Statistical Inference for Diverging Number of Parameters beyond Linear Regression

by

Lu Xia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Yi Li, Co-Chair
Professor Bin Nan, Co-Chair
Professor Moulinath Banerjee
Professor Bhramar Mukherjee

Lu Xia

luxia@umich.edu

ORCID iD: 0000-0003-1561-6871

To Wenbo and My Parents

ACKNOWLEDGEMENTS

Over the past five years, I have had an enjoyable and rewarding PhD life studying biostatistics at the University of Michigan. This dissertation would not have been possible without the support that I received from many people.

First and foremost, I would like to express my sincere gratitude to my advisors, Drs. Bin Nan and Yi Li, for their mentorship. They have been extremely knowledgeable, understanding and supportive, and I have been more than fortunate to receive valuable guidance from them on how to conduct good statistical research. Bin has been strict on academic training but very warm in person, which made our research discussions fruitful and pleasant, and helped me stay calm and focused when facing obstacles. Yi has shown an infectious enthusiasm to help students grow and has kindly offered many opportunities for me to learn interesting research and many other aspects from himself and his collaborators. They have set such great examples as scholars in the field that I will keep looking up to and learning from in my future career.

I would also like to thank my other committee members, Drs. Bhramar Mukherjee and Moulinath Banerjee, for their insightful suggestions on improving my dissertation. Bhramar has always been a caring and inspiring model to me. I am deeply grateful to her for supporting me to work on two projects with her in the summer of 2016, which has greatly broadened my understanding in gene-environment interactions and health disparities among women and children. I am also very grateful to Mouli for sharing his great expertise in high-dimensional statistics that helped improve the presentation

of this dissertation, and for his wonderful lectures in theoretical statistics during the early days of my graduate study.

My sincere appreciation also goes out to Drs. David Christiani and Kevin He. I must thank David and his lab members at Harvard for sharing their expertise on genetics and the Boston Lung Cancer Study, and Kevin on the UNOS kidney transplant data. Without their help, the data applications in this dissertation would not be in shape. I am also deeply greatful to Kevin for organizing the study groups that I benefit a lot from for my research, and for his generous advice on career development.

I would like to say a big thank you to our knowledgeable faculty, friendly staff and fellow students in the Michigan Biostatistics family. Especially, it has been a delightful and enriching experience working with Drs. Jack Kalbfleisch, Kevin He and Yanming Li as a graduate student research assistant. Special thanks to my friends who have been offering help in my dissertation and life, listed in alphabetical order: Zhe Fei, Cui Guo, Lu Tang, Yun Wei, Yuan Yang, Xutong Zhao, Yingchao Zhong, and Yiwang Zhou, and I also thank all members in Dr. Li's lab for their helpful feedback in the weekly group meetings.

Finally, I would like to send my deepest love to my husband, Wenbo, and my parents. I owe them a debt of gratitude for their unconditional love and support over the years, without which I would not have gotten where I am.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABSTRACT

In the big data era, regression models with a large number of covariates have emerged as a common tool to tackle problems arising from business, engineering, genomics, neuroimaging, and epidemiological studies. Drawing statistical inference for these models has sparked much interest over the past few years. Albeit successful for high dimensional linear models, high dimensional inference approaches beyond linear regression are limited and present unsatisfactory performance, theoretically or numerically. In this dissertation, we focus on de-biased lasso, which has been one of the most popular methods for high dimensional inferences. We propose procedures that provide better bias correction and confidence interval coverage, and draw reliable inference for regression parameters in the "large n, diverging p" scenario. In general, we caution against applying de-biased lasso and its variants to models beyond linear regression when parameters outnumber the sample size.

Following an overview outlined in Chapter I, we focus on the generalized linear models (GLMs) in Chapter II. Extensive numerical simulations indicate that de-biased lasso may not adequately remove biases for high dimensional GLMs, and thus yield unreliable confidence intervals. We have further found that several key assumptions, especially the sparsity condition on the inverse Hessian matrix, may not hold for GLMs. In a "large n, diverging p" scenario, we consider an alternative de-biased lasso approach that inverts the Hessian matrix of the concerned model without requiring matrix sparsity, and establish the asymptotic distributions of linear combinations of the estimates. Simulations evidence that our proposed de-biased estimator performs

better in bias correction and confidence interval coverage for a wide range of p/n ratios. We apply our method to the Boston Lung Cancer Study, an epidemiology study on the mechanisms underlying lung cancer, and investigate the joint effects of genetic variants on overall lung cancer risks.

In Chapter III, we draw inference based on the Cox proportional hazards model with a diverging number of covariates. As the existing methods assume sparsity on the inverse of the Fisher information matrix, which may not hold for Cox models, they typically generate biased estimates and under-covered confidence intervals. We modify de-biased lasso by using quadratic programming to approximate the inverse of the information matrix, without posing matrix sparsity assumptions. We establish the asymptotic theory for the estimated regression coefficients when the covariate dimension diverges with the sample size. With extensive simulations, our proposed method provides consistent estimates and confidence intervals with improved coverage probabilities. We apply the proposed method to assess the effects of genetic markers on overall survival of non-small cell lung cancer patients in the aforementioned Boston Lung Cancer Study.

Stratified Cox proportional hazards model, with extensive applications in large scale cohort studies, are useful when some covariates violate the proportional hazards assumption or data are stratified based on factors, such as transplant centers. In Chapter IV, we extend the de-biased lasso approach proposed in Chapter III to draw inference for the stratified Cox model with potentially many covariates. We provide asymptotic results useful for inference on linear combinations of the regression parameters, and demonstrate its utility via simulation studies. We apply the method to analyze the national kidney transplantation data stratified by transplant center, and assess the effects of many factors on graft survival.

CHAPTER I

Introduction

With the advent of big data era, it becomes increasingly common that a large number of covariates are collected to study important and complex scientific problems arising from areas such as engineering, genomics, neuroimaging, and other biomedical studies. For example, in genome-wide association studies, the traditional method is typically to screen marginal associations between single nucleotide polymorphisms (SNPs) and complex traits. However, the marginal approach does not take into account the complicated structural relationships among SNPs. Jointly modeling the effects of SNPs within target genes can pinpoint functionally impactful loci in the coding regions (*Taylor et al.*, 2001; *Repapi et al.*, 2010), better understand the molecular mechanisms underlying complex diseases (*Guan and Stephens*, 2011), reduce false positives around true causal SNPs and improve prediction accuracy (*He and Lin*, 2010). In the Boston Lung Cancer Study (BLCS), which is a large cancer epidemiology cohort investigating molecular mechanisms underlying lung cancer, an analytical goal is to study the joint effects of genetic variants residing in multiple disease related pathway genes on lung cancer risk and patient survival. The results can potentially aid personalized medicine as individualized therapeutic interventions are only possible with proper characterization of relevant SNPs in pharmacogenomics (*Evans and Relling*, 2004).

Statistical methods that can tackle the challenging high-dimensionality of covariates have been increasingly popular in methodological research and real world applications over the past two decades. Variable selection is one of the most popular topics, which usually assumes that there is only a small number of important variables and concerns selecting the most relevant subset of variables to facilitate interpretation and prediction. Some well acknowledged regularization methods for variable selection include the lasso (*Tibshirani*, 1996, 1997), the elastic net (*Zou and Hastie*, 2005), the adaptive lasso (*Zou*, 2006; *Zhang and Lu*, 2007), the Dantzig selector (*Candes and Tao*, 2007; *Li et al.*, 2014) and SCAD (*Fan and Li*, 2001, 2002), among many others.

However, scientific discoveries demand solid statistical evidence based on inference, e.g. confidence interval estimation, hypothesis testing and p-values. In the presence of high-dimensional covariates, conventional methods, such as ordinary least squares, maximum likelihood estimation and maximum partial likelihood estimation, will generate biased parameter estimates and confidence intervals with poor coverage, or even no longer be feasible. Inferential methods suitable for drawing inference on high-dimensional regression models are needed.

Some recent efforts in this direction have received much attention. One stream is conditional inference based on the selected models (*Lee et al.*, 2016), which often neglects the uncertainty in model selection. Inference for the selected variables based on the asymptotic results in *Fan and Li* (2001), *Zou* (2006) and *Zhang and Lu* (2007) shares a similar flavor, and thus is super-efficient. Another main stream concerns de-biasing the lasso estimator, providing inference for every model parameter. Most existing literature on de-biasing the lasso has been developed under linear regression models (*van de Geer et al.*, 2014; *Zhang and Zhang*, 2014; *Javanmard and Montanari*, 2014), as well as some extensions, for example, to simultaneous inference (*Zhang and Cheng*, 2017; *Dezeure et al.*, 2017). Regression models for other types of outcomes, such as binary, count, ordinal and time-to-event data, are very commonly used in

real data analysis. Among the limited literature beyond linear regression (see, for example, *van de Geer et al.* 2014; *Ning and Liu* 2017; *Kong et al.* 2018; *Yu et al.* 2018; *Fang et al.* 2017), we have found that there is severe insufficient bias correction from the lasso estimator, especially for large signals, and the corresponding confidence intervals have poor coverage probabilities. Moreover, the theoretical developments for the "large p, small n" case, where the number of covariates exceeds the sample size, are heavily dependent on assumptions related to the sparsity of the inverse Fisher information matrix, which lack practical interpretation and can hardly hold in general settings beyond linear regression.

In this dissertation, we focus on the challenging high-dimensional models beyond linear regression. We scrutinize the empirical and theoretical limitations of the existing inferential methods beyond linear regression with high-dimensionality, and present methodologies, theories and real data applications based on the idea of de-biasing the lasso with an emphasis on sufficient bias correction and reliable and honest confidence regions in the "large n, diverging p" scenario, where the sample size is still larger than the number of covariates, while the latter is allowed to increase with the sample size. In particular, we consider in Chapter II the generalized linear models (GLMs) that are commonly used to model binary, count and ordinal outcomes, and the Cox proportional hazards model for right censored time-to-event outcomes in Chapter III. In the analysis of large survival studies, stratification also occurs often due, for example, to violation of proportional hazard assumption, stratum effects not being of interest or computational burden. In Chapter IV, we propose an inferential method for stratified Cox proportional hazards model.

CHAPTER II

A Revisit to De-biased Lasso for Generalized Linear Models

2.1 Introduction

It is of great interest, though with enormous challenges, to draw inference when the number of covariates grows with the sample size. When the number of covariates exceeds the sample size, the well known "large p, small n" scenario, maximum likelihood estimation (MLE) is no longer feasible and regularized variable selection methods have been developed over the decades. These include the lasso method (*Tibshirani*, 1996), the elastic net method (*Zou and Hastie*, 2005), and the Dantzig selector (*Candes and Tao*, 2007), among many others. However, these regularized methods yield biased estimates, and thus cannot be directly used for drawing statistical inference, in particular, constructing confidence intervals with a nominal coverage. Even when the number of covariates is smaller than the sample size but can increase with n, conventional methods may still not be trustworthy. *Sur and Candès* (2019) showed that MLE for high-dimensional logistic regression models can overestimate the magnitudes of non-zero effects while underestimating the variances of the estimates when the number of covariates is smaller than, but of the same order as, the sample size. We encountered the same difficulty when applying MLE to the analysis

of the Boston Lung Cancer Study (BLCS) data.

