ORIGINAL ARTICLE

Statistical inference for Cox proportional hazards models with a diverging number of covariates

Lu Xia[1] │ Bin Nan[2] │ Yi Li[3]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA

[2]Department of Statistics, University of California, Irvine, Irvine, California, USA

[3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Yi Li, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA.
Email: yili@umich.edu
Bin Nan, Department of Statistics, University of California, Irvine, Donald Bren Hall, Irvine, CA 92697, USA.
Email: nanb@uci.edu

Abstract

For statistical inference on regression models with a diverging number of covariates, the existing literature typically makes sparsity assumptions on the inverse of the Fisher information matrix. Such assumptions, however, are often violated under Cox proportion hazards models, leading to biased estimates with under-coverage confidence intervals. We propose a modified debiased lasso method, which solves a series of quadratic programming problems to approximate the inverse information matrix without posing sparse matrix assumptions. We establish asymptotic results for the estimated regression coefficients when the dimension of covariates diverges with the sample size. As demonstrated by extensive simulations, our proposed method provides consistent estimates and confidence intervals with nominal coverage probabilities. The utility of the method is further demonstrated by assessing the effects of genetic markers on patients' overall survival with the Boston Lung Cancer Survival Cohort, a large-scale epidemiology study investigating mechanisms underlying the lung cancer.

KEYWORDS

cancer epidemiology, debiased lasso, lung cancer, precision matrix, quadratic programming, sparsity

1 | INTRODUCTION

The Cox proportional hazards model (Cox, 1972), a semi-parametric model with an unspecified baseline hazard function, has been widely used for the analysis of censored time-to-event data. With a fixed dimension of covariates, Cox (1972) proposed the maximum partial likelihood estimation (MPLE) to infer the regression coefficients, and Andersen and Gill (1982) proved the asymptotic distributional results for MPLE using the martingale theory.

Technological advances nowadays have made it possible to collect a large amount of information in biomedical studies. For example, the Boston Lung Cancer Survival Cohort (BLCSC), the motivating study for this work, has acquired abundant clinical, genetic, epigenetic, and genomic data, which enable comprehensive investigations of molecular mechanisms underlying the lung cancer survival (McKay et al., 2017). High-dimensionality of the collected covariates has confronted the traditional parameter estimation and uncertainty quantification based on Cox models. In high-dimensional settings, where the number of covariates p increases with the sample size n or even greater than n, the conventional MPLE is usually ill-conditioned. Penalized estimators have emerged as a powerful tool for simultaneous variable selection and estimation (Antoniadis et al., 2010; Fan and Li, 2002; Gui and Li, 2005; Tibshirani, 1997). Recently, Huang et al. (2013) and Kong and Nan (2014) derived the nonasymptotic oracle inequalities of the lasso estimator in the Cox model. However, none of these works dealt with statistical inference for Cox models with high-dimensional covariates.

Existing literature on inference for high-dimensional models mainly concerns linear regression. Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014) developed inference procedures for linear models, based on debiasing the lasso estimator via low-dimensional projection or inverting the Karush–Kuhn–Tucker condition. van de Geer et al. (2014) extended the debiased lasso idea to generalized linear models, using the nodewise lasso regression. Ning and Liu (2017) focused on hypothesis testing and devised decorrelated score, Wald and likelihood ratio tests for inference on a low-dimensional parameter in generalized linear models based on projection theory.

There has been some existing progress in inference for the Cox model with high-dimensional covariates. Fang et al. (2017) developed decorrelated tests for hypothesis testing of low-dimensional components under high-dimensional Cox models, using ideas similar to Ning and Liu (2017). Kong et al. (2021) extended the debiased lasso approach in van de Geer et al. (2014) to potentially misspecified Cox models, and used the nodewise lasso regression to estimate the inverse information matrix. Yu et al. (2021) proposed a debiased lasso approach, by estimating the inverse information matrix with a CLIME estimator adapted from Cai et al. (2011). Most of these works restricted the number of nonzero elements of each row in the inverse information matrix to be small, that is, ℓ_0 sparsity. However, as found in Xia et al. (2021), the sparse inverse information matrix assumption has no practical interpretation beyond linear regression models, often fails to hold in the Cox model, and these methods cannot perform satisfactorily in high-dimensional Cox model settings. For example, as evidenced by our extensive simulations, these methods cannot correct biases of lasso estimators or construct confidence intervals with desired coverage probabilities, even when the number of regression coefficients is moderate relative to the sample size. Recently, Zhang et al. (2022) proposed a projection-based cross-validation approach, which split the data into two halves and used one half for model selection and the other half for inference by constructing projected partial score functions. This approach relied on model selection consistency, for which Zhang et al. (2022) employed weighted lasso.

Our work is pertaining to the "large n, diverging p" framework where $p < n$ and p is allowed to increase with n to infinity, which reflects the setting of the motivating BLCSC with $n = 561$ and $p = 231$. Under this framework, we draw inference based on Cox models without imposing sparsity to the inverse information matrix. Specifically, we propose a debiased lasso approach via solving a series of quadratic programming problems to estimate the inverse information matrix. We use quadratic programming as a means of balancing the bias-variance trade-off and avoiding the unrealistic ℓ_0 sparsity assumption for the large inverse information matrix in the Cox model. Unlike Zhang et al. (2022), we use lasso estimates as initial values rather than for variable selection, and our proposed inference is not conditional on the selected model. Our work adds to the literature in the following aspects. First, unlike Javanmard and Montanari (2014), our work entails careful treatment of the sum of non-independently nor identically distributed terms in the empirical loss function, and we consider random designs instead of deterministic designs. Second, we find that the tuning parameter selection for the inverse information matrix estimation is crucial for bias correction. For example, a related work (Yu et al., 2021) proposed to select tuning parameters by minimizing the cross-validated difference between the product of the information matrix with its estimated inverse and the identity matrix, but was found to perform poorly. In contrast, we propose a cross-validation procedure to tune parameters by hard thresholding debiased estimates when solving the quadratic programming problems, which yields satisfactory numerical performance.

The article is organized as follows. Section 2 introduces the proposed debiased lasso approach, where the inverse information matrix is estimated via quadratic programming with a novel cross-validation procedure for selecting the tuning parameter. Section 3 lays the theoretical foundation for reliable inference on linear combinations of the Cox regression parameters using debiased lasso estimators. We examine the finite sample performance of our proposed method with simulation studies in Section 4, apply it to analyze the BLCSC data in Section 5, and conclude the paper with a few remarks in Section 6. We state several useful technical lemmas and provide proofs of the main theorems in the Appendix, and defer proofs of all the lemmas to Appendix S1.

2 | METHOD

2.1 | Background and setup

We introduce notation that will be used throughout this article. For a vector $x = (x_1, \ldots, x_r)^T \in \mathbb{R}^r$, $x^{\otimes 0} = 1$, $x^{\otimes 1} = x$ and $x^{\otimes 2} = xx^T$. The ℓ_q-norm for x is $\|x\|_q = (\sum_{j=1}^r |x_j|^q)^{1/q}$, $q \geq 1$, and $\|x\|_0 = \sum_{j=1}^r I(x_j \neq 0)$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m \times r}$, the induced matrix norm is defined as $\|A\|_{q_1,q_2} = \sup_{x \in \mathbb{R}^r, x \neq 0} \|Ax\|_{q_2} / \|x\|_{q_1}$, $q_1, q_2 \geq 1$. In particular, $\|A\|_{1,1} = \max_{1 \leq j \leq r} \sum_{i=1}^m |a_{ij}|$, $\|A\|_{2,2} = \sigma_{\max}(A)$, the largest singular value of A, and $\|A\|_{\infty,\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^r |a_{ij}|$. The element-wise max norm is denoted as $\|A\|_\infty = \max_{i,j} |a_{ij}|$. For two positive sequences $\{d_n\}$ and $\{g_n\}$, we define $d_n \asymp g_n$ if there are two bounded positive constants C and C' such that $C \leq d_n/g_n \leq C'$.

A Cox model stipulates that the hazard function for the underlying failure time T, conditional on a p-dimensional vector of covariates $X = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$, is $h(t|X) = h_0(t) \exp\{X^T \beta^0\}$, where $h_0(t)$ is an unknown baseline hazard function and $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)^T \in \mathbb{R}^p$ is an unknown vector of regression coefficients. With T subject to right censoring, the observed survival time is $Y = \min(T, C)$, where the censoring time C is assumed to be independent of T given X. Let $\delta = 1(T \leq C)$ denote the event indicator. Based on n independent and identically distributed

observations $\{Y_i, X_i, \delta_i\}_{i=1}^n$, the goal of the paper is to estimate and draw inference on the regression coefficients β^0, when $p < n$ but $p \to \infty$ as $n \to \infty$.

2.2 | Debiasing the lasso estimator

When p is fixed, a natural approach for inferring β^0 is through MPLE, which maximizes the log partial likelihood function

$$\frac{1}{n} \sum_{i=1}^n \left[X_i^T \beta - \log \left\{ \frac{1}{n} \sum_{j=1}^n 1(Y_j \geq Y_i) \exp(X_j^T \beta) \right\} \right] \delta_i, \tag{1}$$

However, with a diverging p of our interest, MPLE may suffer from numerical instability and yield unreliable inference; see Section 4.

A more commonly used approach, when p diverges to ∞ as $n \to \infty$, is a lasso estimator, defined to be the minimizer of the following penalized negative log partial likelihood:

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \ell_n(\beta) + \lambda_n \|\beta\|_1 \right\},$$

where $\ell_n(\beta)$ is the negative log partial likelihood function, that is, the negative of (1), and $\lambda_n > 0$ is a tuning parameter to be decided. The first- and second-order derivatives of $\ell_n(\beta)$ with respect to β, that is, the score function and the information matrix, are respectively denoted by

$$\dot{\ell}_n(\beta) = -\frac{1}{n} \sum_{i=1}^n \left\{ X_i - \frac{\widehat{\mu}_1(Y_i; \beta)}{\widehat{\mu}_0(Y_i; \beta)} \right\} \delta_i, \quad \ddot{\ell}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\widehat{\mu}_2(Y_i; \beta)}{\widehat{\mu}_0(Y_i; \beta)} - \left[\frac{\widehat{\mu}_1(Y_i; \beta)}{\widehat{\mu}_0(Y_i; \beta)} \right]^{\otimes 2} \right\} \delta_i,$$

where $\widehat{\mu}_r(t; \beta) = n^{-1} \sum_{j=1}^n 1(Y_j \geq t) X_j^{\otimes r} \exp\{X_j^T \beta\}$, $r = 0, 1, 2$. We also define the weighted average covariate vector $\widehat{\eta}_n(t; \beta) = \widehat{\mu}_1(t; \beta) / \widehat{\mu}_0(t; \beta) = \sum_{j=1}^n 1(Y_j \geq t) \exp\{X_j^T \beta\} X_j \Big/ \sum_{j=1}^n 1(Y_j \geq t) \exp\{X_j^T \beta\}$.

