1 *Simulation studies: large n, diverging p*

We examined the scenario with smaller sample sizes, where we simulated $n = 500$ observations with $p = 20, 100, 200, 300$ covariates in logistic regression models. The rest of the settings were identical to those with $n = 1000$ in the main text. Figures S1 – S3 display the results from three types of covariance structures, including the identity matrix, the autoregressive structure of order 1 or AR(1) with correlation 0.7, and the compound symmtry structure with correlation 0.7, respectively. Figure S3 shows that with $n = 500$, $p = 300$ and the compound symmetry structure, neither of the de-biased lasso methods worked well, which is not surprising given the relatively small sample size and highly correlated covariates.

We also varied the correlation $\rho = 0.2$ in the covariance matrix Σ_x for the autoregressive and compound symmetry structures to reflect the presence of less correlated covariates; see Figure S4 and Figure S5, respectively. These results are close to the independent covariate case. To summarize, our proposed refined de-biased lasso approach, in most cases, can provide the best bias correction and honest confidence intervals.

In Section 4, additional simulation results have been shown to demonstrate that generally $\mu_n = 0$ leads to the best performance empirically in Eq. (5). In the presented logistic regression setting, we simulate $n = 500$ observations and $p = 40, 100, 200, 300, 400$ covariates for 200 times. Covariates follow a multivariate Gaussian distribution with mean zero and AR(1) covariance matrix ($\rho = 0.7$). Only two coefficients are non-zero (1 and 0.5) and the rest are noises. We pre-specify a sequence of values in $[0, 1]$ for the tuning parameter μ_n in Eq. (5), equally spaced in log scale. The de-biased lasso estimator based on Eq. (5) is referred to by "tuning".

Figure S6 (already shown in Section 4 of the main article) and Figure S7 show the simulation results for the coefficients $\xi_j^0 = 1$ and $\xi_j^0 = 0.5$ respectively, where the three columns correspond to average estimation bias, coverage probability for its 95% confidence interval, and the ratio between

its average model-based standard error and empirical standard error, over 200 replications. Since our main focus is good bias correction and honest confidence interval coverage, we find that over a very wide range of number of covariates, $\mu_n = 0$ performs the best empirically.

Web Appendix B.2 *Simulation studies: large p, small n*

We also present simulation studies that feature logistic regression models in the "large p, small n" setting, with $n = 300$ observations and $p = 500$ covariates. For simplicity, covariates are simulated from $N_p(0, \Sigma_x)$, where $\Sigma_{x,ij} = 0.7^{|i-j|}$, and truncated at ± 6. In the true coefficient vector β^0, the intercept $\beta_0^0 = 0$ and β_1^0 varies from 0 to 1.5 with 40 equally spaced increments. To examine the impacts of different true model sizes, we arbitrarily choose $\bar{s}_0$ =2, 4 or 10 additional coefficients from the rest in β^0, and fix them at 1 throughout the simulation. At each value of β_1^0, a total of 500 simulated datasets are generated. We focus on the de-biased estimates and inference for β_1^0 using the method of van de Geer et al. (2014).

Figure S8, with the true model size increasing from the top to the bottom, shows that the de-biased lasso estimate for β_1^0 has a bias which almost linearly increases with the true size of β_1^0. This undermines the credibility of the consequent confidence intervals. Meanwhile, the model-based variance overestimates the true variance for smaller signals and underestimates it for larger signals in the two models with smaller model sizes, as shown by the top two rows in Figure S8. This partially explains the over- and under-coverage for smaller and larger signals, respectively. Due to penalized estimation in node-wise lasso, the variance of the original de-biased lasso estimator is even smaller than the oracle maximum likelihood estimator obtained as if the true model were known; see the bottom two rows in Fig. S8. The empirical coverage probability decreases to about 50% as the signal β_1^0 goes to 1.5, and when the true model size reaches 5; see the middle row in Figure S8. The bias correction is sensitive to the true model size, which becomes worse for larger true models. We have also conducted simulations by changing the covariance structure of

covariates to be independent or compound symmetry with correlation coefficient 0.7 and variance 1, and have obtained similar results.

Web Appendix C. Demographics of the Boston Lung Cancer Survivor Cohort

Table S1 summarizes the demographics of the 1,374 individuals studied in the main text, stratified by their smoking status.

Web Appendix D. Discussion on the difference between sparsity assumptions in our work and van de Geer et al. (2014)

We first notice there are two kinds of sparsity parameters: one for the sparsity of regression coefficients (denoted by s_0), and the other for the sparsity of the inverse of the information matrix, Θ_{ξ^0} (denoted by s_j, the number of non-zero elements in the jth row of Θ_{ξ^0}). The following clarifies the extent to which the assumptions of our Theorem 1 differs from those of van de Geer et al. (2014). For the model sparsity s_0, our assumption $s_0 \log(p)(p/n)^{1/2} \to 0$ is indeed more stringent than $s_0 \log(p)/\sqrt{n} \to 0$ required by van de Geer et al. (2014), whereas for the sparsity of the inverse information matrix, van de Geer et al. (2014) assumed $s_j = o(\sqrt{n/\log(p)})$ for all j and we do not make any assumptions on s_j directly. A related condition set by us is $p^2/n \to 0$, which is weaker than $s_j = o(\sqrt{n/\log(p)})$ by a logarithmic factor if $s_j \asymp p$.