Advances to address these challenges have been made recently. One stream of methods is post-selection inference conditional on selected models (*Lee et al.*, 2016), which ignores the uncertainty associated with model selection. Other super-efficient procedures, such as SCAD (*Fan and Li*, 2001) and adaptive lasso (*Zou*, 2006), share the flavor of post-selection inference. Another school of methods is to draw inference by de-biasing the lasso estimator, termed de-biased lasso or de-sparsified lasso, which relieves the restrictions of post-selection inference and has been shown to possess nice theoretical and numerical properties in linear regression models (*van de Geer et al.* 2014; *Zhang and Zhang* 2014; *Javanmard and Montanari* 2014). When coefficients have group structures, various extensions of de-biased lasso have been proposed (*Zhang and Cheng*, 2017; *Dezeure et al.*, 2017; *Mitra and Zhang*, 2016; *Cai et al.*, 2019).

De-biased lasso has seen applications beyond linear models. For example, *van de Geer et al.* (2014) considered the de-biased lasso approach in generalized linear models (GLMs) and developed the asymptotic normality theory for each component of the coefficient estimates; *Zhang and Cheng* (2017) proposed a multiplier bootstrap procedure to draw inference on a group of coefficients in GLMs, yet without sufficient numerical evidence for the performance; *Eftekhari et al.* (2019) considered a de-biased lasso estimator for a low-dimensional component in a generalized single-index model with an unknown link function and restricted to an elliptically symmetric design.

However, in the GLM setting, our extensive simulations reveal that biases cannot be adequately removed by the existing de-biased lasso methods. Even after de-biasing, the biases are still too large relative to the model based standard errors, and the resulting confidence intervals have much lower coverage probabilities than the nominal level. Scrutiny of the existing theories points to a key assumption: the inverse of the Fisher information matrix is sparse (see *van de Geer et al.* 2014). For linear regression,

this assumption amounts to that the precision matrix for the covariates is sparse. It, however, is unlikely to hold in GLM settings, even when the precision matrix for the covariates is indeed sparse.

This begs a critical question: when can we obtain reliable inference results using de-biased lasso? Deviated from the aforementioned works which mainly focused on hypothesis testing, we are concerned with making reliable inference, such as eliminating estimation bias and obtaining good confidence interval coverage. We consider two scenarios: the "large p, small n" case where $p > n$, and the "large n, diverging p" case where p increases to infinity with n but $p/n \to 0$. In the first scenario, we discuss a key sparsity assumption in GLMs, which is likely to fail and compromise the validity of de-biased lasso. In the second scenario, we consider a natural alternative for further bias correction, by directly inverting the Hessian matrix. We study its theoretical properties and use simulations to demonstrate its advantageous performance to the competitors.

The remainder of the paper is organized as follows. Section 2.2 briefly reviews de-biased lasso in GLMs. In Section 2.3, we exemplify the performance of the original de-biased lasso estimator using simulated examples and elaborate on the theoretical limitations. In Section 2.4, under the "large n, diverging p" regime, we introduce a refined de-biased approach as an alternative to the node-wise lasso estimator for the inverse of the information matrix (*van de Geer et al.*, 2014), and establish asymptotic distributions for any linear combinations of the refined de-biased estimates. We provide simulation results and analyze the Boston Lung Cancer Study that investigates the joint associations of SNPs in nine candidate genes with lung cancer. We conclude with the summarized findings in Section 2.5. Detailed technical proofs are presented in Section 2.6

2.2 Background

2.2.1 Notation

We define commonly used notation. Denote by $\lambda_{\max}$ and $\lambda_{\min}$ the largest and the smallest eigenvalue of a symmetric matrix. For a real matrix $\mathbf{A} = (A_{ij})$, let $\|\mathbf{A}\| = [\lambda_{\max}(\mathbf{A}^T\mathbf{A})]^{1/2}$ be the spectral norm. The induced matrix ℓ_1 norm is $\|\mathbf{A}\|_1 = \max_j \sum_i |A_{ij}|$, and when $\mathbf{A}$ is symmetric, $\|\mathbf{A}\|_1 = \max_i \sum_j |A_{ij}|$. The entrywise ℓ_∞ norm is $\|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|$. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_q$ denotes the ℓ_q norm, $q \geq 1$. We write $x_n \asymp y_n$ if $x_n = \mathcal{O}(y_n)$ and $y_n = \mathcal{O}(x_n)$.

2.2.2 Generalized linear models

Denote by y_i the response variable and $\boldsymbol{x}_i = (1, \widetilde{\boldsymbol{x}}_i^T)^T \in \mathbb{R}^{p+1}$ for $i = 1, \cdots, n$, where the first element in $\boldsymbol{x}_i$ corresponds to the intercept, and the rest elements $\widetilde{\boldsymbol{x}}_i$ represent p covariates. Let $\mathbf{X}$ be an $n \times (p+1)$ covariate matrix with $\boldsymbol{x}_i^T$ being the ith row. We assume that $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ are independently and identically distributed (i.i.d.) copies of $(y, \boldsymbol{x})$. Define the negative log-likelihood function (up to a constant irrelevant to the unknown parameters) when the conditional density of y given $\boldsymbol{x}$ belongs to the linear exponential family:

$$\rho_{\boldsymbol{\xi}}(y, \boldsymbol{x}) \equiv \rho(y, \boldsymbol{x}^T\boldsymbol{\xi}) = -y\boldsymbol{x}^T\boldsymbol{\xi} + b(\boldsymbol{x}^T\boldsymbol{\xi}) \tag{2.1}$$

where $b(\cdot)$ is a known twice continuously differentiable function, $\boldsymbol{\xi} = (\beta_0, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{p+1}$ denotes the vector of regression coefficients and $\beta_0 \in \mathbb{R}$ is the intercept parameter. The unknown true coefficient vector is $\boldsymbol{\xi}^0 = (\beta_0^0, \boldsymbol{\beta}^{0T})^T$.

2.2.3 De-biased lasso

Consider the loss function $\rho_{\boldsymbol{\xi}}(y, \boldsymbol{x}) \equiv \rho(y, \boldsymbol{x}^T\boldsymbol{\xi})$ given in (2.1). Denote its first and second order derivatives with respect to $\boldsymbol{\xi}$ by $\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}}$ and $\ddot{\boldsymbol{\rho}}_{\boldsymbol{\xi}}$, respectively. For any function

$g(y, \boldsymbol{x})$, let $\mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^{n} g(y_i, \boldsymbol{x}_i)$. Then for any $\boldsymbol{\xi} \in \mathbb{R}^{p+1}$, we denote the empirical loss function based on the random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ by $\mathbb{P}_n \rho_{\boldsymbol{\xi}} \equiv \frac{1}{n}\sum_{i=1}^{n} \rho_{\boldsymbol{\xi}}(y_i, \boldsymbol{x}_i)$, and its first and second order derivatives with respect to $\boldsymbol{\xi}$ by $\mathbb{P}_n \dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_{\boldsymbol{\xi}}(y_i, \boldsymbol{x}_i)}{\partial \boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} \equiv \mathbb{P}_n \ddot{\boldsymbol{\rho}}_{\boldsymbol{\xi}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_{\boldsymbol{\xi}}(y_i, \boldsymbol{x}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T}$. Two important population-level matrices are the expectation of the Hessian matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} \equiv \mathbb{E}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} = \mathbb{E}(\mathbb{P}_n \ddot{\boldsymbol{\rho}}_{\boldsymbol{\xi}})$, and its inverse $\boldsymbol{\Theta}_{\boldsymbol{\xi}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1}$. With $\lambda > 0$, the lasso estimator for $\boldsymbol{\xi}^0$ is defined as

$$\widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}=(\beta_0, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \mathbb{P}_n \rho_{\boldsymbol{\xi}} + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.2}$$

To avoid ambiguity, we do not penalize the intercept β_0 in (2.2). The theoretical results such as prediction and ℓ_1 error bounds, however, are the same as those in *van de Geer* (2008) and *van de Geer et al.* (2014) where all the coefficients are penalized (*Bühlmann and van de Geer*, 2011). *van de Geer et al.* (2014) applied the node-wise lasso method to obtain an estimator $\widehat{\boldsymbol{\Theta}}$ for $\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}$, and proposed a de-biased lasso estimator for ξ_j^0 with:

$$\widehat{b}_j \equiv \widehat{\xi}_j - \widehat{\boldsymbol{\Theta}}_j \mathbb{P}_n \dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}},$$

where $\hat{\sigma}_j \equiv \sqrt{\widehat{\boldsymbol{\Theta}}_j \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} \widehat{\boldsymbol{\Theta}}_j^T / n}$ is the model based standard error for $\widehat{b}_j$. Here, $\widehat{\boldsymbol{\Theta}}_j$ is the jth row of $\widehat{\boldsymbol{\Theta}}$.

2.3 The "large p, small n" scenario

Even though the asymptotic theory has been developed for the "large p, small n" scenario (*van de Geer et al.*, 2014), we examine why de-biased lasso performs unsatisfactorily in GLMs.

2.3.1 A simulation study

We present a simulation study that features a logistic regression model with $n = 300$ observations and $p = 500$ covariates. For simplicity, covariates are simulated from $N_p(\mathbf{0}, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_{x,ij} = 0.7^{|i-j|}$, and truncated at ± 6. In the true coefficient vector $\boldsymbol{\beta}^0$, 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 2, 4 or 10 additional coefficients from the rest in $\boldsymbol{\beta}^0$, 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.

Figure 2.1, 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 almost linearly increasing with the true size of β_1^0. This undermines the credibility of the consequent confidence intervals. Meanwhile, the model-based variance does not approximate the true variance well, overestimating the variance for smaller signals and underestimating for larger ones in the two smaller models, as shown by the top two rows in Figure 2.1. This partially explains the over- and under-coverage for smaller and larger signals, respectively. Due to penalized estimation in $\widehat{\mathbf{\Theta}}$, the variance of the 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 Figure 2.1. 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 2.1. 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.