The lasso estimates tend to be more stable because of the penalization. However, as the lasso estimator $\widehat{\beta}$ incurs biases (Javanmard and Montanari, 2014), we consider a debiased lasso approach to remove its bias and draw inference. Analogous to van de Geer et al. (2014) for generalized linear models, we define a debiased lasso estimator for β^0 as

$$\widehat{b} = \left(\widehat{b}_1, \ldots, \widehat{b}_p \right)^T = \widehat{\beta} - \widehat{\Theta} \dot{\ell}_n(\widehat{\beta}), \tag{2}$$

with $-\widehat{\Theta} \dot{\ell}_n(\widehat{\beta})$ serving as the bias correction term, where $\widehat{\Theta}$ is an estimate of the inverse information matrix. A reliable estimator, $\widehat{\Theta}$, is important to ensure the validity of the method. However, existing methods, most of which pose ℓ_0 sparsity assumptions on the true inverse information matrix and estimate sparse matrices $\widehat{\Theta}$ using, for instance, nodewise lasso (Kong et al., 2021) or CLIME (Yu et al., 2021), perform poorly for Cox models in numerical studies. Inspired by the work of Javanmard and Montanari (2014), we propose to estimate each row of $\widehat{\Theta}$ by solving the following optimization problem for m:

$$\min\{m^T \widehat{\Sigma} m : m \in \mathbb{R}^p, \|\widehat{\Sigma} m - e_j\|_\infty \leq \gamma_n\}, j = 1, \ldots, p, \tag{3}$$

where $\gamma_n \geq 0$ is a tuning parameter, e_j is the vector with one at the jth element and zero elsewhere, and the $p \times p$ matrix

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \delta_i \left\{ X_i - \hat{\eta}_n(Y_i; \hat{\beta}) \right\}^{\otimes 2}. \tag{4}$$

We obtain $\hat{\Theta}$ as a $p \times p$ matrix consisting of all p solutions to (3) as its corresponding row vectors. Of note, we use $\hat{\Sigma}$ in (3) in lieu of $\ddot{\ell}_n(\hat{\beta})$, which is for theoretical convenience that becomes evident in Section 3. In fact, under the assumptions in Section 3, we do have $\|\hat{\Sigma} - \ddot{\ell}_n(\hat{\beta})\|_{\infty} = o_P(1)$ with a desirable convergence rate (see the proof of Theorem 1 in the Appendix), and the numerical difference in the resulting debiased lasso estimators is negligible.

The objective $m^T \hat{\Sigma} m$ in (3) is quadratic in m, and the constraint $\|\hat{\Sigma} m - e_j\|_{\infty} \leq \gamma_n$ is equivalent to $\hat{\Sigma} m - e_j \leq \gamma_n 1_p$ and $-\hat{\Sigma} m + e_j \leq \gamma_n 1_p$, which are linear in m. Hence, (3) falls into a standard quadratic programming problem (Boyd & Vandenberghe, 2004). With a smooth shape of $m^T \hat{\Sigma} m$, (3) is unlikely to produce sparse solutions, as opposed to lasso or Dantzig selectors. As in Javanmard and Montanari (2014), we do not impose any sparsity conditions on the true inverse information matrix. Computationally, the proposed (3) can be implemented fairly fast for moderate dimensions and parallelized for high dimensions by using the R function `solve.QP`. Our simulations demonstrate its computational efficiency.

Our approach extends Javanmard and Montanari (2014) in a linear regression setting to survival models. However, as $\hat{\eta}_n(Y_i; \hat{\beta})$ involves all subjects, $\hat{\Sigma}$ given in (4) is no longer a sum of independent and identically distributed terms, posing additional theoretical difficulties. We have addressed these challenges in our proofs.

2.3 | Selection of the tuning parameter

Selecting a proper tuning parameter γ_n is critical for bias correction in $\hat{b}$, which can be illustrated by a simulation study. We simulate $n = 500$ independent subjects, each with $p = 100$ independent covariates generated from $N(0, 1)$. Only two coefficients in β^0 in the Cox model are nonzero, taking values of 1 and 0.3. The underlying survival time Y follows an exponential distribution with a rate of $\exp(X^T \beta^0)$, and the censoring time is simulated from an exponential distribution with a rate of $0.2 \exp(X^T \beta^0)$, resulting in a censoring rate of about 20%. Figure 1 depicts how the estimation bias and the empirical coverage probability from the debiased lasso approach change as γ_n ranges from 0 to 1, revealing that γ_n within the shaded range would yield desirable inference results.

We have found that, when evaluating cross-validation criteria for choosing γ_n, directly plugging in debiased estimates produces highly unstable values because of accumulative errors from inclusion of the estimates for a large number of noise covariates. Instead, we propose a cross-validation procedure by hard-thresholding debiased estimates: splitting data randomly into K folds ($K = 5$ or 10), we use the kth fold to obtain a debiased lasso estimate $\hat{b}^{(k)}$, hard-threshold it and plug in the thresholded values for computing cross-validation criteria. Hard-thresholding is based on multiple testing with, for example, the Bonferroni correction. That is, we take the hard-thresholded values to be $\hat{b}_j^{(k),HT} = \hat{b}_j^{(k)}$ if $\sqrt{n}|\hat{b}_j^{(k)}|/\hat{\Theta}_{jj} > z_{\alpha/(2p)}$, or 0 otherwise, where $z_{\alpha/(2p)}$ is the upper $(\alpha/(2p))$th percentile of $N(0, 1)$, as determined by the asymptotic result given in Theorem 1. Then, letting $\ell^{(k)}$ be the negative log partial likelihood (defined as in (1) but applied
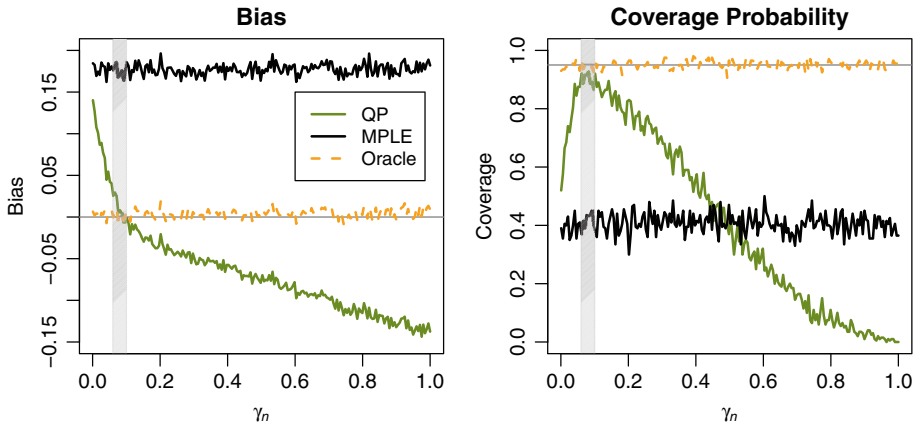
FIGURE 1 Estimation bias and 95% confidence interval coverage probability for $\beta_1^0 = 1$ with the tuning parameter $\gamma_n \in [0, 1]$ in a simulated example with $n = 500$ observations and $p = 100$ independent covariates, based on 200 simulations. The methods in comparison include the proposed debiased lasso with quadratic programming (QP), the maximum partial likelihood estimation (MPLE) and the oracle estimator (Oracle) obtained from fitting the true model.

to the kth testing set) evaluated at $\widehat{b}^{(k),HT}$, we choose γ_n that gives the smallest cross-validated negative partial likelihood, $\sum_{k=1}^{K} n^{(k)} \ell^{(k)}$, where $n^{(k)}$ is the number of observations in the kth testing set.

3 | THEORETICAL RESULTS

We infer $c^T \beta^0$ for a loading vector $c \in \mathbb{R}^p$ or $A\beta^0$ for a loading matrix $A \in \mathbb{R}^{l \times p}$, by studying the asymptotic properties for linear combinations of $\widehat{b}$. Let $s_0 = \|\beta^0\|_0 = \#\{1 \le j \le p : \beta_j^0 \ne 0\}$ be the true model sparsity. Denote the expectation of $\widehat{\mu}_r(t; \beta)$ as $\mu_r(t; \beta) = E[1(Y \ge t)X^{\otimes r} \exp\{X^T \beta\}]$, and define population-level counterparts for $\widehat{\eta}_n(t; \beta)$ as $\eta_0(t; \beta) = \mu_1(t; \beta)/\mu_0(t; \beta)$, and for $\widehat{\Sigma}$ in (4) as $\Sigma_{\beta^0} = E\left[\{X - \eta_0(Y; \beta^0)\}^{\otimes 2} \delta\right]$. Denote by $\Theta_{\beta^0} = \Sigma_{\beta^0}^{-1}$. We enumerate sufficient conditions needed for establishing the theoretical properties of the debiased lasso estimator.

Assumption 1. Covariates are almost surely uniformly bounded, that is, $\|X_i\|_\infty \le K$ for some constant $K < \infty$ for $i = 1, 2, \ldots, n$.

Assumption 2. $|X_i^T \beta^0| \le K_1$ uniformly for all $i = 1, \cdots, n$ with some constant $K_1 < \infty$ almost surely.

Assumption 3. The follow-up time stops at a finite time point $\tau > 0$, where the probability $\pi_0 = P(Y \ge \tau) > 0$.