Below we elaborate on how these sparsity differences lead to different results obtained by our manuscript and van de Geer et al. (2014), which indeed have different inferential objectives, and the rationale why our assumptions fit our inferential objectives.

First, these two works differ in inferential objectives. We aim to infer any linear combinations of the regression parameter, i.e. $\alpha_n^T \xi^0$, where the only constraint on α_n is $\|\alpha_n\|_2 = 1$ (in fact, bounded $\|\alpha_n\|_2$ would suffice). Thus, we have to control the behavior of the $(p+1) \times (p+1)$ matrix $\{\widehat{\Theta} - \Theta_{\xi^0}\}$. In contrast, van de Geer et al. (2014) inferred individual components in ξ^0 one

at a time, making it sufficient to control the rates of $\{\widehat{\Theta}_j - \Theta_{\xi^0,j}\}$ (here the subscript j indicates the jth row of a matrix) for one row at a time, and the node-wise lasso provides such required rates.

Second, besides the essential assumptions that both papers require (our Assumptions 1–4), van de Geer et al. (2014) has another important assumption that we do not need to assume, that is, $\|\mathbf{X}_{\beta^0,-j}\gamma^0_{\beta^0,j}\|_\infty = \mathcal{O}(1)$ (see their Theorem 3.3 (iv)), which results in $\|\mathbf{X}\widehat{\Theta}_j^T\|_\infty = \mathcal{O}_P(K)$ in their condition (C5). In our notation, this assumption would be equivalent to the boundedness on $\|\Theta_{\xi^0}x_i\|_\infty$ and would result in $\|\widehat{\Theta}x_i\|_\infty$ being bounded in probability. However, we have elected not to directly make such assumptions on the inverse of the informative matrix and its estimate as they may be closely related to the sparsity requirement of Θ_{ξ^0} under Assumption 1, and may not hold or be verifiable in GLM settings.

Finally, we clarify that our specified order assumption on s_0 with respect to n and p is to ensure $|n^{1/2}\alpha_n^T\widehat{\Theta}\Delta| = o_P(1)$ in the proof of Theorem 1 (please see Pages 5–6 in Web Appendix A), which is for inference on any linear combinations of regression coefficients. However, if we had aimed for a weaker result of inferring an individual coefficient only as in van de Geer et al. (2014), we would have let $\alpha_n = e_j$ (a p-dimensional vector with the jth element being 1 and all the other elements being zero) corresponding to drawing inference on the effect of the jth covariate, and also with a condition of $\|\widehat{\Theta}x_i\|_\infty = \mathcal{O}_P(1)$ as in van de Geer et al. (2014), we would have had

$$\begin{aligned}
|\sqrt{n}\widehat{\Theta}_j\Delta| &= |\sqrt{n}\frac{1}{n}\sum_{i=1}^n \left\{\ddot{\rho}(y_i, a_i^*) - \ddot{\rho}(y_i, x_i^T\widehat{\xi})\right\}\widehat{\Theta}_j x_i x_i^T(\xi^0 - \widehat{\xi})| \\
&\leqslant \sqrt{n}\frac{1}{n}\sum_{i=1}^n |\ddot{\rho}(y_i, a_i^*) - \ddot{\rho}(y_i, x_i^T\widehat{\xi})| \cdot |\widehat{\Theta}_j x_i| \cdot |x_i^T(\xi^0 - \widehat{\xi})| \\
&\leqslant \sqrt{n}\frac{1}{n}\sum_{i=1}^n c_{Lip}|x_i^T(\xi^0 - \widehat{\xi})| \cdot \mathcal{O}_P(1) \cdot |x_i^T(\xi^0 - \widehat{\xi})| \\
&= \sqrt{n}\mathcal{O}_P(1)\frac{1}{n}\sum_{i=1}^n |x_i^T(\xi^0 - \widehat{\xi})|^2 \\
&= \mathcal{O}_P(\sqrt{n}s_0\lambda^2).
\end{aligned}$$

Therefore, to infer ξ_j^0 alone, we would have reached the same assumption that $s_0\log(p)/\sqrt{n} \to 0$ as in van de Geer et al. (2014) with $\lambda \asymp \sqrt{\log(p)/n}$.

In summary, our work may be meritorious by providing readers with these explicit rates for guaranteeing proper inferences when directly inverting the information matrix.

REFERENCES

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

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45,** 158–195.

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

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge: Cambridge University Press.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Table 1 about here.]