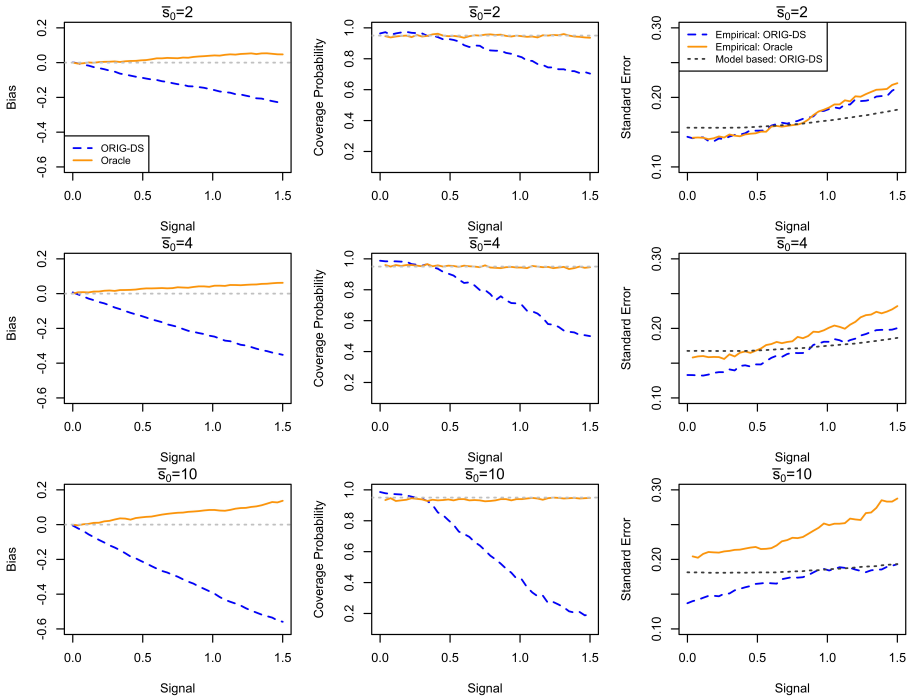

Figure 2.1: Simulation results of logistic regression with sample size $n = 300$ and $p = 500$ covariates. Covariates are first generated from multivariate Gaussian distribution with mean zero, AR(1) covariance structure and correlation 0.7, and truncated at ± 6. Each row presents estimation bias, empirical coverage probability and standard error (both model-based and empirical) of the estimated β_1^0, with 2, 4 and 10 additional signals fixed at 1 from the top to the bottom, respectively. "ORIG-DS" and "Oracle" stand for the original de-biased lasso estimator and the oracle estimator as if the true model were known, respectively.

2.3.2 Reflections on the validity of theoretical assumptions

van de Geer et al. (2014) established the asymptotic properties of the de-biased lasso estimator in GLMs under certain regularity conditions (see Section 3 of *van de Geer et al.* 2014), which are imposed to regularize the behavior of the lasso estimator $\widehat{\boldsymbol{\xi}}$ and the estimated matrix $\widehat{\boldsymbol{\Theta}}$. *van de Geer et al.* (2014) employed the node-wise lasso estimator for $\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}$, which was originally proposed by *Meinshausen and Bühlmann* (2006) for covariance selection in high-dimensional graphs.

We now revisit the de-biased lasso estimator and its decomposition. The first

order Taylor expansion of $\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}$ at $\widehat{\boldsymbol{\xi}}$ gives

$$\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0} = \mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}} + \mathbb{P}_n\ddot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}}(\boldsymbol{\xi}^0 - \widehat{\boldsymbol{\xi}}) + \boldsymbol{\Delta}, \tag{2.3}$$

where $\boldsymbol{\Delta}$ is a $(p+1)$-dimensional vector of remainder terms with its jth element

$$\Delta_j = \frac{1}{n}\sum_{i=1}^{n}\left(\ddot{\rho}(y_i, a_j^*) - \ddot{\rho}(y_i, \boldsymbol{x}_i^T\widehat{\boldsymbol{\xi}})\right)x_{ij}\boldsymbol{x}_i^T(\boldsymbol{\xi}^0 - \widehat{\boldsymbol{\xi}}), \tag{2.4}$$

in which $\ddot{\rho}(y,a) \equiv \dfrac{\partial^2\rho(y,a)}{\partial a^2}$, and a_j^* lies between $\boldsymbol{x}_i^T\widehat{\boldsymbol{\xi}}$ and $\boldsymbol{x}_i^T\boldsymbol{\xi}^0$. It follows that $\boldsymbol{\Delta} = \mathbf{0}$ in linear regression models, but generally non-zero in GLMs. Multiplying both sides of (2.3) by $\widehat{\boldsymbol{\Theta}}_j$ and re-organizing the terms, we obtain the following equality for the jth component

$$\left[\widehat{\xi}_j + \overbrace{\left(-\widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}}\right)}^{I_j} + \overbrace{\left(-\widehat{\boldsymbol{\Theta}}_j\boldsymbol{\Delta}\right)}^{II_j} + \overbrace{\left(\widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\ddot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{e}_j^T\right)\left(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\right)}^{III_j}\right] - \xi_j^0 = -\widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}, \tag{2.5}$$

where $\boldsymbol{e}_j$ is a $(p+1)$-dimensional vector with the jth element being 1 and 0 elsewhere. We define three terms

$$I_j = -\widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}},\ II_j = -\widehat{\boldsymbol{\Theta}}_j\boldsymbol{\Delta},\ \text{and}\ III_j = \left(\widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\ddot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{e}_j^T\right)\left(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\right).$$

They are crucial in studying the bias behavior of the de-biased lasso estimator that can be alternatively expressed as $\widehat{b}_j = \widehat{\xi}_j + I_j$. According to (2.5), as long as $\sqrt{n}\ II_j/\hat{\sigma}_j = o_P(1)$, $\sqrt{n}\ III_j/\hat{\sigma}_j = o_P(1)$, and $\sqrt{n}\ \widehat{\boldsymbol{\Theta}}_j\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}/\hat{\sigma}_j$ is asymptotically normal, the asymptotic normality of $\sqrt{n}\left(\widehat{b}_j - \xi_j^0\right)/\hat{\sigma}_j$ follows directly.

The de-biased lasso approach requires an appropriate inverse matrix estimator with $\mathcal{O}(p^2)$ unknown parameters. In the "large p, small n" scenario, where the number of covariates can be as large as $o(\exp(n^a))$ for some $a > 0$, the $(p+1)\times(p+1)$ inverse

information matrix is not estimable without further assumptions on the structure of $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$. This inevitably needs regularization, and ℓ_1-type regularization is often adopted due to its theoretical readiness. An important assumption on $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ in *van de Geer et al.* (2014) is the ℓ_0 sparsity, i.e. the number of non-zero elements of each row in $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ is small. This assumption is vital for the consistency of $\widehat{\mathbf{\Theta}}_j$ to $\mathbf{\Theta}_{\boldsymbol{\xi}^0,j}$ and consequently the model-based variance, and impacts the negligibility of term III_j in (2.5). In particular, the third bias term in (2.5) III_j is non-negligible if the convergence rate of $\widehat{\mathbf{\Theta}}_j$ to $\mathbf{\Theta}_{\boldsymbol{\xi}^0,j}$, which depends on the ℓ_0 sparsity of the row vector $\mathbf{\Theta}_{\boldsymbol{\xi}^0,j}$ using the node-wise lasso estimation, is not fast enough.

However, these sparsity assumptions have not been clarified in the existing literature, except for linear regression models. In a linear regression model, $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ is the precision matrix for covariates which is free of $\boldsymbol{\xi}^0$, and for multivariate Gaussian covariates, a zero element of $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ implies conditional independence between corresponding covariates. In contrast, the row sparsity assumption on $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ does not have a clear interpretation in GLMs, and may not be valid as it depends on the unknown $\boldsymbol{\xi}^0$. In the information matrix $\mathbf{\Sigma}_{\boldsymbol{\xi}^0}$, its (j,k)-th element is $\mathbb{E}\left[x_{ij}x_{ik}\ddot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)\right] = \mathbb{E}\left[x_{ij}x_{ik}\ddot{b}(\boldsymbol{x}_i^T\boldsymbol{\xi}^0)\right]$. In the most extreme case where all covariates are independent with mean zero, $\mathbf{\Sigma}_{\boldsymbol{\xi}^0,jk} = 0$ for $j \neq k, j = 2, \cdots, p+1$, $k \in \{k : 2 \leq k \leq p+1, \xi_k^0 = 0\}$, and then $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ is sparse if the true model $\{j : 1 \leq j \leq p,\ \beta_j^0 \neq 0\}$ is small. With covariates generally correlated, it is unconceivable that most off-diagonal elements in $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ are zero, because $\ddot{b}(\boldsymbol{x}_i^T\boldsymbol{\xi}^0) = \ddot{b}(\beta_0^0 + \widetilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}^0)$ also depends on the covariates $\widetilde{\boldsymbol{x}}_i$ in a GLM, even when the precision matrix for $\widetilde{\boldsymbol{x}}_i$ is sparse *per se*. This makes the sparsity assumption for $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ obscure in GLMs. To see this, consider the Poisson regression, which has a closed-form expression for $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$. Assume the covariates $\widetilde{\boldsymbol{x}}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma}_x)$ and the mean response conditional on $\widetilde{\boldsymbol{x}}_i$ is $\mu_i = \exp\{\beta_0^0 + \widetilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}^0\}$ under the canonical link. Then,

we have

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0} = \exp\left\{\beta_0^0 + \frac{1}{2}\boldsymbol{\beta}^{0T}\boldsymbol{\Sigma}_x\boldsymbol{\beta}^0\right\}\begin{pmatrix} 1 & \boldsymbol{\beta}^{0T}\boldsymbol{\Sigma}_x \\ \boldsymbol{\Sigma}_x\boldsymbol{\beta}^0 & \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_x\boldsymbol{\beta}^0\boldsymbol{\beta}^{0T}\boldsymbol{\Sigma}_x \end{pmatrix}$$

and

$$\boldsymbol{\Theta}_{\boldsymbol{\xi}^0} = \exp\left\{-\beta_0^0 - \frac{1}{2}\boldsymbol{\beta}^{0T}\boldsymbol{\Sigma}_x\boldsymbol{\beta}^0\right\}\begin{pmatrix} \frac{1}{c} & -\frac{1}{c}\boldsymbol{a}^T\mathbf{A}^{-1} \\ -\frac{1}{c}\mathbf{A}^{-1}\boldsymbol{a} & \mathbf{A}^{-1} + \frac{1}{c}\mathbf{A}^{-1}\boldsymbol{a}\boldsymbol{a}^T\mathbf{A}^{-1} \end{pmatrix},$$

where $\mathbf{A} = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_x\boldsymbol{\beta}^0\boldsymbol{\beta}^{0T}\boldsymbol{\Sigma}_x$, $\boldsymbol{a} = \boldsymbol{\Sigma}_x\boldsymbol{\beta}^0$ and $c = 1 - \boldsymbol{\beta}^{0T}(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{\beta}^0\boldsymbol{\beta}^{0T})^{-1}\boldsymbol{\beta}^0$. In an over-simplified case where covariates are independent ($\boldsymbol{\Sigma}_x = \mathbf{I}_p$) and $\boldsymbol{\beta}^0$ is sparse, $\mathbf{A}^{-1} + \frac{1}{c}\mathbf{A}^{-1}\boldsymbol{a}\boldsymbol{a}^T\mathbf{A}^{-1}$ can be a sparse matrix. However, with often complicated correlation structures between covariates, signal positions and strengths in $\boldsymbol{\beta}^0$, it is difficult to guarantee that $\mathbf{A}^{-1} + \frac{1}{c}\mathbf{A}^{-1}\boldsymbol{a}\boldsymbol{a}^T\mathbf{A}^{-1}$ is sparse.

To summarize, we believe that the sparsity assumption imposed on $\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}$ plays an extremely important role in obtaining the desirable asymptotic properties and finite sample performance of de-biased lasso in GLMs. However, such an assumption is hardly justifiable in a GLM setting. As evidenced by our simulations, the gap between theory and practice likely explains the problematic performance of de-biased lasso in the "large p, small n" scenario. Also note that both bias terms II_j and III_j are not even computable and cannot be recovered, because they involve the unknown $\boldsymbol{\xi}^0$. All point to that de-biased lasso generally does not work well in GLMs in the "large p, small n" scenario.