Assumption 4. Let

$$\tilde{\Sigma}_{\beta^0}(t) = \int_0^t \left\{ \mu_2(u; \beta^0) - \frac{\mu_1(u; \beta^0)\mu_1^T(u; \beta^0)}{\mu_0(u; \beta^0)} \right\} dH_0(u).$$

For any $t \in [0, \tau]$, we assume

$$\frac{c^T \Theta_{\beta^0} \tilde{\Sigma}_{\beta^0}(t) \Theta_{\beta^0} c}{c^T \Theta_{\beta^0} c} - v(t; c) \to 0, \text{ as } n \to \infty,$$

for some fixed function $v(\cdot; c) > 0$.

Assumption 5. The matrix Σ_{β^0} has bounded eigenvalues, that is, there exist two constants $\zeta_{\min}$ and $\zeta_{\max}$ such that $0 < \zeta_{\min} \leq \zeta_{\min}(\Sigma_{\beta^0}) \leq \zeta_{\max}(\Sigma_{\beta^0}) \leq \zeta_{\max} < \infty$, where $\zeta_{\min}(\Sigma_{\beta^0})$ and $\zeta_{\max}(\Sigma_{\beta^0})$ represent the smallest and the largest eigenvalues of Σ_{β^0}.

It is common in the literature of high-dimensional inference to assume bounded covariates as in Assumption 1. Fang et al. (2017) and Kong et al. (2021) also posed Assumption 2 for the Cox model inference, that is, uniform boundedness on the multiplicative hazard. Under Assumption 1, Assumption 2 can be implied by bounded overall signal $\|\beta^0\|_1$. Assumption 3 is usually used for survival models with censored data (Andersen & Gill, 1982). Assumption 4 ensures the convergence of a predictable variation process in the martingale central limit theorem and thus the asymptotic normality of the debiased lasso estimator. $\tilde{\Sigma}_{\beta^0}(t)$ can be viewed as the information matrix up to time point t. It is easy to see that $\tilde{\Sigma}_{\beta^0}(\tau) = \Sigma_{\beta^0}$ and $v(\tau; c) = 1$. This assumption states that the limiting function $v(t; c)$ also depends on $c \in \mathbb{R}^p$, the loading vector of interest, which is reasonable; we discuss a potential replacement for Assumption 4 outside the framework of martingale central limit theorem in the Appendix S1. The bounded eigenvalue condition on Σ_{β^0} in Assumption 5 is standard in inference for high-dimensional models.

Theorem 1. *Assume that the two tuning parameters satisfy $\lambda_n \asymp \sqrt{\log(p)/n}$ and $\gamma_n \asymp \|\Theta_{\beta^0}\|_{1,1} s_0 \lambda_n$. Furthermore, assume $\|\Theta_{\beta^0}\|_{1,1}^2 p s_0 \log(p)/\sqrt{n} \to 0$ as $n \to \infty$. Under Assumptions 1–5, for any $c \in \mathbb{R}^p$ such that $\|c\|_2 = 1$ and $\|c\|_1 \leq a_*$ with some absolute constant $a_* < \infty$, we have*

$$\sqrt{n} c^T (\hat{b} - \beta^0)/(c^T \hat{\Theta} c)^{1/2} \xrightarrow{D} N(0, 1).$$

Theorem 1 provides the basis for simultaneous inference on the regression coefficients, and does not involve the ℓ_0 sparsity of the rows of the true inverse information matrix Θ_{β^0}. As a trade-off, $\|\Theta_{\beta^0}\|_{1,1}$ plays an important role in the theory and can be viewed as a relaxation of the ℓ_0 sparsity.

In the following, Corollary 1(i) discusses the type I error and the power of testing $H_0 : c^T \beta^0 = a_0$ based on Theorem 1, and Corollary 1(ii) ensures that the corresponding confidence intervals achieve nominal coverage probability asymptotically.

Corollary 1. *Suppose that the assumptions in Theorem 1 hold.*

(i) *To test a null hypothesis $H_0 : c^T \beta^0 = a_0$ versus an alternative hypothesis $H_1 : c^T \beta^0 = a_1$, where $a_1 \neq a_0$, with a known $c \in \mathbb{R}^p$ and constant $a_0 \in \mathbb{R}$, let the test statistic $T = \sqrt{n}(c^T \hat{b} - a_0)/(c^T \hat{\Theta} c)^{1/2}$. We construct a test function*

$$\phi(T) = \begin{cases} 1 & \text{if } |T| > z_{\alpha/2}, \\ 0 & \text{if } |T| \leq z_{\alpha/2}, \end{cases}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of $N(0,1)$. Then, the type I error rate for the test $\phi(T)$ satisfies $P(\phi(T) = 1|H_0) \to \alpha$, and the power under the alternative H_1 satisfies $P(\phi(T) = 1|H_1) \to 1$ as $n \to \infty$.

(ii) *The two-sided level α confidence interval for $c^T\beta^0$ can be constructed as $CI(\alpha) = [c^T\hat{b} - z_{\alpha/2}(c^T\hat{\Theta}c/n)^{1/2}, \ c^T\hat{b} + z_{\alpha/2}(c^T\hat{\Theta}c/n)^{1/2}]$. Then $P(c^T\beta^0 \in CI(\alpha)|H_0) \to 1 - \alpha$ as $n \to \infty$.*

With Theorem 1 and the Cramér–Wold device, we can also conduct simultaneous inference on multiple linear combinations, that is, $A\beta^0$ for some $l \times p$ matrix A, as summarized in the following Theorem 2, with Assumption 4 replaced by its multivariate version, Assumption 6. Similarly, Corollary 2 provides the asymptotic results for hypothesis testing and confidence region in this setting.

Assumption 6. Let $\tilde{\Sigma}_{\beta^0}(t)$ be the same as in Assumption 4. For a fixed combination matrix of interest $A \in \mathbb{R}^{l \times p}$, it holds that

$$\frac{\omega^T A\Theta_{\beta^0}\tilde{\Sigma}_{\beta^0}(t)\Theta_{\beta^0}A^T\omega}{\omega^T A\Theta_{\beta^0}A^T\omega} - v'(t; A^T\omega) \to 0, \ \text{as } n \to \infty,$$

for any vector $\omega \in \mathbb{R}^l$ and any $t \in [0, \tau]$, where $v'(\cdot; A^T\omega) > 0$ is some fixed function depending on $A^T\omega$.

Theorem 2. *Let A be an $l \times p$ matrix of full row rank such that the number of rows l is fixed, $\|A\|_{\infty,\infty} = \mathcal{O}(1)$ and $A\Theta_{\beta^0}A^T \to F$ for some fixed $l \times l$ matrix F. Assume that the two tuning parameters $\lambda_n \asymp \sqrt{\log(p)/n}$ and $\gamma_n \asymp \|\Theta_{\beta^0}\|_{1,1}s_0\lambda_n$, and that $\|\Theta_{\beta^0}\|_{1,1}^2 ps_0 \log(p)/\sqrt{n} \to 0$ as $n \to \infty$. Under Assumptions 1–3, 5 and 6, we have*

$$\sqrt{n}A(\hat{b} - \beta^0) \xrightarrow{D} N(0, F).$$

Corollary 2. *Suppose the assumptions in Theorem 2 hold.*

(i) *For the $l \times p$ matrix A in Theorem 2, under the null hypothesis $H_0 : A\beta^0 = a_0$ for some $a_0 \in \mathbb{R}^l$, the statistic $T' = n(A\hat{b} - a^0)^T\hat{F}^{-1}(A\hat{b} - a^0) \xrightarrow{D} \chi_l^2$, where $\hat{F} = A\hat{\Theta}A^T$.*

(ii) *For $\alpha \in (0,1)$, let the confidence region for $A\beta^0$ be $CR(\alpha) = \{a \in \mathbb{R}^l : n(A\hat{b} - a)^T\hat{F}^{-1}(A\hat{b} - a) \le \chi_{l,\alpha}^2\}$, where $\chi_{l,\alpha}^2$ is the upper αth percentile from χ_l^2. Then $P(A\beta^0 \in CR(\alpha)|H_0) \to 1 - \alpha$ as $n \to \infty$.*

Proofs of Theorems 1 and 2 are provided in the Appendix. Corollaries 1 and 2 are directly obtained from Theorems 1 and 2, and their proofs are omitted.

4 | NUMERICAL EXPERIMENTS

For a total of $n = 500$ subjects, we simulate $p = 20, 100, 200$ covariates, respectively, and generate these covariates from $N(0, \Sigma)$, where $\Sigma = I_p$ and AR(1) with the correlation parameter of 0.5 as two different setups. Each covariate is truncated at ± 2.5. Concerning the specifications of the true regression coefficients β^0, the first element β_1^0 varies from 0 to 2 with an equal step size of 0.2, four of the other elements are arbitrarily chosen to take values of 1, 1, 0.5 and 0.5, and the rest are set

to be zero. The underlying survival times T and the censoring times C are independently generated from an exponential distribution with hazard $h(t|X) = \exp\{X^T \beta^0\}$, and from Uniform(1, 20), respectively. Under each simulation configuration, 200 datasets are generated.

The methods in comparison include: (i) QP: our proposed debiased lasso with quadratic programming for matrix $\hat{\Theta}$; (ii) NW: the debiased lasso with node-wise lasso for matrix $\hat{\Theta}$ in Kong et al. (2021); (iii) CLIME: debiased lasso with CLIME for matrix $\hat{\Theta}$ in Yu et al. (2021); (iv) Decor: decorrelated Wald test in Fang et al. (2017); (v) TPCV: projection-based cross-validation approach in Zhang et al. (2022); and (vi) Oracle: the estimator when the true model is known a priori.

For the lasso estimator, we use 10-fold cross-validation, implemented by the R package glmnet, to select the tuning parameter λ_n; see more details in Simon et al. (2011). Five-fold cross-validation is used for tuning parameter selection in CLIME, QP, and NW. For the hard-thresholding step used to select γ_n as described in Section 2.3, we adopt the Bonferroni correction with the adjusted p-value threshold $0.1/p$, where p is the number of covariates.