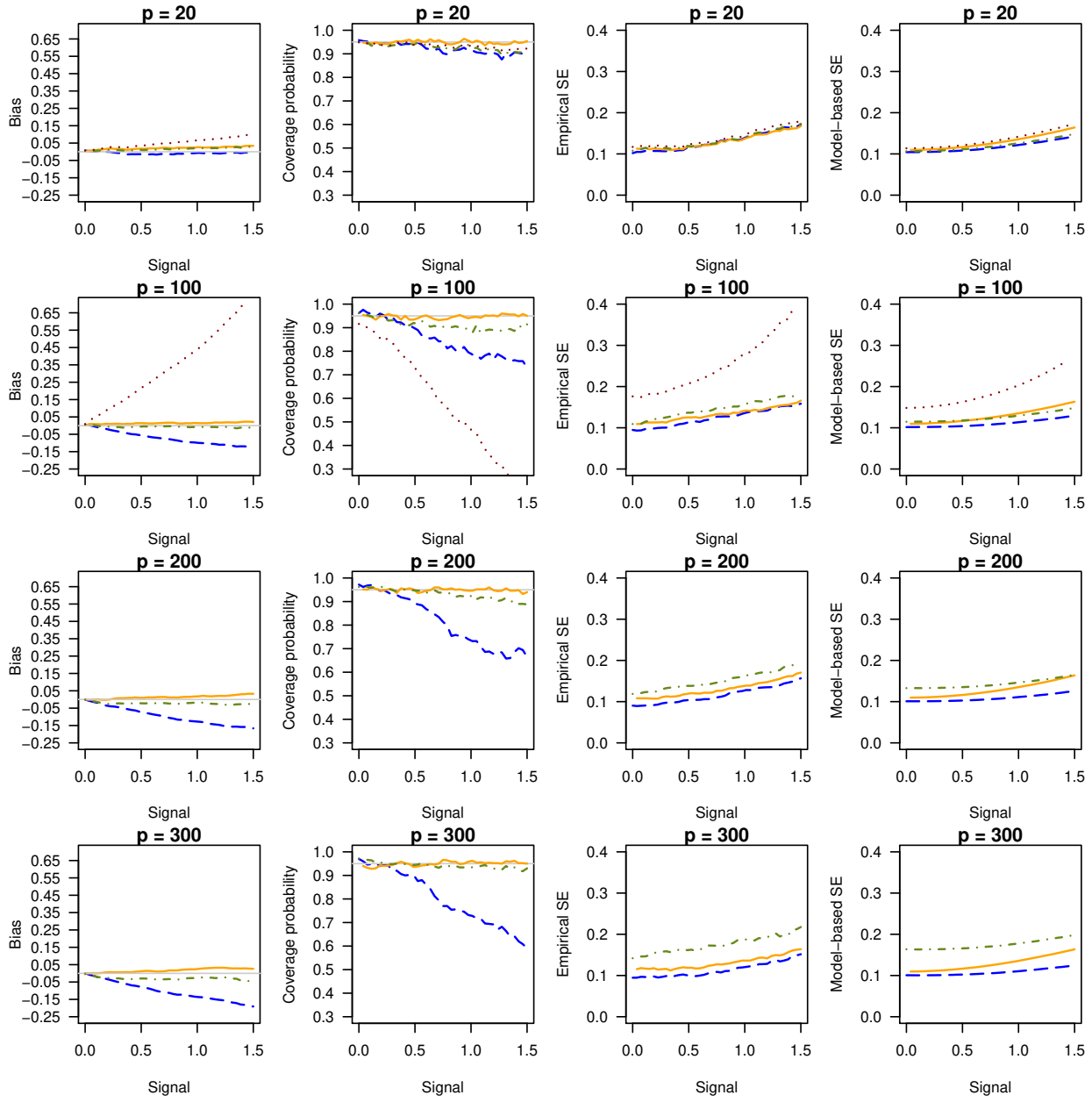

Figure S1. Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(0_p, I)$ before being truncated at ± 6. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