2.4 The "large n, diverging p" scenario

We next study de-biased lasso in GLMs when $p < n$ but p diverges to infinity with n by eliminating more biases, where, under certain conditions, the Hessian matrix is

invertible with probability going to one. Therefore, directly inverting the Hessian matrix serves as a natural alternative to the node-wise lasso for $\widehat{\boldsymbol{\Theta}}$. In the following, we study the properties of this alternative estimator. Denote $\widetilde{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}$ to distinguish it from the node-wise lasso estimator $\widehat{\boldsymbol{\Theta}}$. Similarly, $\widetilde{\boldsymbol{\Theta}}_j$ represents the jth row of $\widetilde{\boldsymbol{\Theta}}$.

Similar to (2.5), we have the following equality using $\widetilde{\boldsymbol{\Theta}}$:

$$\left[\widehat{\boldsymbol{\xi}} + \left(-\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}}\right) + \left(-\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}\right) + \left(\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\ddot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}} - \mathbf{I}\right)\left(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\right)\right] - \boldsymbol{\xi}^0 = -\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}. \quad (2.6)$$

With $\widetilde{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}$, the new term III_j in (2.6) equals 0 for all j, which is no longer a source of bias compared to the original de-biased lasso. Then (2.6) becomes

$$\left[\widehat{\boldsymbol{\xi}} + \left(-\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}}\right) + \left(-\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}\right)\right] - \boldsymbol{\xi}^0 = -\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}. \quad (2.7)$$

The new de-biased lasso estimator based on $\widetilde{\boldsymbol{\Theta}}$ is

$$\widetilde{\boldsymbol{b}} \equiv \widehat{\boldsymbol{\xi}} - \widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\xi}}},$$

which is designed to further correct biases compared to the original de-biased estimator. We will show that any linear combinations of $\widetilde{\boldsymbol{b}}$, including each coefficient estimate as a special case, are asymptotically normally distributed.

2.4.1 Theoretical results

Without loss of generality, we assume that each covariate has been standardized to have mean zero and variance 1. Let s_0 denote the number of non-zero elements in $\boldsymbol{\xi}^0$. Let $\mathbf{X}_{\boldsymbol{\xi}} = \boldsymbol{W}_{\boldsymbol{\xi}}\mathbf{X}$ be the weighted design matrix, where $\boldsymbol{W}_{\boldsymbol{\xi}}$ is a diagonal matrix with elements $\omega_i(\boldsymbol{\xi}) = \sqrt{\ddot{\rho}(y_i, x_i^T\boldsymbol{\xi})}$, $i = 1, \cdots, n$. Recall that for any $\boldsymbol{\xi} \in \mathbb{R}^{p+1}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} = \mathbf{X}_{\boldsymbol{\xi}}^T\mathbf{X}_{\boldsymbol{\xi}}/n$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbb{E}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}})$. The ψ_2-norms (see *Vershynin* 2010) introduced below are useful for characterizing the convergence rate of $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}$. For a random variable

Z, its ψ_2-norm is defined as

$$\|Z\|_{\psi_2} = \sup_{r \geq 1} r^{-1/2} (\mathbb{E}|Z|^r)^{1/r}.$$

We call Z a sub-Gaussian random variable if $\|Z\|_{\psi_2} \leq M < \infty$ for a constant $M > 0$. For a random vector $\boldsymbol{Z}$, its ψ_2-norm is defined as

$$\|\boldsymbol{Z}\|_{\psi_2} = \sup_{\|\boldsymbol{a}\|_2 = 1} \|\langle \boldsymbol{Z}, \boldsymbol{a} \rangle\|_{\psi_2}.$$

A random vector $\boldsymbol{Z} \in \mathbb{R}^{p+1}$ is called sub-Gaussian if the inner product $\langle \boldsymbol{Z}, \boldsymbol{a} \rangle$ is sub-Gaussian for all $\boldsymbol{a} \in \mathbb{R}^{p+1}$. Let $L_p = ||\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 \omega_1(\boldsymbol{\xi}^0)||_{\psi_2}$, which characterizes the probabilistic tail behavior of the weighted covariates. We make the following assumptions.

(C1) The elements in $\mathbf{X}$ are bounded, i.e. there exists a constant $K > 0$ such that $\|\mathbf{X}\|_\infty \leq K$.

(C2) $\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}$ is positive definite and its eigenvalues are bounded and bounded away from 0, i.e. there exist two absolute constants $c_{\min}$ and $c_{\max}$ such that $0 < c_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) \leq c_{\max} < \infty$.

(C3) The derivatives $\dot{\rho}(y, a) \equiv \dfrac{\partial}{\partial a} \rho(y, a)$ and $\ddot{\rho}(y, a) = \dfrac{\partial^2}{\partial a^2} \rho(y, a)$ exist for all (y, a). For some δ-neighborhood ($\delta > 0$), $\ddot{\rho}(y, a)$ is Lipschitz such that for some absolute constant $c_{Lip} > 0$,

$$\max_{a_0 \in \{\boldsymbol{x}_i^T \boldsymbol{\xi}^0\}} \sup_{|a - a_0| \vee |\widehat{a} - a_0| \leq \delta} \sup_{y \in \mathcal{Y}} \frac{|\ddot{\rho}(y, a) - \ddot{\rho}(y, \widehat{a})|}{|a - \widehat{a}|} \leq c_{Lip}.$$

The derivatives are bounded in the sense that there exist two constants $K_1, K_2 >$

0 such that

$$\max_{a_0 \in \{\boldsymbol{x}_i^T \boldsymbol{\xi}^0\}} \sup_{y \in \mathcal{Y}} |\dot{\rho}(y, a_0)| \leq K_1,$$

$$\max_{a_0 \in \{\boldsymbol{x}_i^T \boldsymbol{\xi}^0\}} \sup_{|a - a_0| \leq \delta} \sup_{y \in \mathcal{Y}} |\ddot{\rho}(y, a)| \leq K_2.$$

(C4) $\|\mathbf{X}\boldsymbol{\xi}^0\|_\infty$ is bounded.

(C5) The matrix $\mathbb{E}(\mathbf{X}^T\mathbf{X}/n)$ is positive definite and its eigenvalues are bounded and bounded away from 0.

It is common to assume bounded covariates as in (C1) and bounded eigenvalues of the information matrix as in (C2) in high-dimensional inference literature (*van de Geer et al.*, 2014; *Ning and Liu*, 2017). $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are sub-Gaussian random vectors under (C1), but we do not impose a boundedness assumption on their ψ_2-norm, which may depend on p (*Vershynin*, 2010, 2012). (C2) refers to a compatibility condition that is sufficient to derive the rate of convergence for $\widehat{\boldsymbol{\xi}}$. (C3) assumes local properties of the derivatives of the general loss $\rho(y, \boldsymbol{x}^T\xi)$ (*van de Geer et al.*, 2014). (C4) is commonly assumed (*van de Geer et al.*, 2014; *Ning and Liu*, 2017) and ensures the quadratic margin behavior of the excess risk and is useful to obtain the rate for $\|\mathbf{X}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0)\|_2^2/n$ (*Bühlmann and van de Geer*, 2011). (C5) is a mild requirement in high-dimensional regression analysis with random designs. A similar condition can be found in *Wang* (2011).

Theorem II.1 establishes the asymptotic normality result for any linear combinations of $\widetilde{\boldsymbol{b}}$, based on which inference can be drawn. The proof is given in Section 2.6, as well as useful lemmas.

Theorem II.1. *Assume that* $L_p^4 \dfrac{p^2 \log p}{n} \to 0$, $\sqrt{p \log(p)} s_0 \lambda \to 0$, *and* $\sqrt{np} s_0 \lambda^2 \to 0$ *as* $n \to \infty$. *Let* $\widetilde{\boldsymbol{b}} = \widehat{\boldsymbol{\xi}} - \widetilde{\boldsymbol{\Theta}} \mathbb{P}_n \dot{\rho}_{\widehat{\boldsymbol{\xi}}}$ *and* $\boldsymbol{\alpha}_n \in \mathbb{R}^{p+1}$ *with* $||\boldsymbol{\alpha}_n||_2 = 1$. *Under (C1) - (C5), we have*

$$\frac{\sqrt{n}\boldsymbol{\alpha}_n^T(\widetilde{\boldsymbol{b}} - \boldsymbol{\xi}^0)}{\sqrt{\boldsymbol{\alpha}_n^T \widetilde{\boldsymbol{\Theta}} \boldsymbol{\alpha}_n}} \xrightarrow{d} N(0, 1).$$

From Theorem II.1, one can construct $100 \times (1-r)$th confidence intervals for $\boldsymbol{\alpha}_n^T \boldsymbol{\xi}^0$ as

$$\left[\boldsymbol{\alpha}_n^T \boldsymbol{\xi}^0 - z_{r/2} \sqrt{\boldsymbol{\alpha}_n^T \widetilde{\boldsymbol{\Theta}} \boldsymbol{\alpha}_n / n}, \boldsymbol{\alpha}_n^T \boldsymbol{\xi}^0 + z_{r/2} \sqrt{\boldsymbol{\alpha}_n^T \widetilde{\boldsymbol{\Theta}} \boldsymbol{\alpha}_n / n} \right],$$

where $z_{r/2}$ is the upper $(r/2)$th quantile of the standard normal distribution.

Remark II.2. For the lasso approach, $\lambda \asymp \sqrt{\log(p)/n}$, we then only need $L_p^2 \sqrt{\frac{p^2 \log p}{n}} \to 0$ and $\sqrt{np} s_0 \lambda^2 \to 0$ as $n \to \infty$, because $\sqrt{p \log(p)} s_0 \lambda \to 0$ and $\sqrt{np} s_0 \lambda^2 \to 0$ are equivalent.

Remark II.3. Theorem II.1 reveals that the required rate for p relative to n depends on the factor L_p and can be further simplified. The dependence on L_p results from that the convergence rate of $\widetilde{\boldsymbol{\Theta}}$ is related to $L_p = \| \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 \omega_1(\boldsymbol{\xi}^0) \|_{\psi_2}$. In *Javanmard and Montanari* (2014) for linear models and *Ning and Liu* (2017) for GLMs, L_p is assumed to be a constant irrelevant to p. When covariates follow a multivariate Gaussian distribution in a linear model, $L_p = \mathcal{O}(1)$ holds, then it only requires that $\frac{p^2 \log p}{n} \to 0$. However, in general, L_p may grow with p, and it can be shown that the utmost bound $L_p = \mathcal{O}(\sqrt{p})$. Specifically, by definition, $L_p = \| \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 \omega_1(\boldsymbol{\xi}^0) \|_{\psi_2} = \sup_{\boldsymbol{z} \in B^{p+1}} \| \langle \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 w_1(\boldsymbol{\xi}^0), \boldsymbol{z} \rangle \|_{\psi_2}$, where B^{p+1} is the unit ball in $\mathbb{R}^{p+1}$. Then we have

$$\begin{aligned} |\langle \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 w_1(\boldsymbol{\xi}^0), \boldsymbol{z} \rangle| &\leq \|\boldsymbol{z}\|_2 \cdot \| \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 w_1(\boldsymbol{\xi}^0) \|_2 \\ &\leq \| \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \| \cdot \| \boldsymbol{x}_1 w_1(\boldsymbol{\xi}^0) \|_2 \\ &\leq c_{\min}^{-\frac{1}{2}} \sqrt{K_2(p+1)} K. \end{aligned}$$

Therefore, $L_p \leq c_{\min}^{-\frac{1}{2}} \sqrt{K_2(p+1)} K$. This results in the most stringent rate requirement $\frac{p^4 \log p}{n} \to 0$, implying $\sqrt{np} s_0 \lambda^2 = o(1)$ when $\lambda \asymp \sqrt{\log(p)/n}$.