We compare these methods with respect to the bias of the estimated β_1^0 (the parameter of main interest), its model-based standard error, coverage probability with a significance level of $\alpha = .05$ and mean squared error. Figures 2 and 3 show the results for the independent and the AR(1) covariance structures, respectively. When $p = 20$, our proposed method (QP) and the decorrelated Wald test (Decor) perform nearly as well as the oracle estimator (Oracle) and MPLE. When the dimension is relatively large compared to the sample size, that is, $p = 100, 200$, next to Oracle, the proposed estimator (QP) displays the smallest biases and the confidence intervals with coverage probabilities closest to the nominal level 95% for both covariance structures. On the other hand, NW, CLIME, Decor, and MPLE incur substantial biases as the true value of β^0 increases. In addition, owing to the estimation of Θ_{β^0} using penalized approaches, the model-based standard error estimates using NW and CLIME are shrunk toward zero, underestimating the true variation. As such, the four competing methods, except TPCV, present improper confidence interval coverage probabilities, whereas our proposed method retains nearly unbiased estimates with coverage probabilities close to the nominal level. For inference on the low-dimensional parameter β_1^0, TPCV shows the smallest estimation bias and the confidence intervals with coverage probability closest to 95% as the true signal β_1^0 increases. However, TPCV has higher type 1 error rates when p is large; for instance, when $p = 200$ in the independent covariate setting, the type 1 error rates are 7.1% and 4.1% for TPCV and QP, respectively. Additional simulation results on simultaneous inference for all coefficients via false discovery control are provided in Appendix S1; the empirical false discovery proportions for QP, NW, and CLIME are well below the nominal level.

As the model sparsity s_0 may impact the performance of the proposed approach, we conduct additional simulations to examine its influence by varying s_0 to 3, 5, 15. With $n = 500$ subjects and $p = 100$ covariates simulated with AR(1) covariance ($\rho = 0.3$), the target parameter β_1^0 is varied from 0 to 2 as before, and the $(s_0 - 1)$ coefficients are arbitrarily chosen to be realized from a uniform distribution on $[-1, -0.5] \bigcup [0.5, 1]$. The other setup remains the same. Figure 4 shows that, when the model is more sparse ($s_0 = 3, 5$), the proposed method QP performs satisfactorily; when $s_0 = 15$, there is slight a decline in the coverage probability by QP. However, QP still outperforms NW, CLIME, and Decor; as TPCV was designed to estimate a low dimensional set of parameters, whereas our method estimates all p variables simultaneously, this may explain the slightly compromised performance of our method compared to that of TPCV in a finite sample setting.

We next compare the time spent on computing $\hat{\Theta}$ alone (Table 1) among `solve.QP` in the R package `quadprog` for the proposed quadratic programming procedure, and two commonly used R functions for CLIME, namely, `clime` in the package `clime` and `sugm` in the package `flare`.
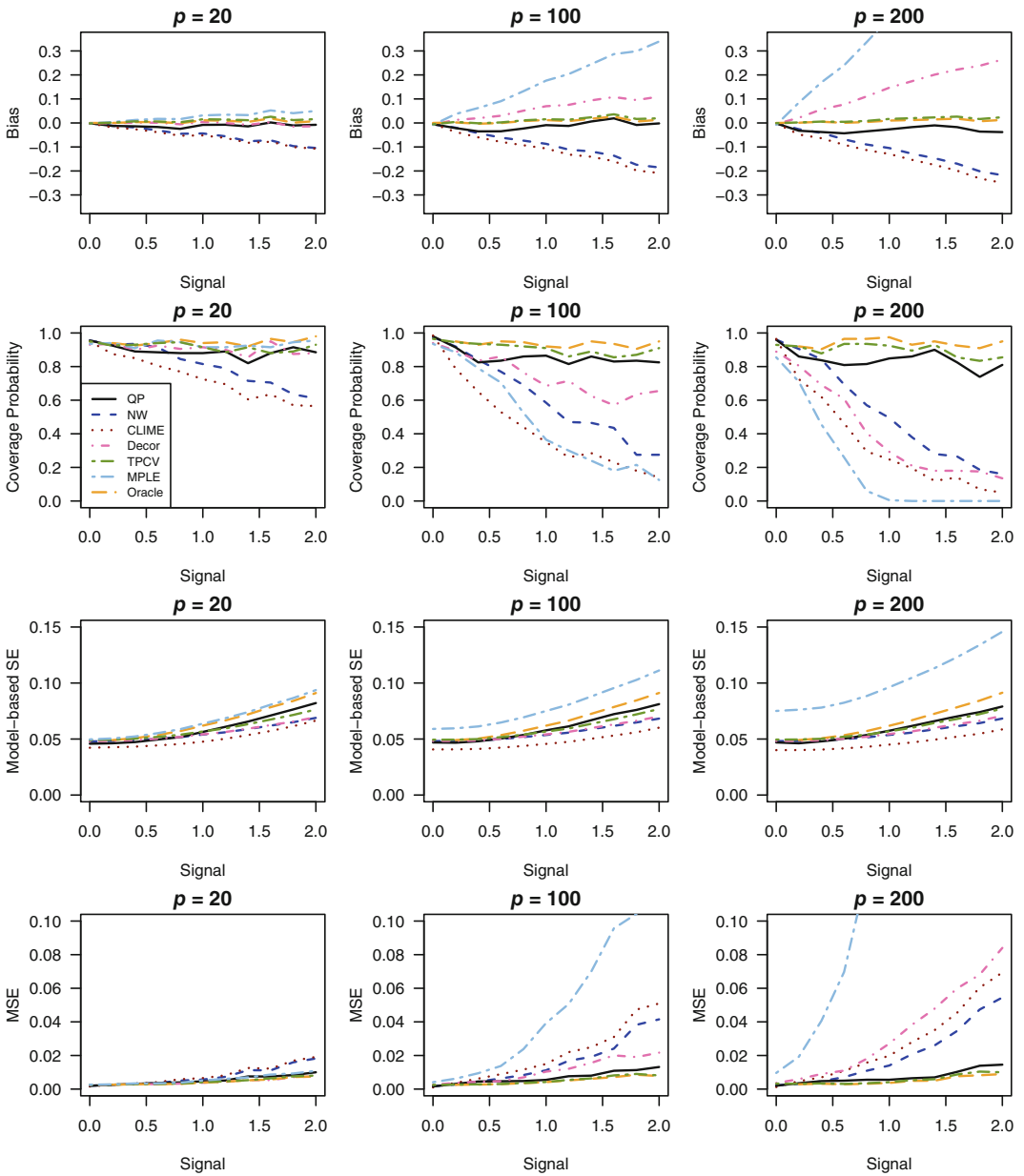
FIGURE 2 Estimation bias, coverage probability, model-based standard error and mean squared error, based on 200 simulations, each with $n = 500$ observations and independent covariance structure for covariates

Three candidate values of γ_n, namely, 0.3, 1, and 2 times of $\sqrt{\log(p)/n}$, are used for demonstration. We fix $\beta_1^0 = 1$ and simulate $n = 500$ observations, with covariates having an AR(1) covariance structure and the rest of the setting being identical to what is described in the first paragraph of this section. The time columns in Table 1 report the average computing time over 10 replications on a MacBook with 2.7 GHz Intel Core i5 processor and 8 GB memory, and the ratio columns compare the average computing time of each programming procedure to that of solve.QP for each simulation setting, respectively. Under all of the scenarios examined, our proposed
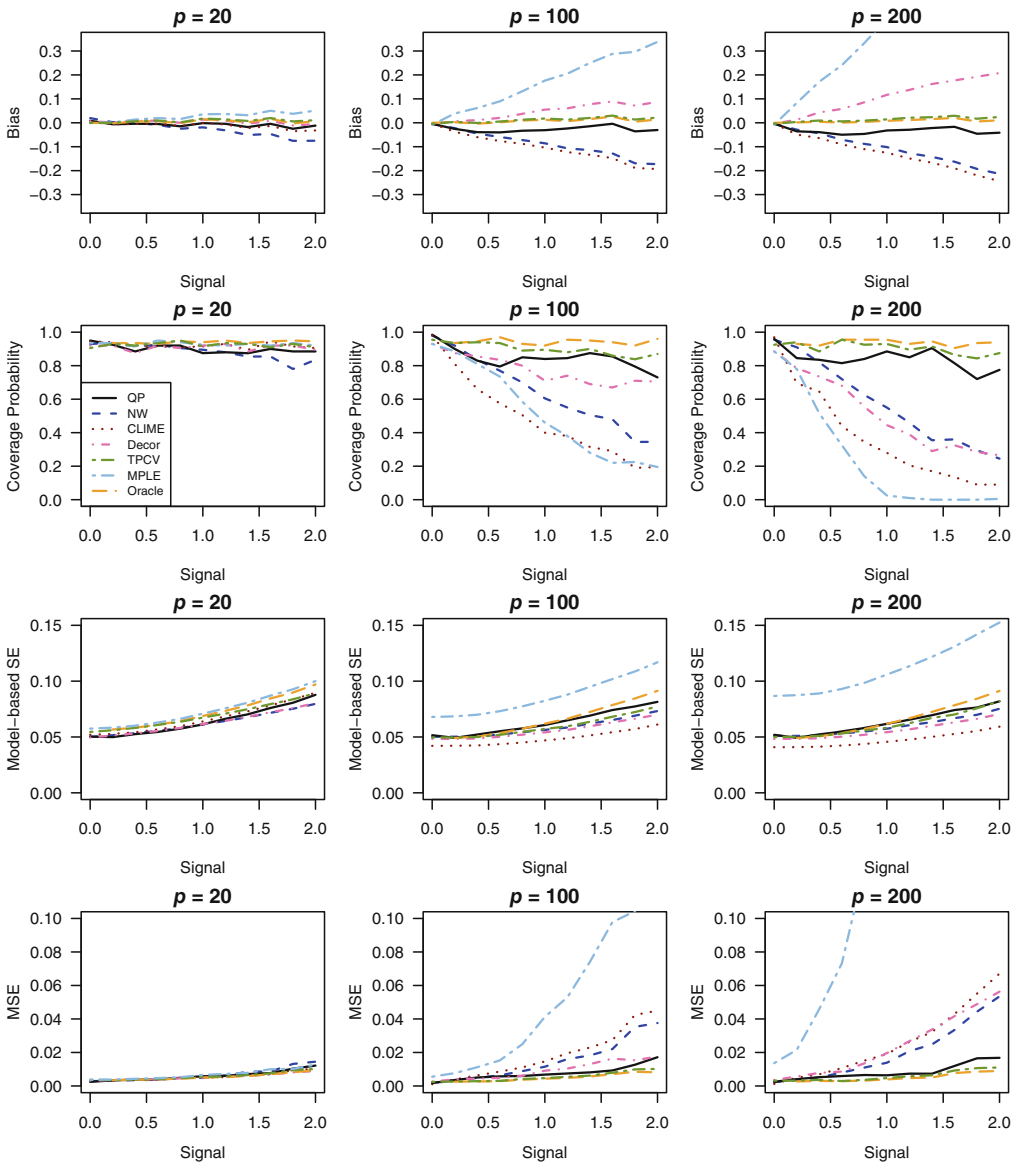
FIGURE 3 Estimation bias, coverage probability, model-based standard error and mean squared error, based on 200 simulations, each with $n = 500$ observations and AR(1) covariance structure for covariates ($\rho = 0.5$)

implementation with `solve.QP` is the most computationally efficient; for large dimensions, for example, $p = 200$, `clime` takes the longest time per dataset on average.