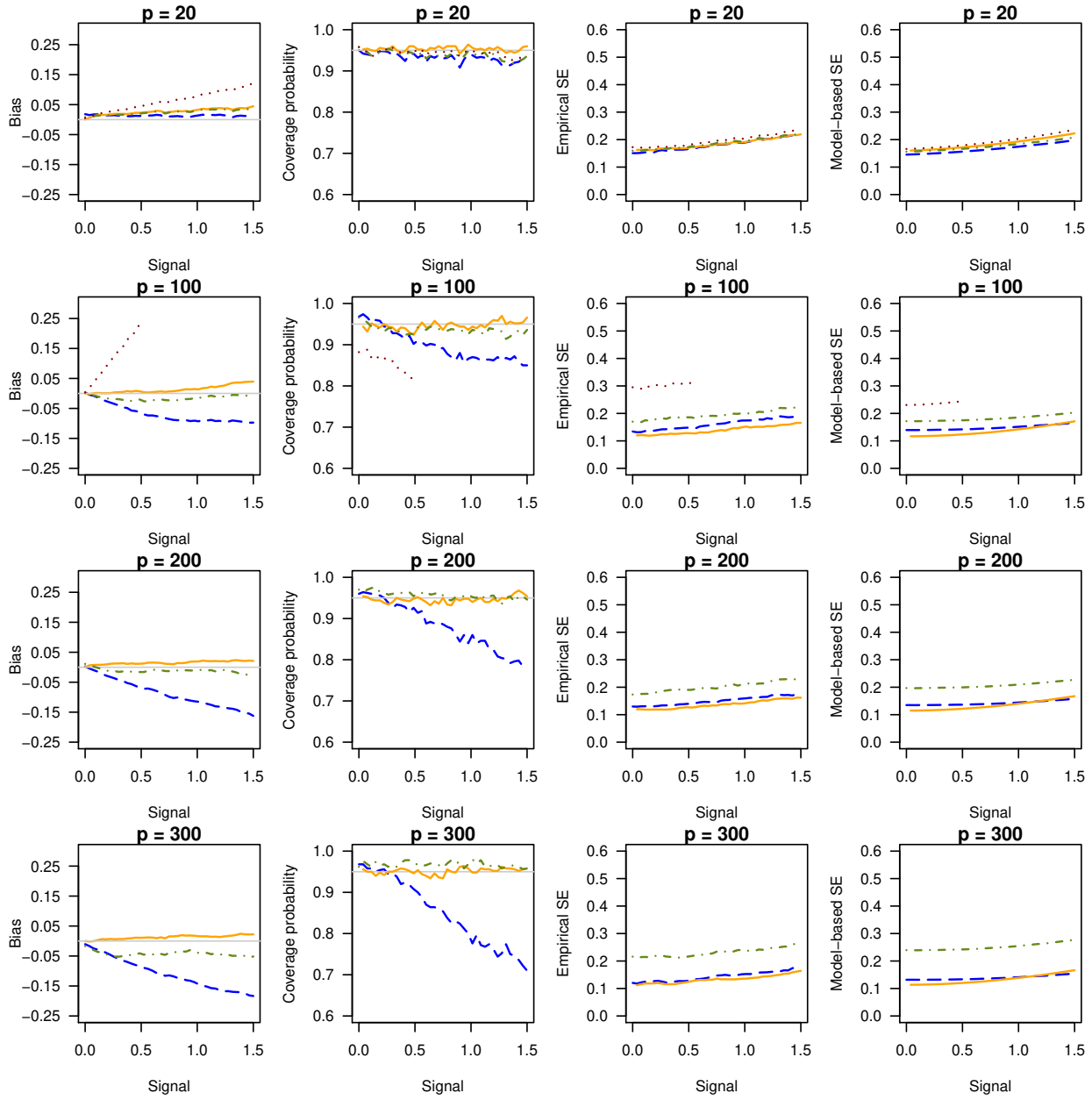

Figure S2. Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(0_p, \Sigma_x)$ before being truncated at ± 6, where Σ_x has an autoregressive covariance structure of order 1 with $\rho = 0.7$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