Remark II.4. In Theorem II.1, p is assumed to grow slowly with n so that $p \ll n$. This assumption is not uncommon in the literature. *Fan and Peng* (2004) assumed $p^5/n \to 0$ for a non-concave penalized maximum likelihood estimator to establish the

oracle property and the asymptotic normality for selected variables. Yet the estimates in *Fan and Peng* (2004) are super-efficient, which is not our focus. Without parameter regularization, *Wang* (2011) assumed $p^3/n \to 0$ to derive asymptotic normality for the solutions to generalized estimating equations with binary outcomes and clustered data, which reduces to the usual logistic regression when simplified to a singleton in each cluster. *Wang* (2011) studied a fixed design case, and proved the asymptotic normality for a different quantity $\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})^{-1/2} \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})$; see Theorem 3.8 in *Wang* (2011). When p/n is not negligible (e.g. > 0.1), simulations show that MLE yields biased and highly variable estimates, and is outperformed by our proposed $\widetilde{b}$.

2.4.2 Simulation results

We investigate the performance of our alternative de-biased estimator $\widetilde{\boldsymbol{b}}$ in the "large n, diverging p" scenario, and focus on biases in estimates and coverage probabilities of confidence intervals. The estimators in comparison are

(i) the original de-biased lasso estimator $\widehat{b}_j$ obtained by using the node-wise lasso estimator $\widehat{\boldsymbol{\Theta}}$ in *van de Geer et al.* (2014) (*ORIG-DS*);

(ii) the refined de-biased lasso approach based on the inverse matrix estimation $\widetilde{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}$, $\widetilde{b}_j$, as described in this section (*REF-DS*);

(iii) the conventional MLE (*MLE*).

As simulations using logistic and Poisson regression models yield similar results, we only report those from logistic regression. A total of $n = 1,000$ observations and $p = 40, 100, 300, 400$ covariates are simulated. We assume that in $\boldsymbol{x}_i = (1, \widetilde{\boldsymbol{x}}_i^T)^T$, $\widetilde{\boldsymbol{x}}_i$ are independently generated from $N_p(\mathbf{0}_p, \boldsymbol{\Sigma}_x)$ then truncated at ± 6, and $y_i|\boldsymbol{x}_i \sim Bernoulli(\mu_i)$, where $\mu_i \equiv \exp(\boldsymbol{x}_i^T \boldsymbol{\xi}^0)/\{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\xi}^0)\}$. The intercept $\beta_0^0 = 0$, and β_1^0 varies from 0 to 1.5 with 40 equally spaced increments. Four additional arbitrarily chosen elements of $\boldsymbol{\beta}^0$ take non-zero values, two at 0.5 and the other two at 1, and

then are fixed throughout the simulation. In some settings, *MLE* estimates do not exist due to divergence and thus are not shown. The covariance matrix $\mathbf{\Sigma}_x$ for $\widetilde{\boldsymbol{x}}_i$ takes one of the following three forms: identity matrix, AR(1) with correlation $\rho = 0.7$, and compound symmetry with correlation $\rho = 0.7$. The tuning parameter in the ℓ_1-norm penalized regression is selected by 10-fold cross-validation, and the tuning parameter for the node-wise lasso estimator $\widehat{\mathbf{\Theta}}$ is selected using 5-fold cross-validation. Both tuning parameter selection procedures are implemented using `glmnet` (*Friedman et al.*, 2010). For every β_1^0 value, we summarize the average bias, empirical coverage probability, empirical standard error and model-based estimated standard error over 200 replications.

Figure 2.2 presents the simulation results for estimating β_1^0 under the AR(1) covariance structure. The three methods in comparison behave similarly when only 40 covariates are present, with *MLE* showing slightly larger biases for larger signals. *MLE* displays much more biases than the other two methods when 100 covariates are present, and does not always exist in some settings as the number of covariates increases. When *MLE* does exist, it shows more variability than *ORIG-DS* and *REF-DS*, and lower coverage probabilities. There is a systematic bias in *ORIG-DS*, which increases with the signal strength of β_1^0. For large signals, the model-based standard error of *ORIG-DS* slightly underestimates the true variability. These factors contribute to the poor coverage probabilities of *ORIG-DS* when signal size is not too close to zero. Among all the competing methods, *REF-DS* presents the smallest biases and has an empirical coverage probability closest to the nominal level across different settings, though *REF-DS* exhibits slightly higher variability than *ORIG-DS*. This is possibly because *REF-DS* does not utilize penalization when inverting the matrix. Under the null $\beta_1^0 = 0$, both *ORIG-DS* and *REF-DS* have coverage probabilities close to 95% and preserve the type 1 error.

Figure 2.3, in the independent covariate case, shows similar patterns to the case

when Σ_x is AR(1) with variance 1 and correlation $\rho = 0.7$. The model-based standard errors estimated by *ORIG-DS* for large signal values are, in most cases, even smaller than those by *Oracle* when $p = 300$ and 400, since $\widehat{\mathbf{\Theta}}$ is estimated using penalized regression and the resulting variance estimates tend to be biased downward. When each pair of covariates has the same correlation $\rho = 0.7$ in the compound symmetry case, estimation biases from *ORIG-DS* persist in Figure 2.4, and the seemingly improved coverage probabilities, especially for *ORIG-DS*, are due in part to higher variability.

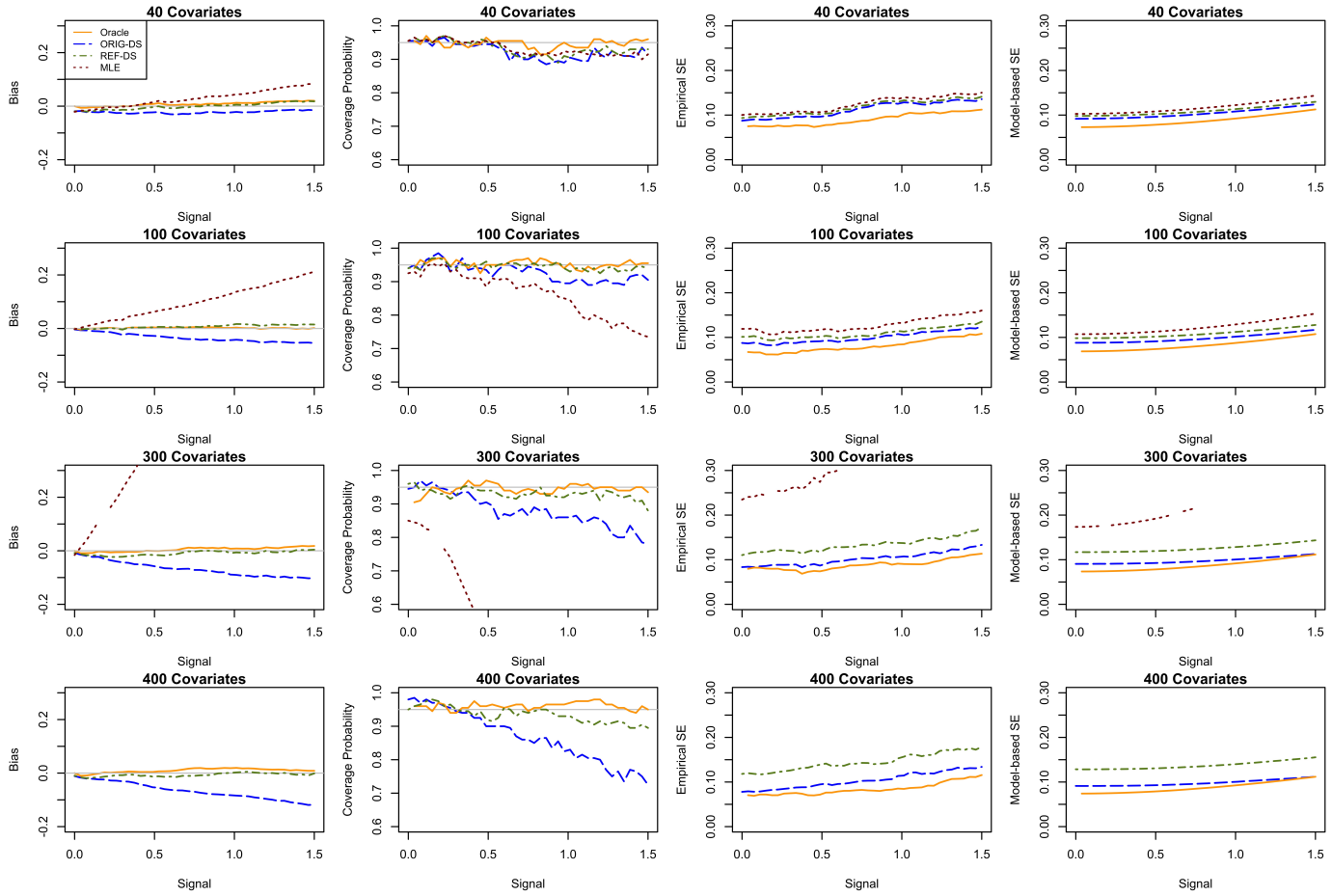

Figure 2.2: Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being AR(1) with $\rho = 0.7$. The sample size is $n = 1,000$ and the number of covariates $p = 40, 100, 300, 400$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