5 | BOSTON LUNG CANCER DATA ANALYSIS

Lung cancer is the leading cause of cancer deaths in the United States, and nonsmall cell lung cancer (NSCLC), accounting for approximately 80%–85% among all the lung cancer cases, is the most common histological type of lung cancer (Houston et al., 2018). Identification of genetic
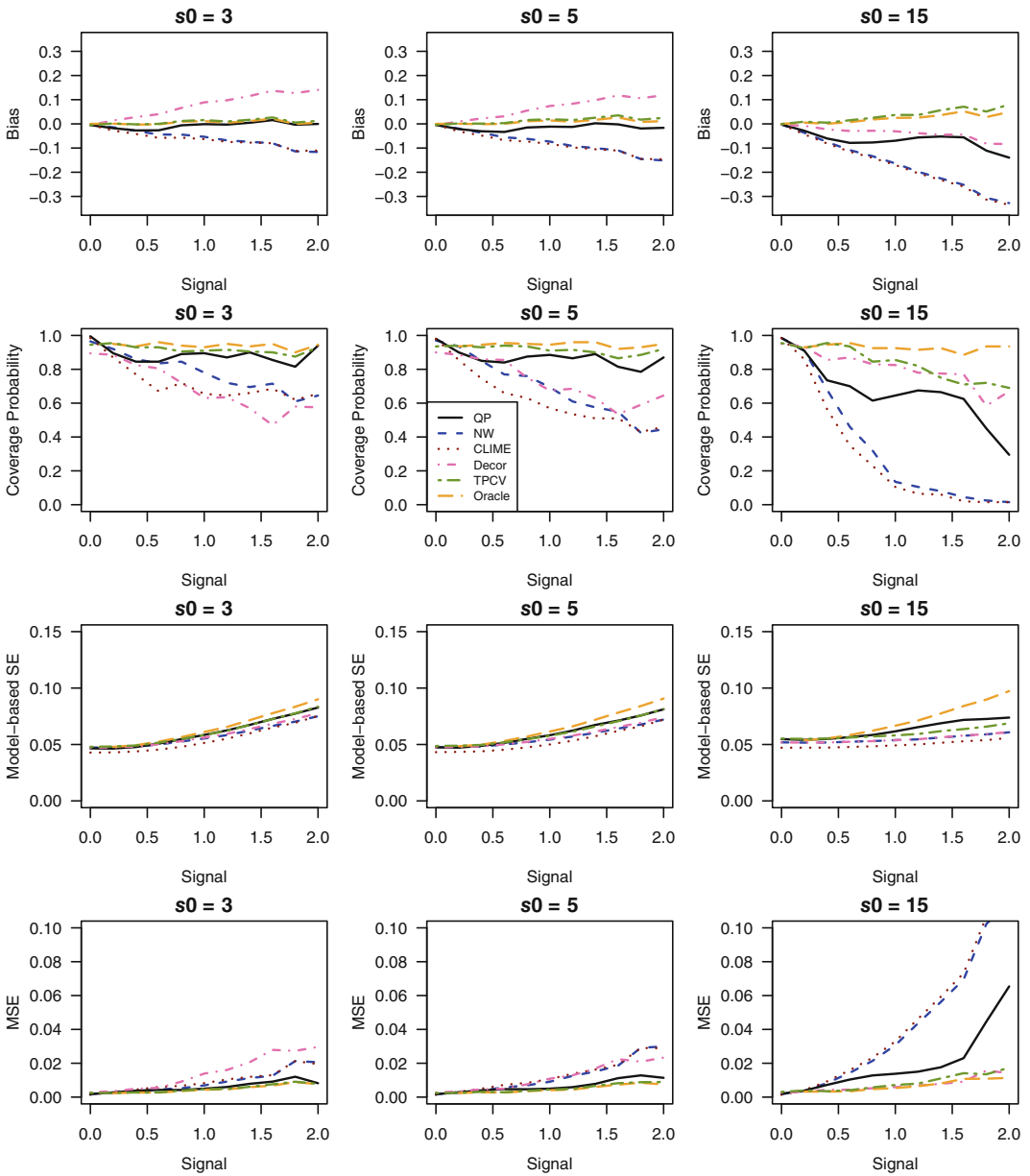
FIGURE 4 Estimation bias, coverage probability, model-based standard error and mean squared error, based on 200 simulations, each with $n = 500$ observations and AR(1) covariance structure ($\rho = 0.3$) for $p = 100$ covariates. The model sparsity parameter $s_0 = 3, 5, 15$

variants associated with lung cancer patient survival sparks modern translational cancer research, and has the potential to refine prognosis and promote individualized treatment and clinical care. Despite numerous studies investigating potential predisposing genes to lung cancer risks, studies on patient survival usually have small sample sizes and the reported genetic markers associated with lung cancer survival have been poorly replicated (Bossé & Amos, 2018). The BLCSC is a large epidemiology cohort for investigating the molecular cause underlying lung cancer, where lung

TABLE 1 Comparison of the computational time spent on computing $\hat{\Theta}$. Time (s) is averaged over 10 replications under each setting. Time ratio is with respect to the proposed method implemented using `solve.QP`.

	solve.QP		clime		flare	
$p = 20$	Time	Ratio	Time	Ratio	Time	Ratio
$\gamma_n = 0.3\sqrt{\log(p)/n}$	0.0016	1.0	0.0392	24.5	0.1898	118.6
$\gamma_n = \sqrt{\log(p)/n}$	0.0015	1.0	0.0373	24.9	0.1597	106.5
$\gamma_n = 2\sqrt{\log(p)/n}$	0.0012	1.0	0.0338	28.2	0.1522	126.8
$p = 100$						
$\gamma_n = 0.3\sqrt{\log(p)/n}$	0.3159	1.0	4.3452	13.8	5.8860	18.6
$\gamma_n = 1\sqrt{\log(p)/n}$	0.0922	1.0	3.4164	37.1	2.0754	22.5
$\gamma_n = 2\sqrt{\log(p)/n}$	0.0665	1.0	2.6281	39.5	0.3663	5.5
$p = 200$						
$\gamma_n = 0.3\sqrt{\log(p)/n}$	4.3886	1.0	64.7047	14.7	52.2224	11.9
$\gamma_n = 1\sqrt{\log(p)/n}$	0.9039	1.0	47.0320	52.0	21.7229	24.0
$\gamma_n = 2\sqrt{\log(p)/n}$	0.6196	1.0	33.0308	53.3	2.5536	4.1

cancer cases have been enrolled at Massachusetts General Hospital and the Dana-Farber Cancer Institute from 1992 to present. We apply the proposed debiased lasso method (QP) to a BLCSC cohort with genetic data and simultaneously investigate the joint effects of certain genotyped SNPs on NSCLC patient overall survival.

Included in the analysis are $n = 561$ NSCLC patients with available diagnosis dates, follow-up times and genotypes on Axiom arrays. Among all these patients, 437 (77.9%) died and 124 (22.1%) were censored. The range of the observed survival time is from 6 days to 8584 days, and the restricted mean survival and censoring times at $\tau = 8584$ days are 2124 (SE: 105) and 4397 (SE: 187) days, respectively. Patient characteristics, including age at diagnosis, race, education level, gender, smoking status, histological type, cancer stage, and treatment received, are provided in Appendix S1.

A conventional marginal association analysis (Tang et al., 2020) found two potentially functional SNPs in the genes *HDAC2* and *PPARGC1A* that were significantly associated with NSCLC overall survival. Using the target gene approach, we focus on 32 genes in the CARM ER pathway, which is the largest pathway Tang et al. (2020) considered and described in their supplementary document and contains the two reported genes *HDAC2* and *PPARGC1A*, plus 9 genes that Xia et al. (2021) studied to investigate whether the susceptibility loci are also associated with patient survival. We extract 312 genotyped SNPs from the 32 genes in the CARM ER pathway and the nine target genes described in Xia et al. (2021) from the BLCSC data (minor allele frequency > 0.01, genotype call rate > 95%). After a pruning step using PLINK (Purcell et al., 2007) to avoid multi-colinearity caused by SNPs with high linkage disequilibrium, the number of SNPs is reduced to 217. SNPs are coded by the number of copies of the minor allele, that is, 0, 1, or 2, and assumed to have additive effects on the log hazard ratio. Therefore, the subset of the BLCSC data we analyze include $n = 561$ NSCLC patients and $p = 231$ covariates.