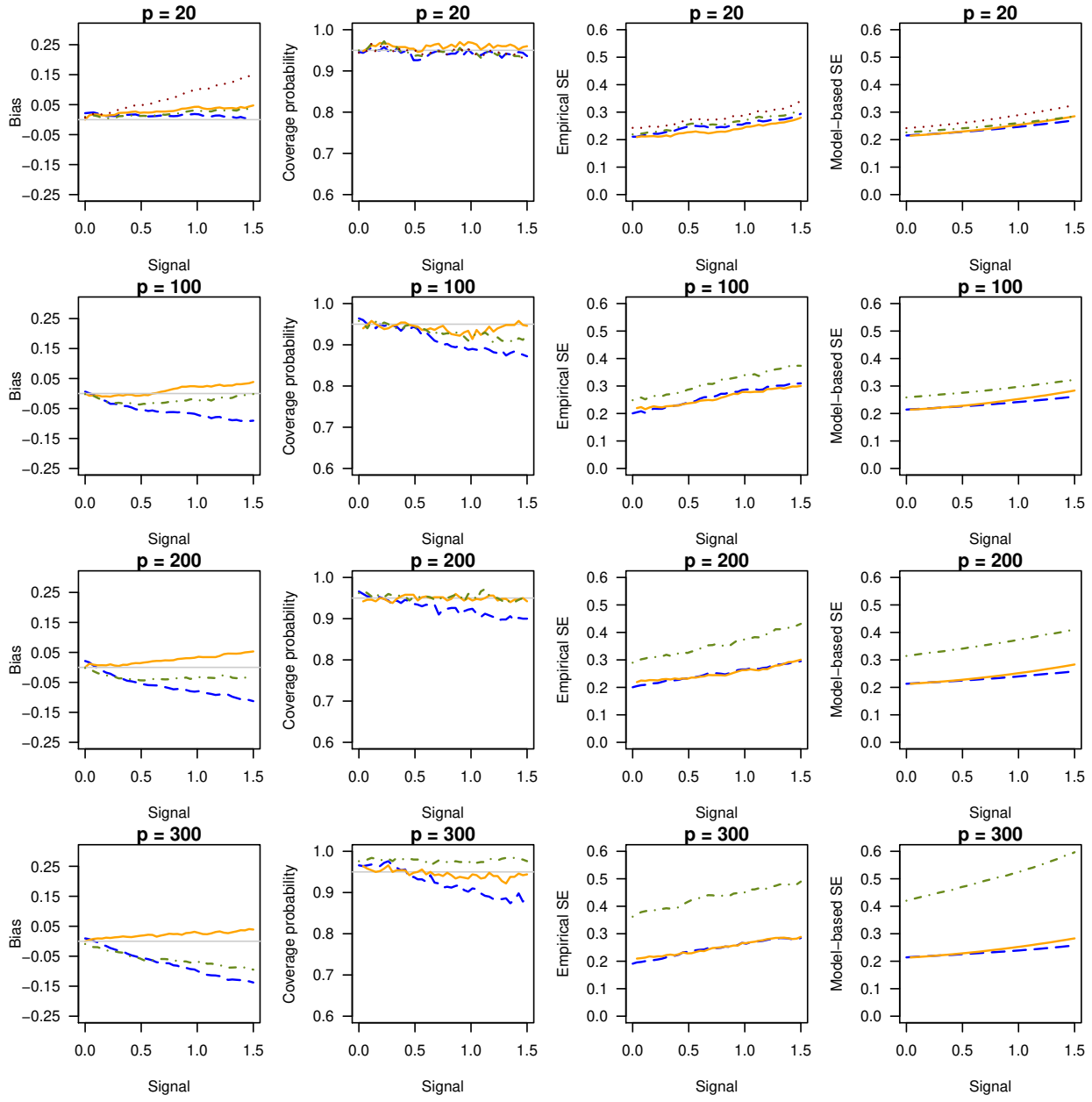

Figure S3. Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(0_p, \Sigma_x)$ before being truncated at ± 6, where Σ_x has a compound symmetry structure with $\rho = 0.7$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

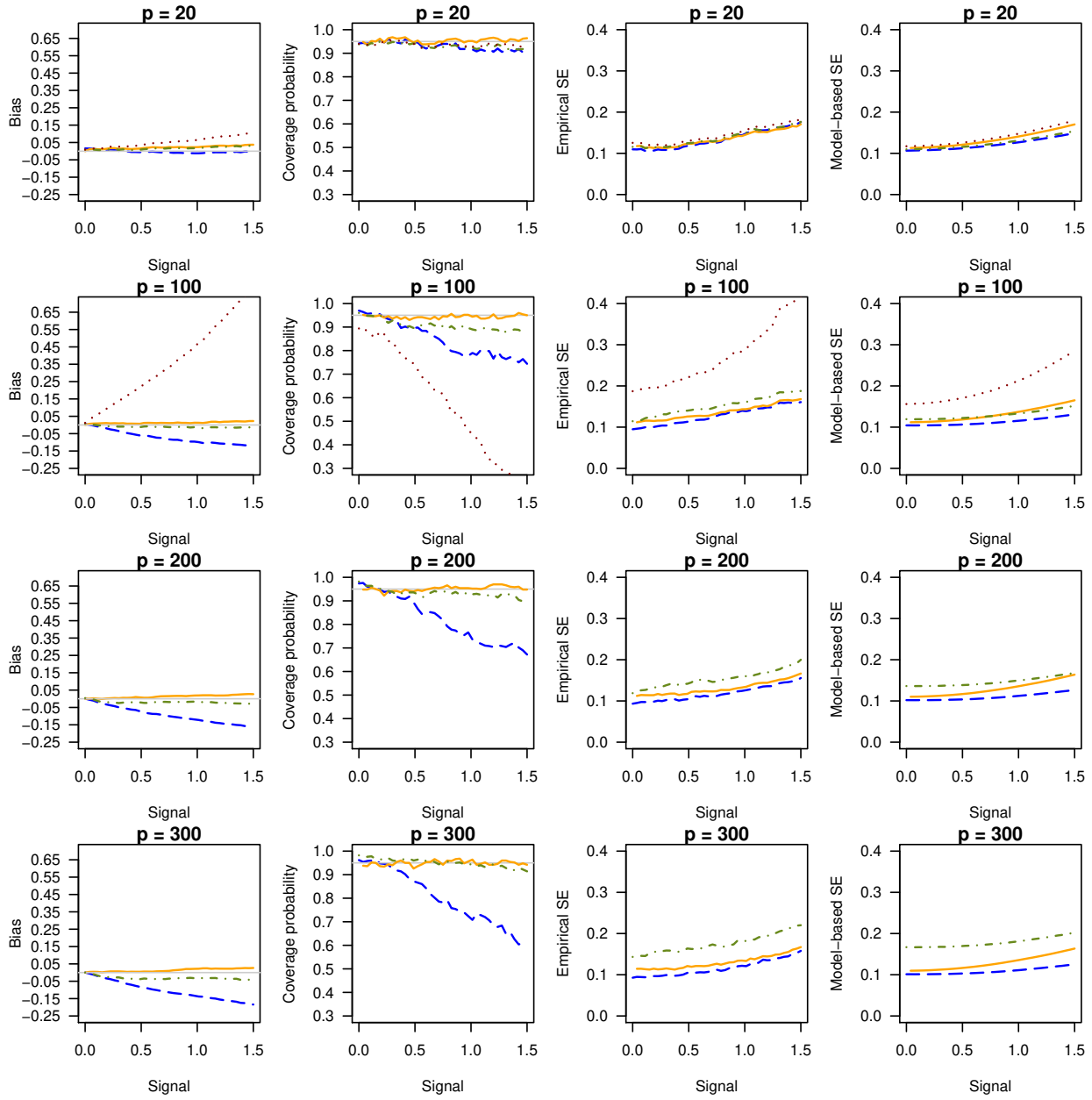

Figure S4. Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(0_p, \Sigma_x)$ before being truncated at ± 6, where Σ_x has an autoregressive covariance structure of order 1 with $\rho = 0.2$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