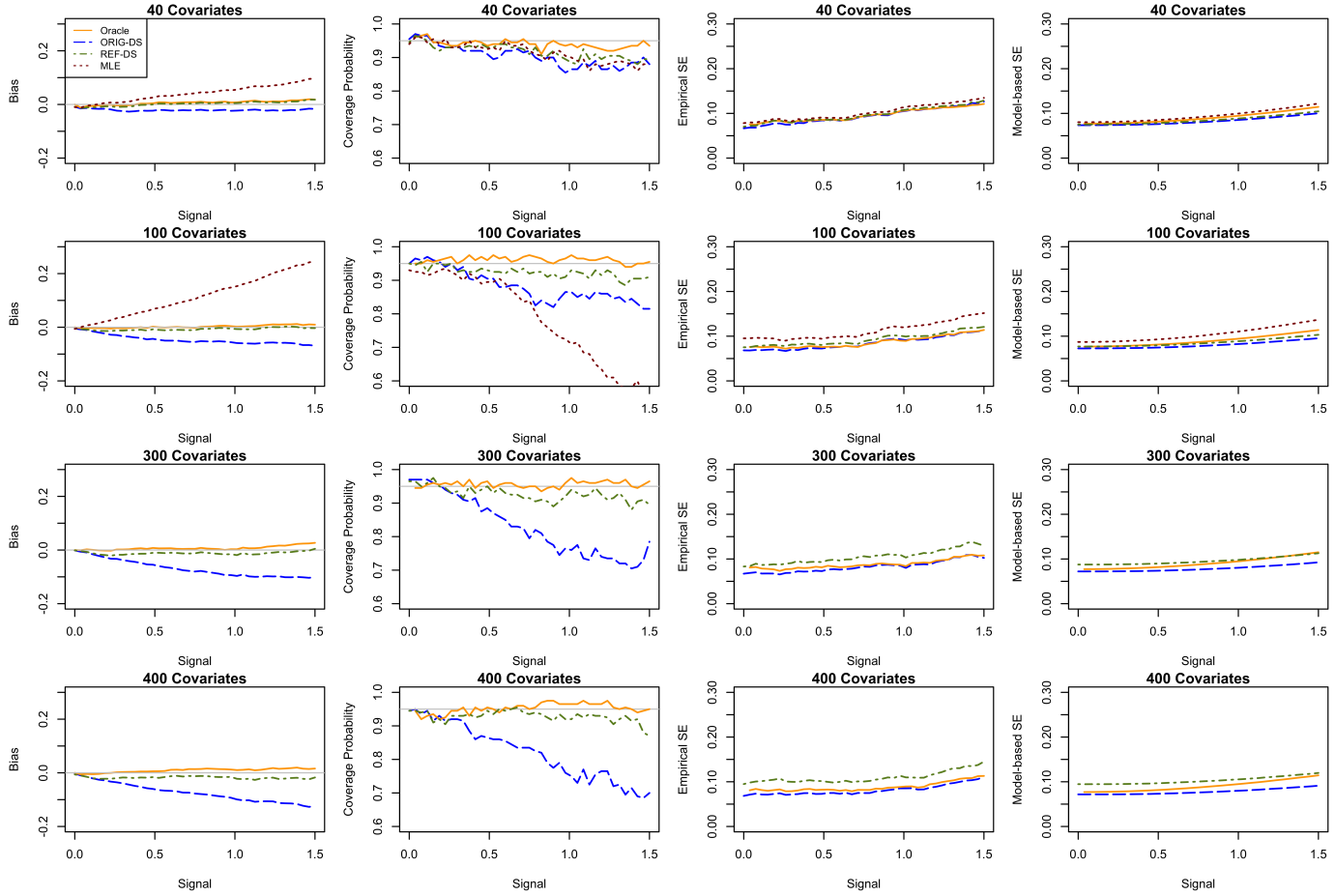

Figure 2.3: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with $\Sigma_x = I_p$. The sample size is $n = 1,000$ and the number of covariates $p = 40, 100, 300, 400$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

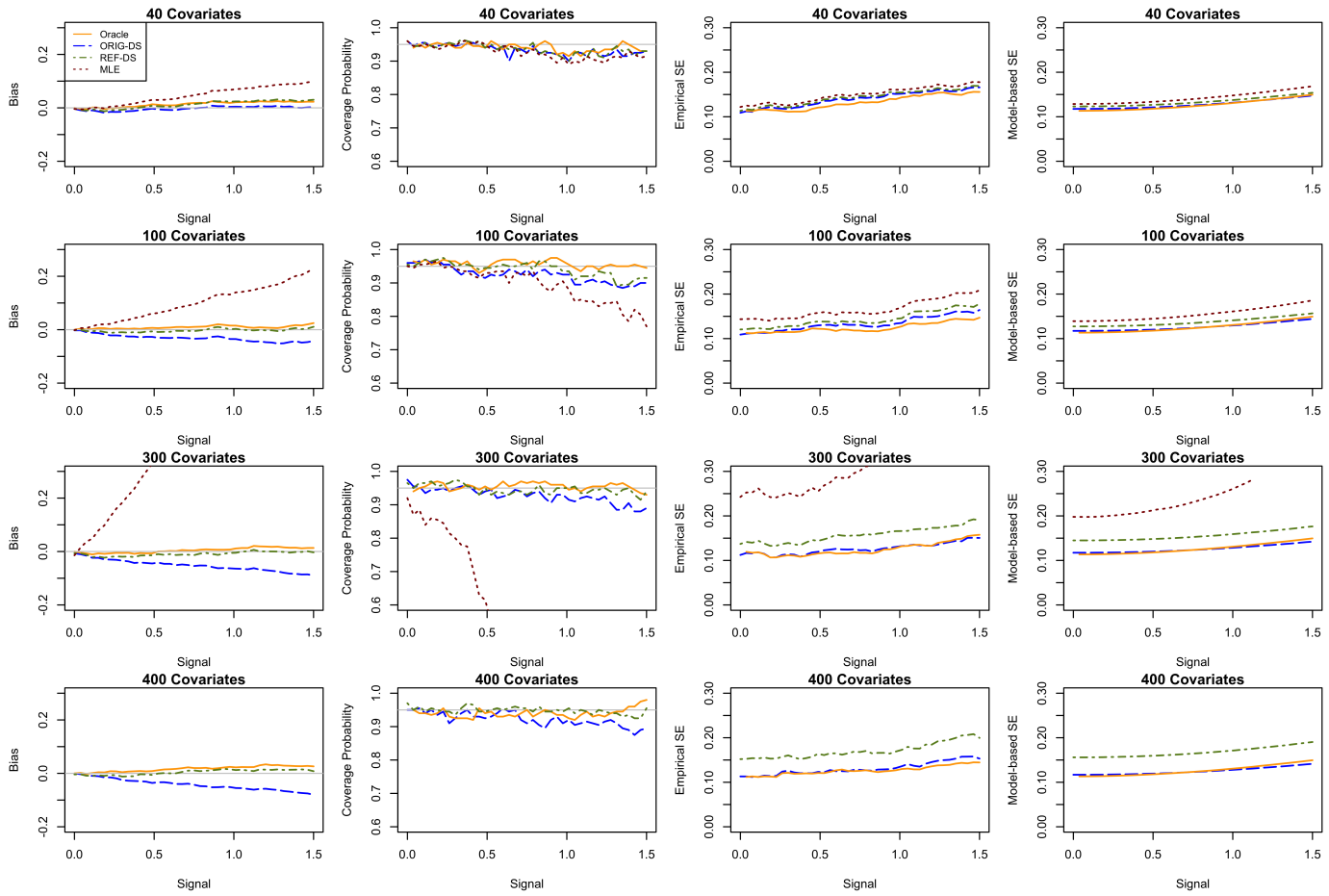

Figure 2.4: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being compound symmetry with $\rho = 0.7$. The sample size is $n = 1,000$ and the number of covariates $p = 40, 100, 300, 400$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

Additional simulation results are provided under the same simulation setups to those with $n = 1,000$, except with a smaller sample size of $n = 500$ and $p = 20, 100, 200, 300$ covariates in the logistic regression model. Figures 2.5 - 2.7 display the results from three types of covariance structures. Figure 2.7 shows that when $n = 500$, $p = 300$, neither de-biased lasso methods can work well in this difficult setup, which is not surprising given the relatively large p/n ratio and high correlation between each pair of the covariates ($\rho = 0.7$). We also varied the correlation to $\rho = 0.2$ in the covariance matrix $\mathbf{\Sigma}_x$ for AR(1) and compound symmetry structures to reflect the smaller correlation among covariates; see Figure 2.8 and Figure 2.9, respectively. These results are similar to the independent covariate case, despite that each covariate is correlated with some or all other covariates to a non-negligible extent. To summarize, *REF-DS*, in most cases, can provide the best bias correction and honest confidence intervals.

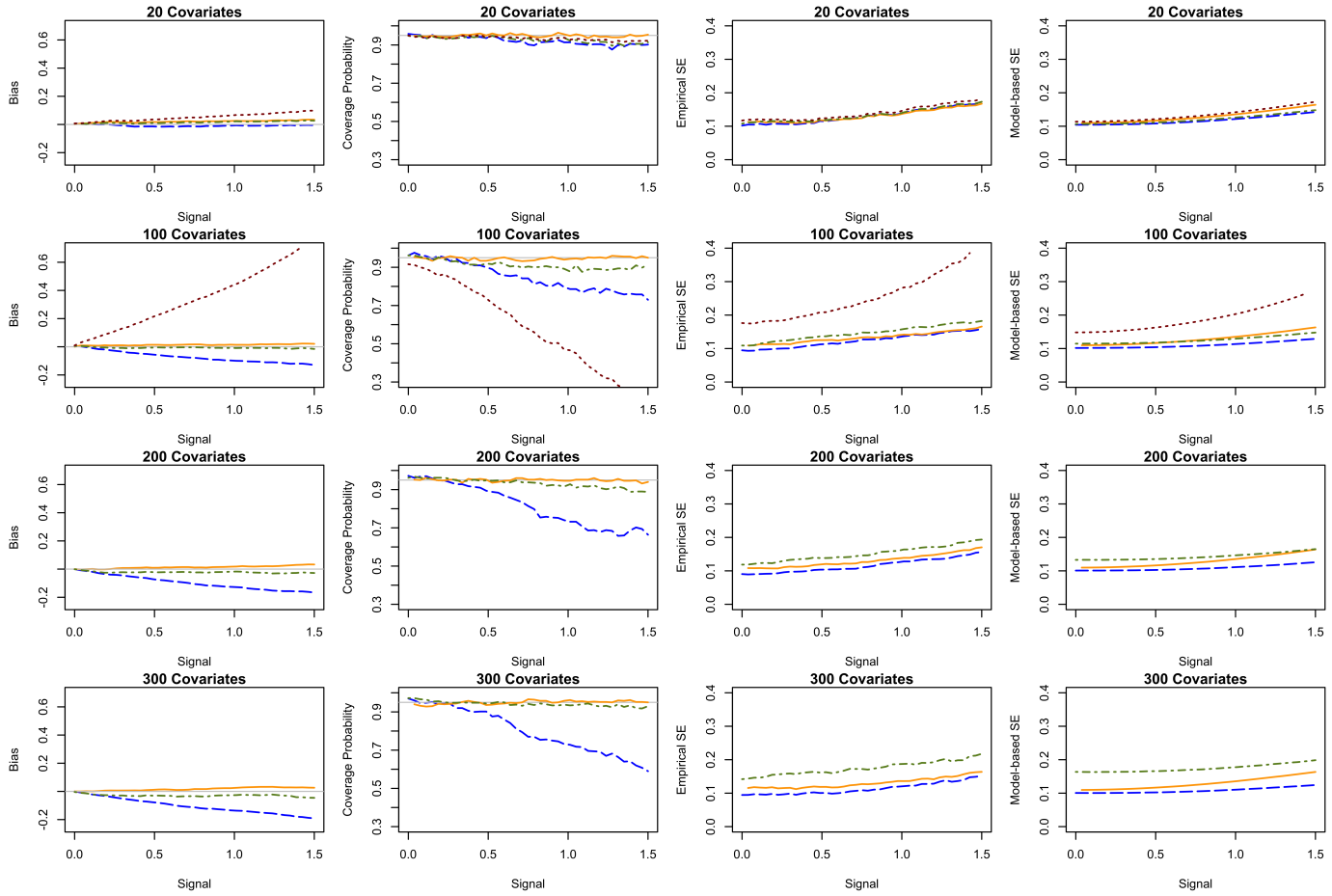

Figure 2.5: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with $\Sigma_x = I$, the identity matrix. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