Table 2 summarizes the coefficient estimates in the Cox proportional hazards model for all patient characteristics and the top 10 SNPs ranked by the p-values from the proposed method (QP). Results of two methods, QP versus MPLE, are listed side by side. In general, QP results in

TABLE 2 Coefficient estimates in the Cox proportional hazards model for the Boston Lung Cancer Survival Cohort data

Variable	Note	QP				MPLE			
		Est	SE	p-Value	95% CI	Est	SE	p-Value	95% CI
Race	Others versus Caucasian	−0.163	0.201	0.416	(−0.557, 0.231)	0.065	0.561	0.908	(−1.034, 1.163)
Education	HS versus No HS	−0.018	0.091	0.840	(−0.198, 0.161)	−0.142	0.253	0.574	(−0.637, 0.353)
	College versus No HS	−0.037	0.076	0.625	(−0.185, 0.111)	−0.085	0.218	0.698	(−0.513, 0.343)
Gender	Male versus Female	0.314	0.075	< 0.001	(0.166, 0.461)	0.439	0.166	0.008	(0.114, 0.763)
Age	Standardized	0.155	0.038	< 0.001	(0.081, 0.230)	0.400	0.090	< 0.001	(0.224, 0.577)
Smoker	Yes versus No	0.103	0.142	0.470	(−0.176, 0.381)	0.066	0.299	0.825	(−0.519, 0.651)
Histology	AD versus LCC	−0.259	0.076	0.001	(−0.409, −0.11)	−0.467	0.294	0.112	(−1.043, 0.109)
	SCC versus LCC	0.065	0.094	0.488	(−0.120, 0.251)	−0.030	0.314	0.923	(−0.646, 0.585)
	Unspecified versus LCC	0.046	0.132	0.729	(−0.213, 0.304)	−0.119	0.384	0.756	(−0.871, 0.633)
Stage	Late versus Early	0.352	0.081	< 0.001	(0.193, 0.510)	0.553	0.190	0.004	(0.180, 0.926)
Surgery	Yes versus No	−1.102	0.085	< 0.001	(−1.269, −0.936)	−2.115	0.226	< 0.001	(−2.557, −1.672)
Chemotherapy	Yes versus No	0.025	0.078	0.753	(−0.128, 0.177)	−0.239	0.220	0.278	(−0.671, 0.193)
Radiation	Yes versus No	0.047	0.077	0.548	(−0.105, 0.198)	0.248	0.198	0.211	(−0.140, 0.636)
Treatment record	Missing versus Not	0.099	0.176	0.573	(−0.245, 0.443)	0.347	0.428	0.417	(−0.492, 1.186)
SNP	Pos	Gene	Est	SE	p-Value	95% CI	Est	SE	p-Value	95% CI
AX-11672686	8:27324822	CHRNA2	0.186	0.054	0.001	(0.081, 0.291)	0.185	0.402	0.645	(−0.603, 0.973)
AX-11673610	12:66762242	GRIP1	0.313	0.092	0.001	(0.133, 0.494)	0.773	0.220	< 0.001	(0.343, 1.203)
AX-11264571	13:32906729	BRCA2	0.206	0.061	0.001	(0.086, 0.325)	0.450	0.164	0.006	(0.129, 0.772)
AX-40031129	16:3860539	CREBBP	−0.566	0.242	0.019	(−1.040, −0.092)	−1.504	0.623	0.016	(−2.726, −0.282)
AX-11235551	16:3832471	CREBBP	−0.130	0.057	0.022	(−0.242, −0.019)	−0.495	0.309	0.110	(−1.101, 0.112)
AX-11639833	5:88088439	MEF2C	−0.121	0.056	0.031	(−0.231, −0.011)	−0.145	0.120	0.228	(−0.381, 0.091)
AX-11326149	15:78867482	CHRNA5	0.102	0.051	0.046	(0.002, 0.202)	1.273	0.366	0.001	(0.555, 1.991)
AX-11376755	21:16340289	NRIP1	−0.101	0.052	0.052	(−0.202, 0.001)	−0.281	0.120	0.019	(−0.516, −0.046)
AX-40181207	17:41218805	BRCA1	−0.524	0.272	0.054	(−1.056, 0.009)	−2.386	0.750	0.001	(−3.856, −0.916)
AX-30854303	12:66761377	GRIP1	0.094	0.054	0.081	(−0.011, 0.199)	0.102	0.117	0.380	(−0.126, 0.331)
...									

Abbreviations: AD, Adenocarcinoma; CI, confidence interval; Est, coefficient estimate; HS, high school; LCC, large cell carcinoma; Pos, physical location based on Assembly GRCh37/hg19; SCC, squamous cell carcinoma; SE, standard error estimate.

points estimates of smaller magnitudes and smaller standard errors compared to MPLE, which is consistent with our observation in the simulated example. MPLE is numerically very unstable when the dimension p is large compared to the sample size n. The numerical instability arises primarily from inverting the Hessian matrix, which may be closer to being singular. On the contrary, Lasso provides a more stabilized initial estimator. As a result, the debiased lasso estimator is numerically more stable than MPLE with narrower confidence intervals. When the dimension p is very small, the difference between the two methods becomes negligible.

Among various patient characteristics, QP found that the adenocarcinoma subtype is significantly associated with better patient survival than large cell carcinoma, consistent with the results of Janssen-Heijnen and Coebergh (2001), which was, however, not detected by MPLE. QP further identified that AX-11672686 in *CHRNA2*, AX-11673610 in *GRIP2,* and AX-11264571 in *BRCA2* are the three most significant SNPs associated with NSCLC patient survival, after adjusting for all the other demographic and genetic risk factors. Interestingly, AX-11672686 was found to be associated with nicotine dependence by Wang et al. (2014). AX-11264571 has been found to be associated with breast cancer (Qiu et al., 2010) and may also be associated with lung cancer susceptibility, although not achieving genome-wide significance in Yu et al. (2011). AX-11673610 or *GRIP1* seems to be a new finding as, to our knowledge, they have yet been reported in the lung cancer literature (Bossé & Amos, 2018)

To understand the impact of the socioeconomic status on cancer survival, we test for the association between education level (no high school, high school, or at least 1–2 years of college) and lung cancer patient survival. With a loading matrix $A_{2 \times p} = (e_2, e_3)^T$ corresponding to the contrast of the effects of high school graduate and at least 1–2 years of college with the reference level of no high school, the test statistic is 0.259 with a *p*-value of .879, suggesting no statistical evidence for the association between education level and NSCLC patient survival, after adjusting all other demographic characteristics and genetic markers. The results confirm a large-scale clinical trial on lung cancer patients which reported "education level was not predictive of survival" (Herndon et al., 2008).

In summary, these results illustrate the utility of our method in providing reliable inference for scientific discovery and interpretation, while more in-depth biological investigations are warranted to validate our findings.

6 | CONCLUDING REMARKS

We have proposed a debiased lasso approach for reliable estimation and inference in the Cox proportional hazards model when $p < n$ but is allowed to diverge to ∞ with n. Unlike existing methods (Fang et al., 2017; Kong et al., 2021; Yu et al., 2021), we resort to a quadratic programming procedure for estimating the inverse information matrix, without imposing an unrealistic sparsity assumption on it. The proposed debiased lasso estimator is asymptotically unbiased and normally distributed under mild regularity conditions. Our simulations demonstrate that, when p is very small, the proposed method behaves similarly to the conventional MPLE; when p is relatively large, it outperforms the competitors in bias correction and confidence interval coverage.

The rationale behind using a quadratic programming procedure to estimate the inverse information matrix has stemmed from the work in linear models by Javanmard and Montanari (2014). The main advantages of this approach are: (i) we do not need to impose unrealistic ℓ_0 sparsity assumptions on the true inverse information matrix Θ_{β^0}; (ii) it can be implemented by using existing packages and parallelized. A limitation of the proposed method is that the developed theory

in Section 3 applies only to the "large n, diverging p" scenario; our numerical experience seems to suggest that, although the proposed method can be numerically carried out when $p > n$, only small γ_n values are feasible for quadratic programming in (3) to have stable solutions.

Selection of tuning parameters has been a persistent issue in high-dimensional estimation and inference. By theoretical and empirical studies, we have found that the proposed inferential method is not very sensitive to the tuning parameter λ_n, provided that λ_n lies within a reasonable range for sustaining reasonable lasso estimates. In contrast, choices of γ_n may more influence the results. Currently, we use a single tuning parameter γ_n to govern the debiased estimates for all β_j's as shown in (3). Along the line of Cai et al. (2016) and to adapt to the variability of individual coefficients, we may estimate the rows of $\hat{\Theta}$ via a series of optimization problems for $j = 1, \cdots, p$:

$$\min_{m \in \mathbb{R}^p} \left\{ m^T \hat{\Sigma} m : |(\hat{\Sigma} m - e_j)_k| \le \gamma_{n,jk}, 1 \le k \le p \right\},$$

where $\gamma_{n,jk}$'s are adaptively estimated via a properly designed procedure. However, it would involve complicated theoretical analysis in the setting of Cox models, for example, deriving the rates of $\gamma_{n,jk}$ and designing an adaptive procedure for estimating $\gamma_{n,jk}$. This level of complexity might be beyond the scope of this article, and we will pursue it more thoroughly in the future.

Lastly, we touch upon the important issue of drawing inference with $p > n$, though not a main focus of this paper. Several methods (Fang et al., 2017; Kong et al., 2021; Yu et al., 2021) had been developed for handling "$p > n$" inference problems; however, our analytical and simulation studies have pinpointed their possible limitations in providing sufficient bias correction and reliable confidence intervals even within the "large n, diverging p" framework, likely due to the sparsity assumptions on the inverse information matrix that may not hold in survival settings. One possible solution, by going beyond the debiased lasso framework, is to perform repeated data splitting for model selection and estimation on two separate parts of the data and smooth the resulting estimates from multiple splits; see Fei and Li (2021) for inference on high-dimensional generalized linear models. The validity of the method hinges upon the sure screening property for the initial model selection, and we will explore its use in survival settings in the future.

SUPPORTING INFORMATION
Additional information for this article is available online in the Supporting Information section, including proofs of the technical lemmas and additional numerical results. The R and Rcpp code files for the proposed approach can be found on https://github.com/luxia-bios/CoxInference/.

ORCID
Yi Li https://orcid.org/0000-0003-1720-2760

REFERENCES
Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, *10*, 1100–1120.

Antoniadis, A., Fryzlewicz, P., & Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, *37*, 531–552.