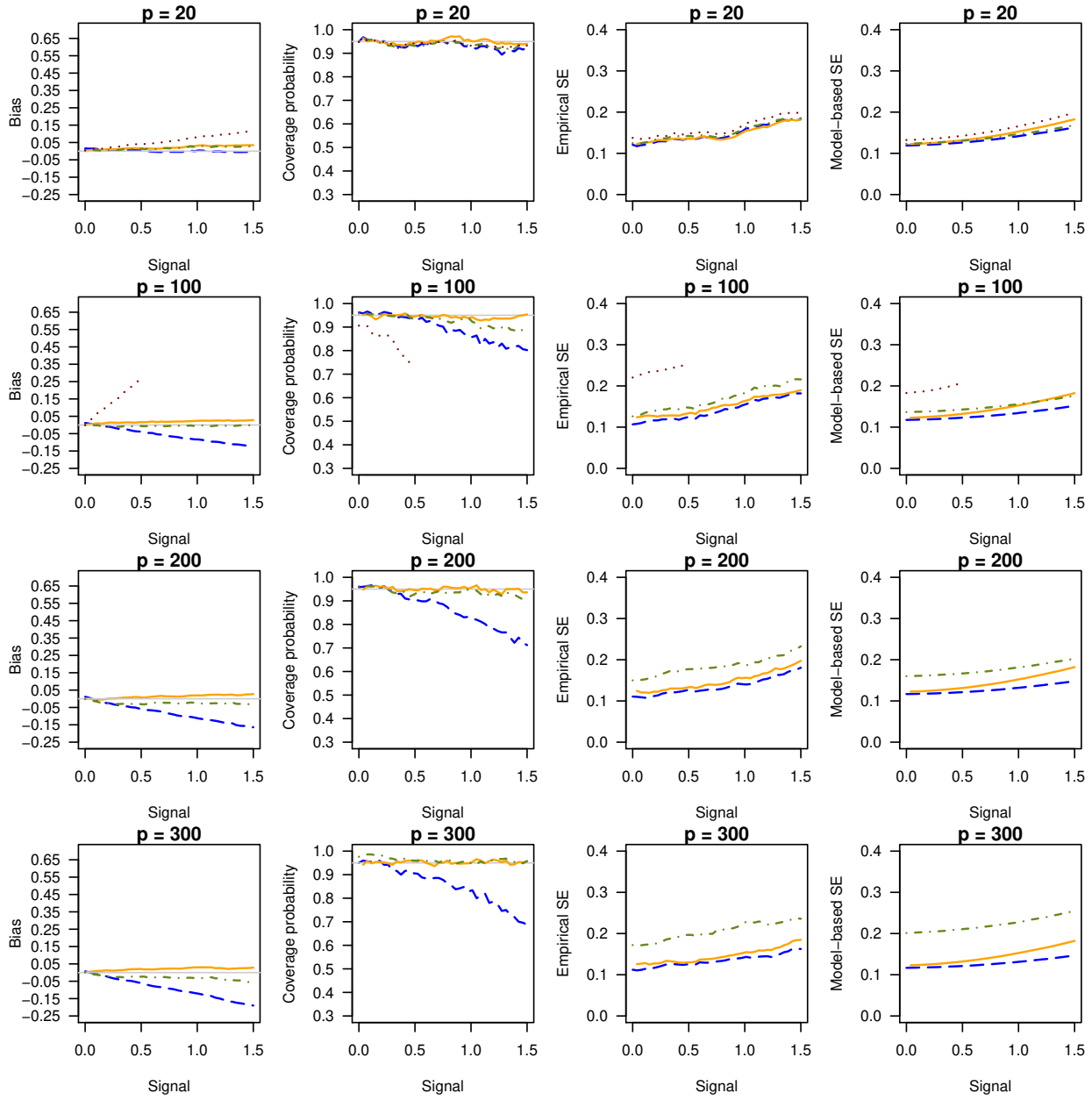

Figure S5. Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(0_p, \Sigma_x)$ before being truncated at ± 6, where Σ_x has a compound symmetry structure with $\rho = 0.2$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