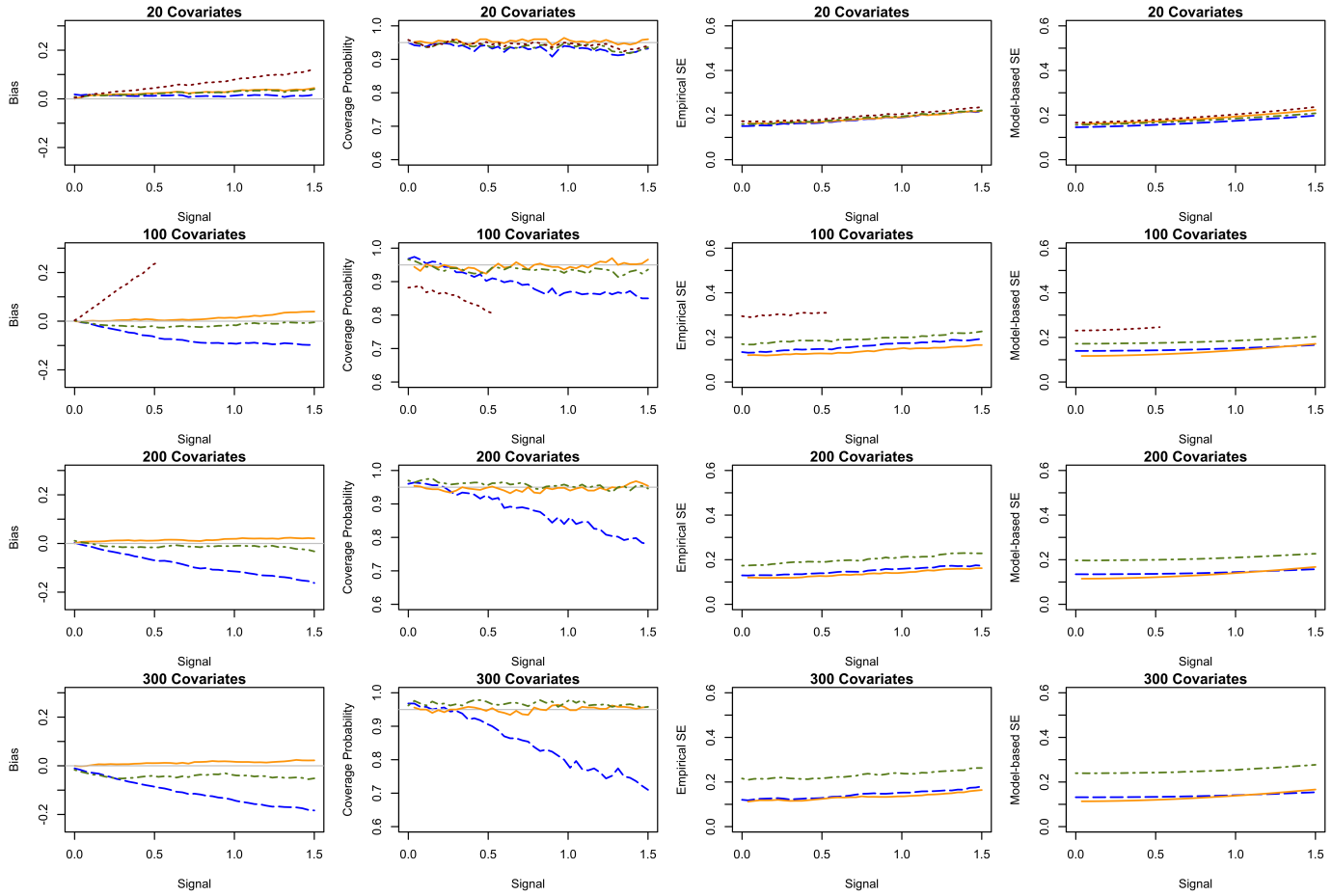

Figure 2.6: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being AR(1) with $\rho = 0.7$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

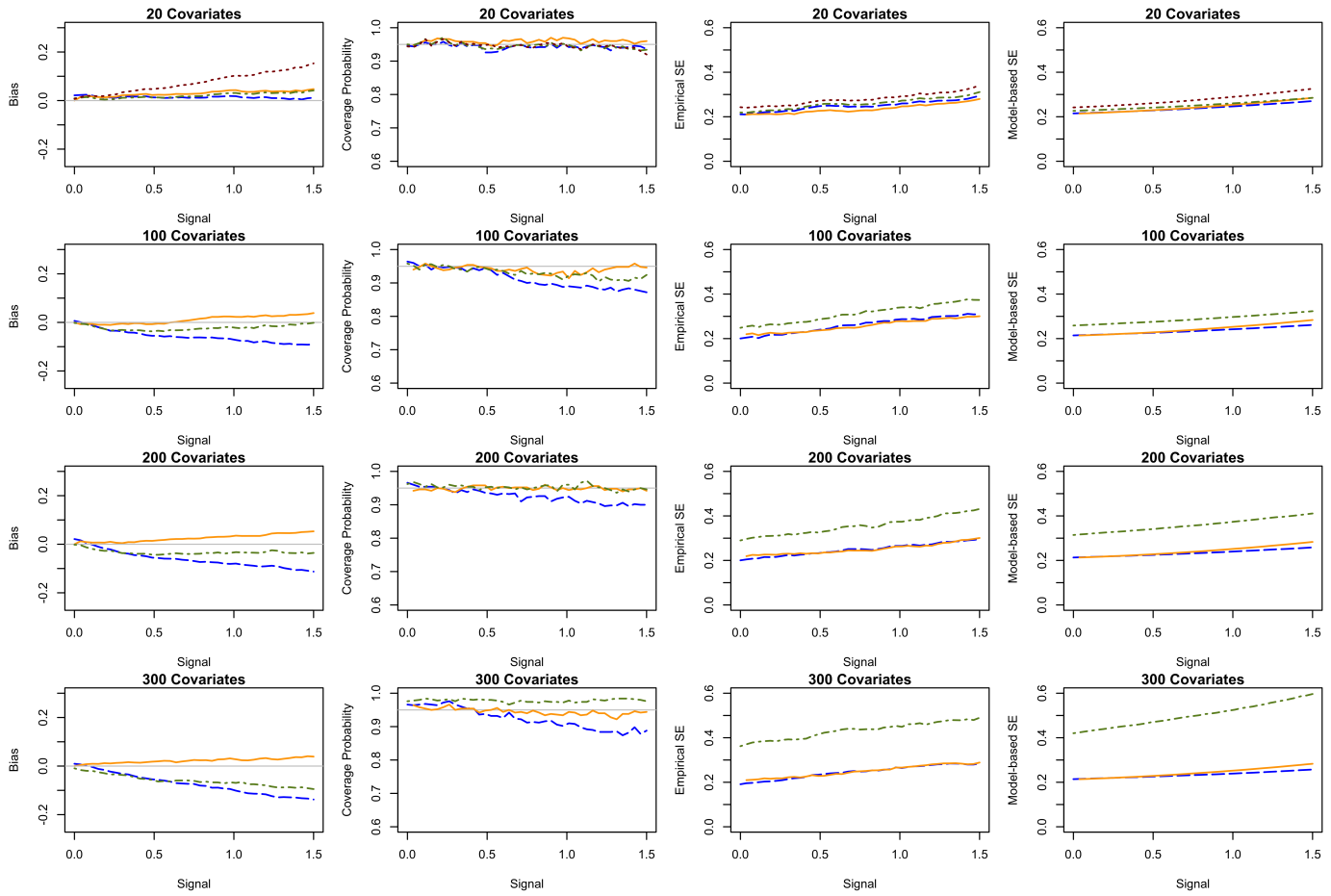

Figure 2.7: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being compound symmetry with $\rho = 0.7$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

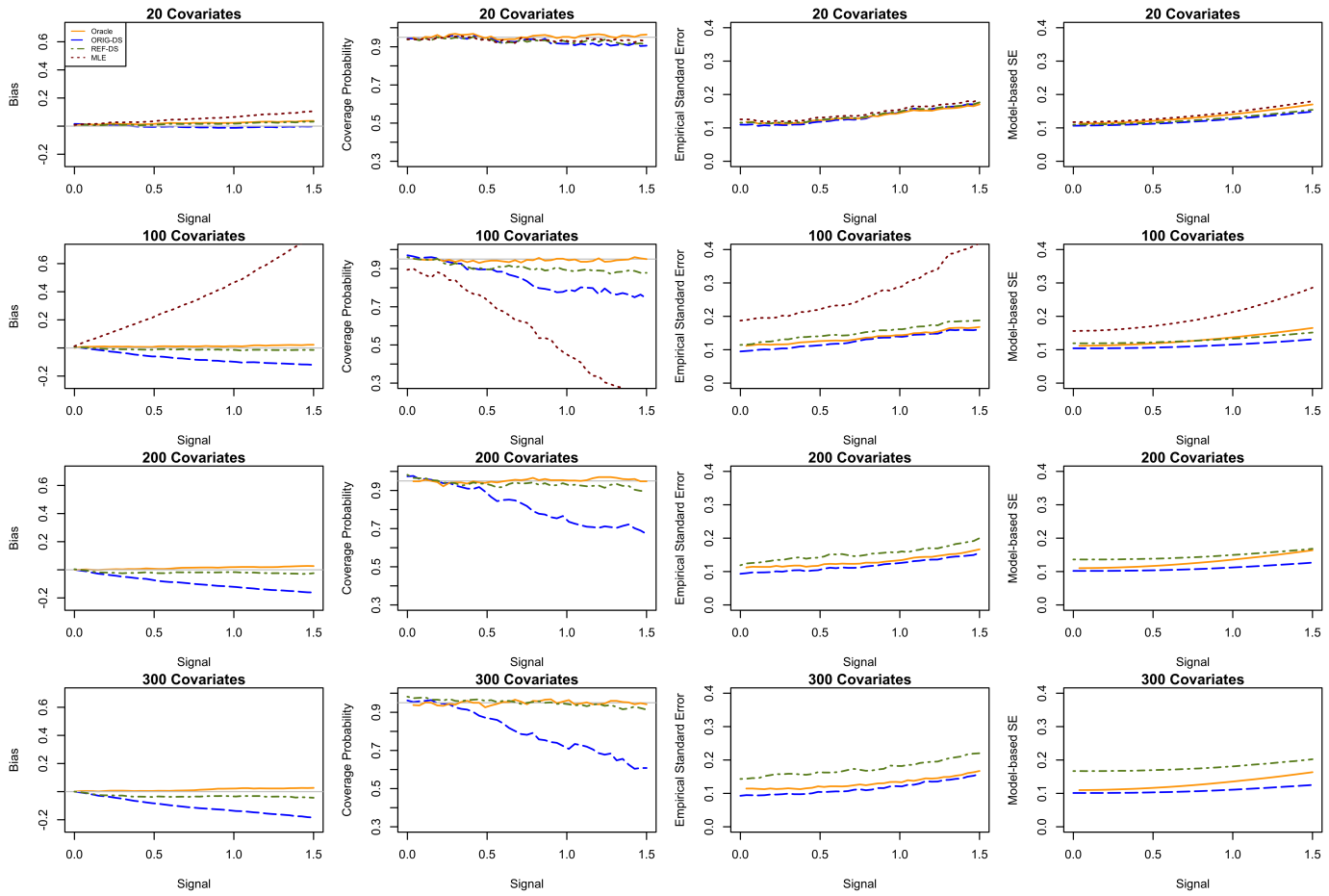

Figure 2.8: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being AR(1) with a small correlation $\rho = 0.2$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