Bossé, Y., & Amos, C. I. (2018). A decade of GWAS results in lung cancer. *Cancer Epidemiology, Biomarkers & Prevention*, *27*, 363–379.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cai, T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*, 594–607.

Cai, T. T., Liu, W., & Zhou, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, *44*, 455–488.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*, 187–202.

Fan, J., & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*, 74–99.

Fang, E. X., Ning, Y., & Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1415–1437.

Fei, Z., & Li, Y. (2021). Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *Journal of Machine Learning Research*, *22*, 1–32.

Gui, J., & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, *21*, 3001–3008.

Herndon, J. E., Kornblith, A. B., Holland, J. C., & Paskett, E. D. (2008). Patient education level as a predictor of survival in lung cancer clinical trials. *Journal of Clinical Oncology*, *26*, 4116–4123.

Houston, K. A., Mitchell, K. A., King, J., White, A., & Ryan, B. M. (2018). Histologic lung cancer incidence rates and trends vary by race/ethnicity and residential county. *Journal of Thoracic Oncology*, *13*, 497–509.

Huang, J., Sun, T., Ying, Z., Yu, Y., & Zhang, C.-H. (2013). Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, *41*, 1142–1165.

Janssen-Heijnen, M. L. G., & Coebergh, J.-W. W. (2001). Trends in incidence and prognosis of the histological subtypes of lung cancer in North America, Australia, New Zealand and Europe. *Lung Cancer*, *31*, 123–137.

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

Kong, S., & Nan, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. *Statistica Sinica*, *24*, 25–42.

Kong, S., Yu, Z., Zhang, X., & Cheng, G. (2021). High-dimensional robust inference for Cox regression models using desparsified lasso. *Scandinavian Journal of Statistics*, *48*, 1068–1095.

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., … and Amos, C. I. (2017) Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, *49*, 1126–1132.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … and Sham, P. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575.

Qiu, L.-X., Yao, L., Xue, K., Zhang, J., Mao, C., Chen, B., … and Hu, X.-C. (2010) BRCA2 N372H polymorphism and breast cancer susceptibility: A meta-analysis involving 44,903 subjects. *Breast Cancer Research and Treatment*, *123*, 487–490.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1–13.

Tang, D., Zhao, Y. C., Qian, D., Liu, H., Luo, S., Patz, E. F., … and Wei, Q. (2020) Novel genetic variants in HDAC2 and PPARGC1A of the CREB-binding protein pathway predict survival of non-small-cell lung cancer. *Molecular Carcinogenesis*, *59*, 104–115.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*, 385–395.

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

Wang, S., van der Vaart, A. D., Xu, Q., Seneviratne, C., Pomerleau, O. F., Pomerleau, C. S., ... and Li, M. D. (2014) Significant associations of CHRNA2 and CHRNA6 with nicotine dependence in European American and African American populations. *Human Genetics*, *133*, 575–586.

Xia, L., Nan, B., & Li, Y. (2021). Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics* in press. https://doi.org/10.1111/biom.13587

Yu, H., Zhao, H., Wang, L.-E., Han, Y., Chen, W. V., Amos, C. I., Rafnar, T., Sulem, P., Stefansson, K., Landi, M. T., & Caporaso, N. (2011). An analysis of single nucleotide polymorphisms of 125 DNA repair genes in the Texas genome-wide association study of lung cancer with a replication for the XRCC4 SNPs. *DNA Repair*, *10*, 398–407.

Yu, Y., Bradic, J., & Samworth, R. J. (2021). Confidence intervals for high-dimensional Cox models. *Statistica Sinica*, *31*, 243–267.

Zhang, C.-H., & 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)*, *76*, 217–242.

Zhang, H., Huang, J., & Sun, L. (2022). Projection-based and cross-validated estimation in high-dimensional Cox model. *Scandinavian Journal of Statistics*, *49*, 353–372.

APPENDIX

We first present the useful lemmas for proving the main theorems, with detailed proofs deferred to Appendix S1. Some of these lemmas present important results in their own right. The proofs of the Theorem 1 and Theorem 2 are presented following the lemmas.

Additional notation from counting processes and martingale theory is defined for the proofs. Under the Cox model, define the counting process $N_i(t) = 1(Y_i \leq t, \delta_i = 1)$ and its compensator $A_i(t; \beta) = \int_0^t 1(Y_i \geq s) \exp(X_i^T \beta) dH_0(s)$, where $H_0(t) = \int_0^t h_0(s) ds$ is the cumulative baseline hazard function, $i = 1, \cdots, n$. Let $M_i(t; \beta) = N_i(t) - A_i(t; \beta)$, and $M_i(t; \beta^0)$ is a martingale with respect to the filtration $\mathcal{F}_i(t) = \sigma\{N_i(s), 1(Y_i \geq s), X_i : s \in (0, t]\}$. It follows that $\hat{\eta}_n(t; \beta)$, and in particular, $\hat{\eta}_n(t; \beta^0)$, is predictable with respect to the filtration $\mathcal{F}(t) = \sigma\{N_i(s), 1(Y_i \geq s), X_i : s \in (0, t], i = 1, \cdots, n\}$, an observation useful for derivations. Notation-wise, we do not distinguish between the usual expectation and the outer expectation.

Lemma A1 below characterizes the difference between $\hat{\eta}_n(t; \beta^0)$ and $\eta_0(t; \beta^0)$, which facilitates the proof of the asymptotic distribution for the leading term $\sqrt{n} c^T \Theta_{\beta^0} \dot{\ell}_n(\beta^0)$ as well as the establishment of the convergence rate for $\hat{\Sigma} - \Sigma_{\beta^0}$.

Lemma A1. *Under Assumptions 1–3, we have*

$$\sup_{t \in [0, \tau]} |\hat{\mu}_0(t; \beta^0) - \mu_0(t; \beta^0)| = \mathcal{O}_P(\sqrt{\log(p)/n}),$$

$$\sup_{t \in [0, \tau]} \|\hat{\mu}_1(t; \beta^0) - \mu_1(t; \beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}),$$

$$\sup_{t \in [0, \tau]} \|\hat{\eta}_n(t; \beta^0) - \eta_0(t; \beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

Lemma A2 establishes the asymptotic distribution for the leading term $-c^T \Theta_{\beta^0} \dot{\ell}_n(\beta^0)$ in the decomposition of $c^T(\hat{b} - \beta^0)$.

Lemma A2. *Assume $p^2 \log(p)/n \to 0$. Under Assumptions 1–5, for any $c \in \mathbb{R}^p$ such that $\|c\|_2 = 1$ and $\|c\|_1 \leq a_*$ with some absolute constant $a_* < \infty$,*

$$\frac{\sqrt{n} c^T \Theta_{\beta^0} \dot{\ell}_n(\beta^0)}{\sqrt{c^T \Theta_{\beta^0} c}} \xrightarrow{D} N(0, 1).$$

Lemma A3 provides theoretical properties of the lasso estimator in the Cox model. This is a direct result from theorem 1 in Kong and Nan (2014), and thus the proof is omitted.

Lemma A3. *Under Assumptions 1–5, for the lasso estimator $\hat{\beta}$, we have*

$$\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_P(s_0 \lambda_n), \quad \frac{1}{n} \sum_{i=1}^{n} |X_i^T(\hat{\beta} - \beta^0)|^2 = \mathcal{O}_P(s_0 \lambda_n^2),$$

where $s_0 = |\{j : \beta_j^0 \neq 0, j = 1, \cdots, p\}|$ is the true model size.

Lemma A4. *Under Assumptions 1–5, if $\lambda_n \asymp \sqrt{\log(p)/n}$, with probability going to 1, we have $\|\Theta_{\beta^0} \hat{\Sigma} - I_p\|_\infty \leq \gamma_n$, for $\gamma_n \asymp \|\Theta_{\beta^0}\|_{1,1} s_0 \lambda_n$.*

Lemma A4 shows that, unlike in a linear regression model where the tuning parameter in the constraint takes the order of $\sqrt{\log(p)/n}$, the Cox model requires a potentially larger γ_n for the feasibility of Θ_{β^0} depending on $\|\Theta_{\beta^0}\|_{1,1}$, because the information matrix involves the regression coefficients.

Lemma A5. *Assume $\limsup_{n \to \infty} p \gamma_n \leq 1 - \epsilon'$ for some $\epsilon' \in (0, 1)$. Then, under the assumptions in Lemma A4, $\|\hat{\Theta} - \Theta_{\beta^0}\|_\infty = \mathcal{O}_P(\gamma_n \|\Theta_{\beta^0}\|_{1,1})$.*

Lemma A6. *Under Assumptions 1–3 and 5, for each $t > 0$,*

$$P\left(\|\dot{\ell}_n(\beta^0)\|_\infty > t\right) \leq 2p e^{-nt^2/(8K^2)}.$$

Now we complete the proofs of Theorem 1 and Theorem 2.

Proof of Theorem 1. The first-order Taylor expansion of $\dot{\ell}_{nj}(\hat{\beta})$, the jth component in $\dot{\ell}_n(\hat{\beta})$, at β^0, is

$$\dot{\ell}_{nj}(\hat{\beta}) = \dot{\ell}_{nj}(\beta^0) + [\ddot{\ell}_{nj}(\tilde{\beta}^{(j)})]^T(\hat{\beta} - \beta^0), \tag{A1}$$

where $\tilde{\beta}^{(j)}$ lies between $\hat{\beta}$ and β^0, and $\ddot{\ell}_{nj}(\beta)$ denotes the jth column in the Hessian matrix $\ddot{\ell}_n(\beta)$. Let the $p \times p$ matrix $B_n = (\ddot{\ell}_{n1}(\tilde{\beta}^{(1)}), \ldots, \ddot{\ell}_{np}(\tilde{\beta}^{(p)}))^T$. Suppose $c \in \mathbb{R}^p$ is a p-dimensional vector, and the parameter of interest is $c^T \beta^0$. Plugging (A1) in (2), we have

$$c^T(\hat{b} - \beta^0) = -c^T \Theta_{\beta^0} \dot{\ell}_n(\beta^0) - c^T(\hat{\Theta} - \Theta_{\beta^0}) \dot{\ell}_n(\beta^0)$$
$$- c^T(\hat{\Theta}\hat{\Sigma} - I_p)(\hat{\beta} - \beta^0) + c^T \hat{\Theta}(\hat{\Sigma} - B_n)(\hat{\beta} - \beta^0). \tag{A2}$$

The first term in (A2) is the leading part and is asymptotically normal as shown in Lemma A2, and the others will be proved to be asymptotically negligible.