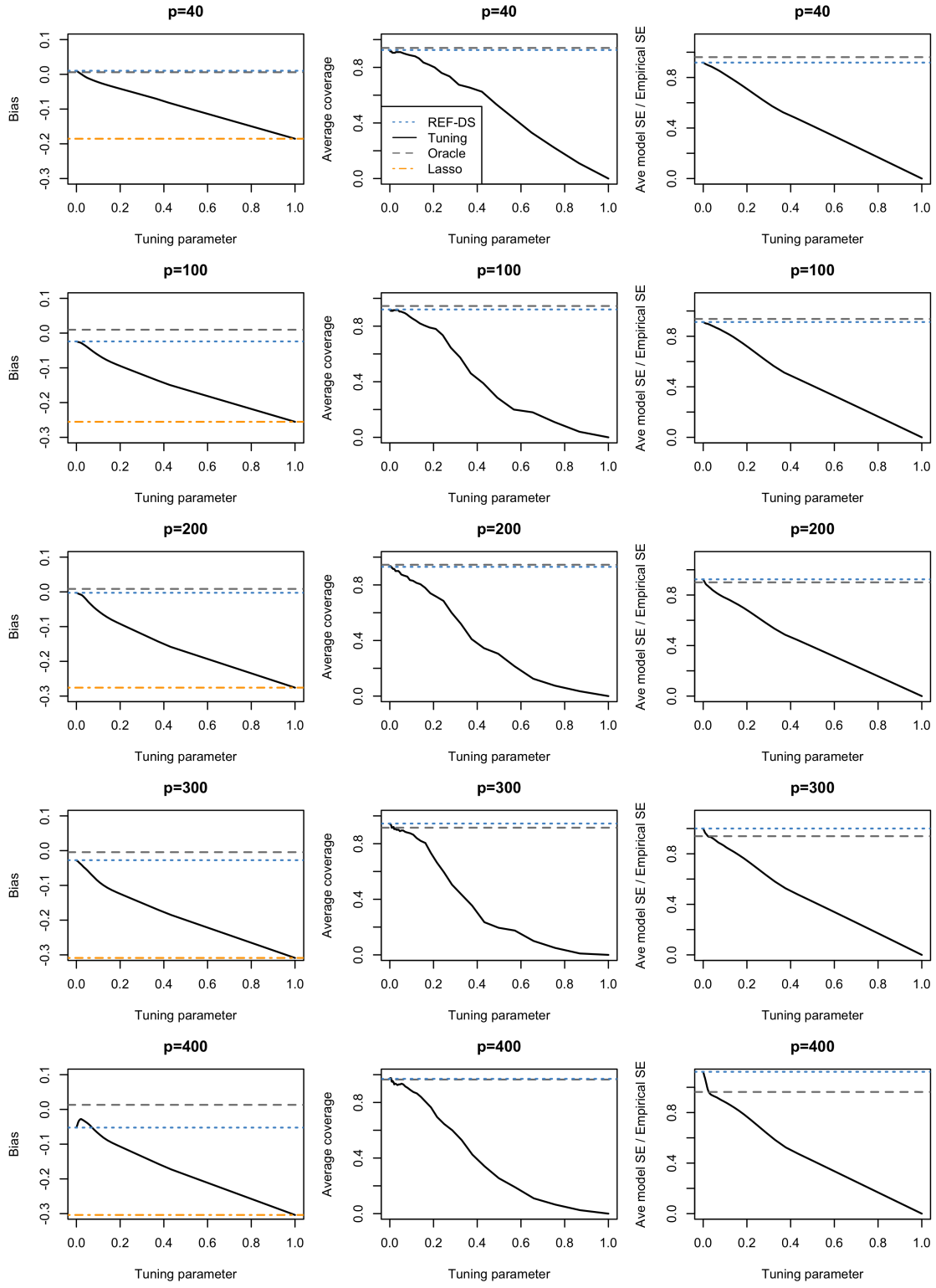

Figure S6. Simulation results that verify the selection of the tuning parameter $\mu_n = 0$ in Eq. (5) for $\xi_j^0 = 1$.