First, we show that $\sqrt{n}c^T(\widehat{\Theta} - \Theta_{\beta^0})\dot{\ell}_n(\beta^0) = o_P(1)$. By Lemma A5 and Lemma A6,

$$
\begin{aligned}
|\sqrt{n}c^T(\widehat{\Theta} - \Theta_{\beta^0})\dot{\ell}_n(\beta^0)| &\leq \sqrt{n}\|c\|_1 \cdot \|\widehat{\Theta} - \Theta_{\beta^0}\|_{\infty,\infty} \cdot \|\dot{\ell}_n(\beta^0)\|_\infty \\
&\leq \sqrt{n}a_* \mathcal{O}_P(p\gamma_n\|\Theta_{\beta^0}\|_{1,1})\mathcal{O}_P(\sqrt{\log(p)/n}) \\
&= \mathcal{O}_P(\|\Theta_{\beta^0}\|_{1,1}p\gamma_n\sqrt{\log(p)}) \\
&= o_P(1).
\end{aligned}
$$

Second, we show that $\sqrt{n}c^T(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0) = o_P(1)$. By Lemma A3,

$$
\begin{aligned}
|\sqrt{n}c^T(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0)| &\leq \sqrt{n}\|c\|_1\|(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0)\|_\infty \\
&\leq \sqrt{n}a_*\|\widehat{\Theta}\widehat{\Sigma} - I_p\|_\infty\|\widehat{\beta} - \beta^0\|_1 \\
&\leq \sqrt{n}a_*\gamma_n\|\widehat{\beta} - \beta^0\|_1 \\
&= \mathcal{O}_P(\sqrt{n}\gamma_n s_0 \lambda_n) \\
&= o_P(1).
\end{aligned}
$$

Next, we show that $\sqrt{n}c^T\widehat{\Theta}(\widehat{\Sigma} - B_n)(\widehat{\beta} - \beta^0) = o_P(1)$. Note that

$$
\widehat{\Sigma} - B_n = (\widehat{\Sigma} - \Sigma_{\beta^0}) + (\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0)) + (\ddot{\ell}_n(\beta^0) - B_n). \tag{A3}
$$

By the proof of Lemma A4, we see that with $\lambda_n \asymp \sqrt{\log(p)/n}$, $\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty = \mathcal{O}_P(s_0\lambda_n)$. We rewrite

$$
\begin{aligned}
\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0) &= \mathbb{E}\int_0^\tau \{X_i - \eta_0(t;\beta^0)\}^{\otimes 2}e^{X_i^T\beta^0}1(Y_i \geq t)h_0(t)dt \\
&\quad - \int_0^\tau \left\{\widehat{\mu}_2(t;\beta^0) - \frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}h_0(t)dt \\
&\quad - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2}\right\}dM_i(t) \\
&= \int_0^\tau \{\mu_2(t;\beta^0) - \widehat{\mu}_2(t;\beta^0)\}h_0(t)dt \\
&\quad + \int_0^\tau \left\{\frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \frac{\mu_1(t;\beta^0)\mu_1^T(t;\beta^0)}{\mu_0(t;\beta^0)}\right\}h_0(t)dt \\
&\quad - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2}\right\}dM_i(t). \tag{A4}
\end{aligned}
$$

Similar to the proof in Lemma A1, we can show that $\sup_{t\in[0,\tau]}\|\widehat{\mu}_2(t;\beta^0) - \mu_2(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$, and thus $\|\int_0^\tau\{\mu_2(t;\beta^0) - \widehat{\mu}_2(t;\beta^0)\}h_0(t)dt\|_\infty \leq \sup_{t\in[0,\tau]}\|\widehat{\mu}_2(t;\beta^0) - \mu_2(t;\beta^0)\|_\infty\int_0^\tau h_0(t)dt = \mathcal{O}_P(\sqrt{\log(p)/n})$. Since

$$
\frac{\widehat{\mu}_1\widehat{\mu}_1^T}{\widehat{\mu}_0} - \frac{\mu_1\mu_1^T}{\mu_0} = \frac{\widehat{\mu}_1\widehat{\mu}_1^T}{\widehat{\mu}_0\mu_0}(\mu_0 - \widehat{\mu}_0) + \frac{1}{\mu_0}[(\widehat{\mu}_1 - \mu_1)\widehat{\mu}_1^T + \mu_1(\widehat{\mu}_1 - \mu_1)^T],
$$

in the second term of (A4), by Assumption 1 and Lemma A1,

$$\left\| \int_0^\tau \left\{ \frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \frac{\mu_1(t;\beta^0)\mu_1^T(t;\beta^0)}{\mu_0(t;\beta^0)} \right\} h_0(t)dt \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

$n^{-1}\sum_{i=1}^n \int_0^\tau \left\{ \mu_2(t;\beta^0)/\mu_0(t;\beta^0) - [\mu_1(t;\beta^0)/\mu_0(t;\beta^0)]^{\otimes 2} \right\} dM_i(t)$ is a sum of n independent and identically distributed mean zero terms, and each term $\left\| \int_0^\tau \left\{ \mu_2(t;\beta^0)/\mu_0(t;\beta^0) - [\mu_1(t;\beta^0)/\mu_0(t;\beta^0)]^{\otimes 2} \right\} dM_i(t) \right\|_\infty$ is bounded by $2K^2(1+e^{K_1}H_0(\tau))$ uniformly for all i and $t \in [0,\tau]$. Similar to the proof of $\|A_n\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$ in Lemma A4, by Hoeffding's concentration inequality,

$$\left\| \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)} \right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

It is easy to see that

$$\sup_{t \in [0,\tau]} \left\| \left\{ \frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} \right]^{\otimes 2} \right\} - \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)} \right]^{\otimes 2} \right\} \right\|_\infty = \mathcal{O}_P\left(\sqrt{\frac{\log(p)}{n}} \right).$$

Then

$$\left\| \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} \right]^{\otimes 2} \right\} dM_i(t) \right.$$

$$\left. - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)} \right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty = \mathcal{O}_P\left(\sqrt{\frac{\log(p)}{n}} \right),$$

and thus for the third term in (A4),

$$\left\| \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} \right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

Therefore, by (A4), $\|\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$.

For the (j,k)th element in $\ddot{\ell}_n(\beta)$, denoted as $\ddot{\ell}_{njk}(\beta)$, by the mean value theorem, we have

$$\ddot{\ell}_{njk}(\tilde{\beta}^{(j)}) - \ddot{\ell}_{njk}(\beta^0) = (\tilde{\beta}^{(j)} - \beta^0)^T \frac{\partial \ddot{\ell}_{njk}(\beta)}{\partial \beta} \Bigg|_{\beta = \overline{\beta}^{(jk)}},$$

where $\overline{\beta}^{(jk)}$ lies in the segment between $\tilde{\beta}^{(j)}$ and β^0. Under Assumptions 1–3, when $\|\beta - \beta^0\|_1 \leq \delta'$ for $\delta' > 0$ small enough, $\left\| \partial \ddot{\ell}_{njk}(\beta)/\partial\beta \right\|_\infty$ is bounded by some constant related to δ' uniformly for all (j,k). Since $s_0\lambda_n = o(1)$, we have $\|B_n - \ddot{\ell}_n(\beta^0)\|_\infty \leq \mathcal{O}_P(\|\widehat{\beta} - \beta^0\|_1) = \mathcal{O}_P(s_0\lambda_n)$.

Combining the three parts in (A3), we have that for $\lambda_n \asymp \sqrt{\log(p)/n}$, $\|\widehat{\Sigma} - B_n\|_\infty = \mathcal{O}_P(s_0 \lambda_n)$. Then

$$
\begin{aligned}
|\sqrt{n} c^T \widehat{\Theta}(\widehat{\Sigma} - B_n)(\widehat{\beta} - \beta^0)| &\leq \sqrt{n} \|c\|_1 \|\widehat{\Theta}\|_{\infty,\infty} \|\widehat{\Sigma} - B_n\|_\infty \|\widehat{\beta} - \beta^0\|_1 \\
&\leq \mathcal{O}_P(\sqrt{n} \|\Theta_{\beta^0}\|_{1,1}(s_0 \lambda_n)^2) \\
&= o_P(1).
\end{aligned}
$$

We show that the variance estimator is consistent, that is, $c^T(\widehat{\Theta} - \Theta_{\beta^0})c \to_P 0$ as $n \to \infty$.

$$
\begin{aligned}
|c^T(\widehat{\Theta} - \Theta_{\beta^0})c| &\leq \|c\|_1^2 \|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty \\
&\leq a_*^2 \mathcal{O}_P(\gamma_n \|\Theta_{\beta^0}\|_{1,1}) = o_P(1).
\end{aligned}
$$

Finally, by the arguments above and Slutsky's theorem, it holds that $\sqrt{n} c^T(\widehat{b} - \beta^0)/(c^T \widehat{\Theta} c)^{1/2} \xrightarrow{D} N(0, 1)$. ∎

Proof of Theorem 2. We prove Theorem 2 using the Cramér–Wold device. For any $\omega \in \mathbb{R}^l$, where the dimension l is a fixed integer free of n and p, let $c = A^T \omega$ in Theorem 1. Essentially, we only require $\|c\|_1 = \|A^T \omega\|_1$ is upper bounded, and it is not essential to force $\|c\|_2 = 1$. Since $\|A\|_{\infty,\infty} = \mathcal{O}(1)$ (by assumption) and $\|\omega\|_1 = \mathcal{O}(1)$ (fixed l), then $\|A^T \omega\|_1 \leq \|A^T\|_{1,1} \|\omega\|_1 = \|A\|_{\infty,\infty} \|\omega\|_1 = \mathcal{O}(1)$. ∎