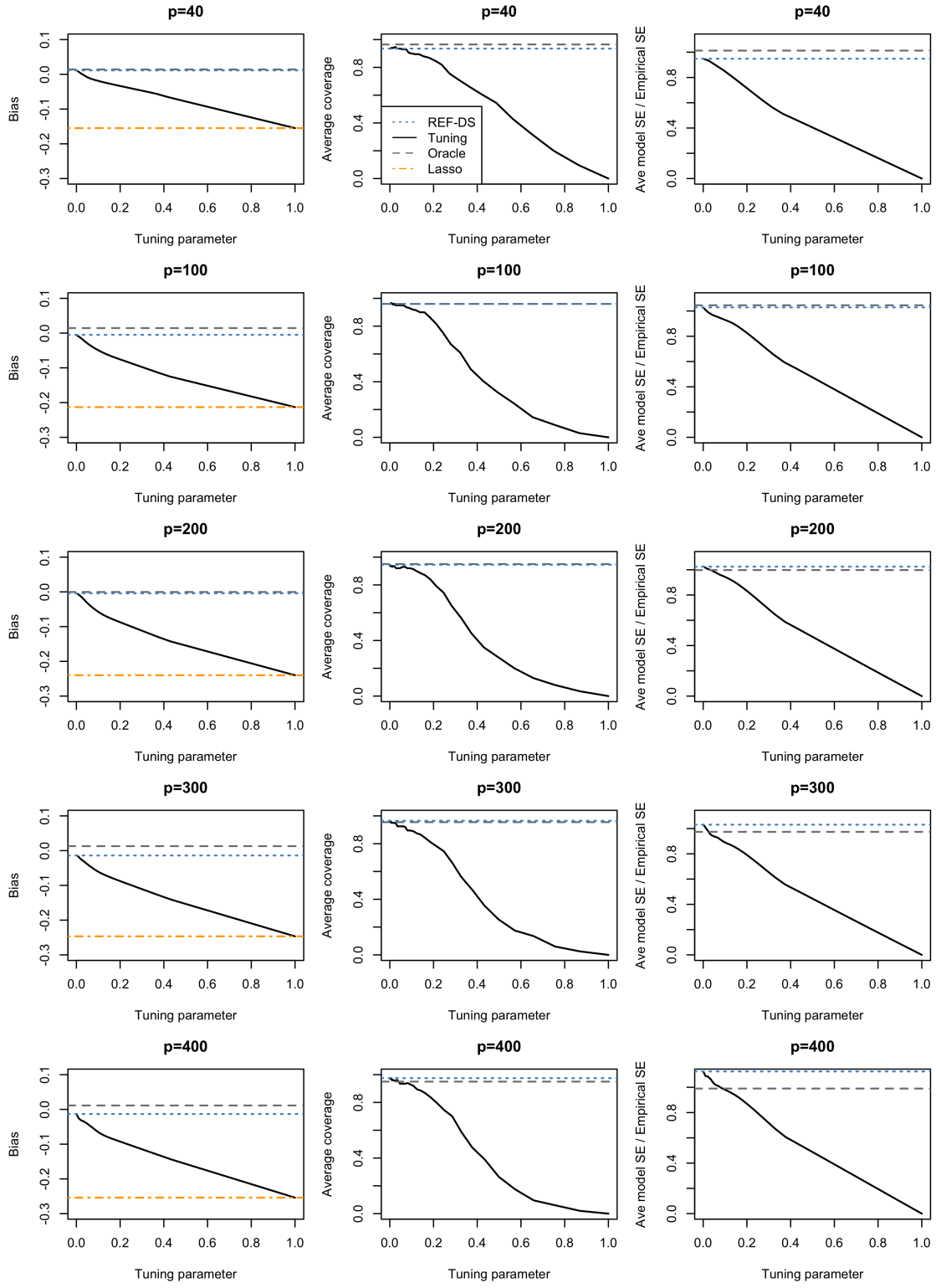

Figure S7. Simulation results that verify the selection of the tuning parameter $\mu_n = 0$ in Eq. (5) for $\xi_j^0 = 0.5$.

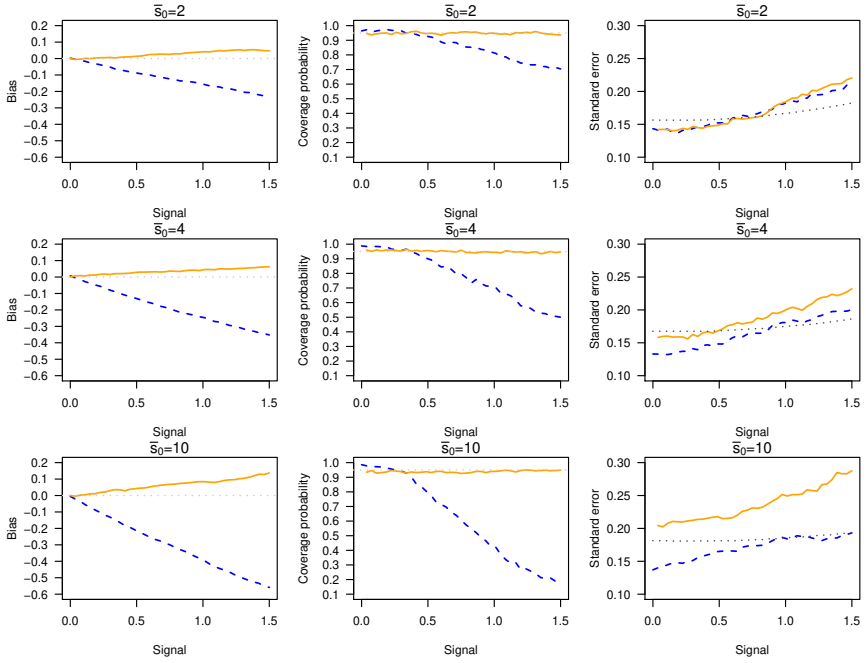

Figure S8. Simulation results of a logistic regression with sample size $n = 300$ and $p = 500$ covariates. Covariates are simulated from $N_p(0_p, \Sigma_x)$ before being truncated at ± 6, where Σ_x has an autoregressive covariance structure of order 1 with $\rho = 0.7$. The left column presents estimation bias, the middle column presents empirical coverage probability, and the right column presents standard error, both model-based and empirical, of the estimated β_1^0. Horizontal panels correspond to models with 2, 4 and 10 additional signals fixed at 1 from the top to the bottom, respectively. In the left and middle columns, blue dashed lines represent the original de-biased lasso approach by van de Geer et al. (2014), and orange solid lines represent the oracle estimator. In the right column, blue dashed lines and black dotted lines represent the empirical standard error and the model-based standard error from the method of van de Geer et al. (2014), respectively, and orange solid lines for the empirical standard error of the oracle estimator.

Table S1
Characteristics of the individuals in the analytical data set of the Boston Lung Cancer Survivor Cohort

Information	Overall Count (%) / Mean (SD[1])	Among smokers Count (%) / Mean (SD)	Among non-smokers Count (%) / Mean (SD)
Total	1374 (100%)	1077 (100%)	297 (100%)
Lung cancer			
Yes	651 (47.4%)	595 (55.2%)	56 (18.9%)
No	723 (52.6%)	482 (44.8%)	241 (81.1%)
Education			
No high school	153 (11.1%)	139 (12.9%)	14 (4.7%)
High school graduate	374 (27.2%)	309 (28.7%)	65 (21.9%)
At least 1-2 years of college	847 (61.7%)	629 (58.4%)	218 (73.4%)
Gender			
Female	845 (61.5%)	644 (59.8%)	201 (67.7%)
Male	529 (38.5%)	433 (40.2%)	96 (32.3%)
Age	60.0 (10.6)	60.7 (10.2)	57.7 (11.7)