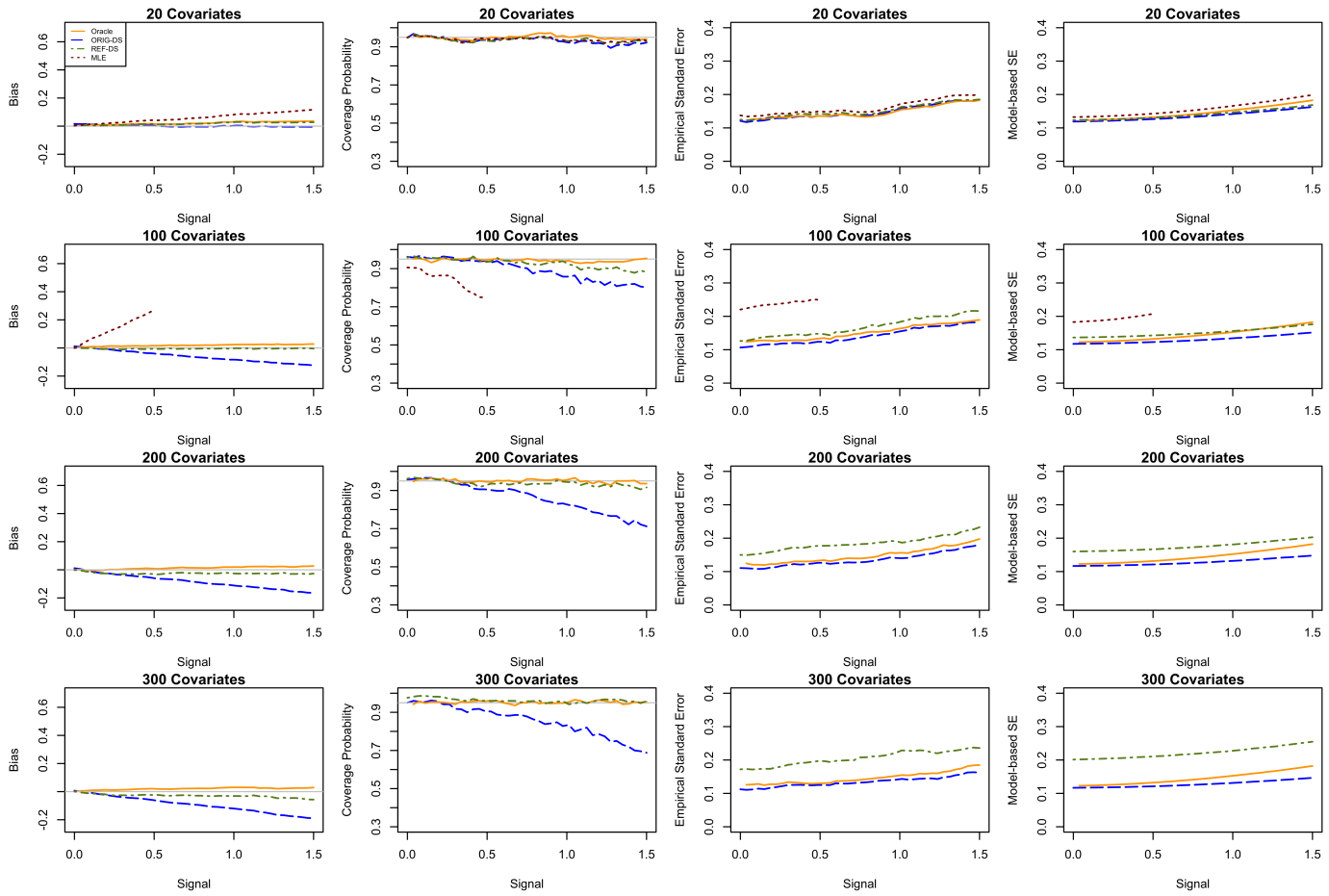

Figure 2.9: Simulation results: Bias, coverage probability, empirical standard error, and model-based estimated standard error for β_1^0 in a logistic regression. Covariates are simulated with Σ_x being compound symmetry with a small correlation $\rho = 0.2$. The sample size is $n = 500$ and the number of covariates $p = 20, 100, 200, 300$. MLE under the true model, denoted as "*Oracle*", is plotted as a reference.

2.4.3 Application to the Boston Lung Cancer Study

Lung cancer is the leading cause of cancer death in the United States. Boston Lung Cancer Study (BLCS) is a large epidemiology cohort for investigating the molecular cause of lung cancer, including lung cancer cases enrolled at Massachusetts General Hospital and the Dana-Farber Cancer Institute from 1992 to present[1]. We applied *REF-DS*, together with *ORIG-DS* and *MLE*, to a subset of the BLCS data and simultaneously examined the joint effects of SNPs in nine target genes on the overall risk of lung cancer.

Genotypes from Axiom array and clinical information were originally collected on 1,459 individuals. Table 2.1 summarizes the demographic information of the study population of 1,374 individuals, and by smoking status as well. Out of the 1,459 individuals, 14 (0.96%) had missing smoking status, 8 (0.55%) had missing race information, and 1,386 (95%) were Caucasian. We included a final number of $n = 1,374$ Caucasians, where $n_0 = 723$ were controls and $n_1 = 651$ were cases, with known lung cancer status ("1" for cases and "0" for controls) and smoking status ("1" for ever smoker and "0" for never). Among the 1,077 smokers, 595 had lung cancer, and the number of cases was 56 out of the 297 non-smokers. Other demographic characteristics of the study population, including education level (no high school, high school graduate, or at least 1-2 years of college), gender and age, are summarized in Table 2.1. Using the target gene approach, we focused on the following genes: *AK5* on region 1p31.1, *RNASET2* on region 6q27, *CHRNA2* and *EPHX2* on region 8p21.2, *BRCA2* on region 13q13.1, *SEMA6D* and *SECISBP2L* on region 15q21.1, *CHRNA5* on region 15q25.1, and *CYP2A6* on region 19q13.2. These genes have been reported in *McKay et al.* (2017) to harbor SNPs associated with the overall lung cancer risks. In our dataset, each SNP was coded as 0,1,2, reflecting the number of copies of the

[1]See the webpage `https://maps.cancer.gov/overview/DCCPSGrants/abstract.jsp?applId=9320074&term=CA209414`

minor allele, and was assumed to have "additive effects". After applying filters on the minor allele frequency, genotype call rate (percentage of missingness), and excluding SNPs that were highly correlated, 103 SNPs remained in the model.

Table 2.1: Demographic characteristics of the population under study in the Boston Lung Cancer Study

Information	Overall Count (%) / Mean (SD)	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)

The final analyzable dataset consisted of 1,374 individuals, 103 SNPs, and demographic information including education history, age and gender. Since existing studies suggest smoking can modify associations between lung cancer risks and SNPs, for example, those residing in region 15q25.1 (*Gabrielsen et al.*, 2013; *Amos et al.*, 2008), we conducted analysis stratified by smoking status. Within the smoker and non-smoker groups, we fitted separate logistic regression models, adjusting for educational history, gender and age (centered at the mean). In total, there were 107 variables for stratified analysis among 1,077 smokers and 297 non-smokers. As a reference, we conducted marginal analysis, which examined one SNP at a time while adjusting for demographic information. Marginal and joint analyses have distinct interpretations and can generate different estimates.

We applied these methods to draw inference on all of the 107 predictors, and comparisons of the results of the BLCS data analysis may shed light on the molecular mechanism underlying lung cancer. For ease of presentation, Table 2.2 lists the regression coefficient estimates, model-based estimated standard errors and 95% confidence intervals (CIs) for demographic variables and 11 SNPs in the stratified

analysis for an illustration. Some of these SNPs had at least one 95% CI (calculated by the three methods) that excluded 0 among either the smokers or the non-smokers; others showed differences among the estimating methods. Details of the remaining SNPs were omitted due to the space limitation. Since the number of the non-smokers was only about one third of the smokers, the *MLE* estimates had the largest standard errors and tended to break down among the non-smokers (see, for example, AX-62479186 in Table 2.2(b)), whereas the two de-biased lasso methods gave more reasonable estimates. The estimates by *REF-DS* and *ORIG-DS* shared more similarity in the smokers (Table 2.2(a)) than in the non-smokers (Table 2.2(b)). Overall, *ORIG-DS* had slightly narrower confidence intervals than *REF-DS*, probably due to penalized estimation for $\widehat{\mathbf{\Theta}}$. These results generally agreed with our simulation results.

Additional differences between *ORIG-DS* and *REF-DS* lied in opposite directions obtained for the estimated effects of some SNPs, such as AX-38419741 and AX-15934253 in Table 2.2(a), and AX-42391645 in Table 2.2(b). Among the non-smokers, the 95% CI for AX-31620127 in *SEMA6D* by *REF-DS* was all positive and excluded 0, while the CI by *ORIG-DS* included 0; the story for AX-88907114 in *CYP2A6* was just opposite (Table 2.2(b)).

CHRNA5 is a gene known for predisposition to nicotine dependence (*Halldén et al.*, 2016; *Hung et al.*, 2008; *Amos et al.*, 2008; *Thorgeirsson et al.*, 2008; *Gabrielsen et al.*, 2013). Though AX-39952685 and AX-88891100 in *CHRNA5* were not significant at level 0.05 in marginal analysis among the smokers, their 95% CIs in Table 2.2(a) excluded 0 by all of the three methods. Indeed AX-88891100, or rs503464 mapped to the same physical location in dbSNP[2], was found to "decrease *CHRNA5* promoter-derived luciferase activity" (*Doyle et al.*, 2011). The same SNP was also reported to be significantly associated with nicotine dependence at baseline, as well

[2]See the webpage `https://www.ncbi.nlm.nih.gov/snp`

as response to varenicline, bupropion, nicotine replacement therapy for smoking cessation (*Pintarelli et al.*, 2017). AX-39952685 was found to be strongly correlated with SNP AX-39952697 in *CHRNA5*, which was mapped to the same physical location as rs11633585 in dbSNP. All of these markers were found to be significantly associated with nicotine dependence (*Stevens et al.*, 2008). The stratified analysis also suggested molecular mechanisms of lung cancer differ between smokers and non-smokers, but affirmative conclusions need additional confirmatory studies. In summary, jointly modeling the genetic effects on lung cancer risks can help understand underlying mechanisms and personalized therapies, which necessitates the use of reliable inference tools.

Table 2.2: The association between SNPs and lung cancer risk in stratified analysis. Coefficient estimates in logistic regression models are reported for demographic variables and 11 SNPs (a) among the smokers, and (b) among the non-smokers. The other SNPs are omitted from the table.

(a) Smokers

				REF-DS			*ORIG-DS*			*MLE*		
			Demographic variable	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
			Education: No high school	0.44	0.20	(0.05, 0.83)	0.48	0.19	(0.11, 0.85)	0.52	0.22	(0.09, 0.94)
			Education: High school graduate	0.00	0.15	(-0.29, 0.28)	-0.02	0.14	(-0.29, 0.25)	-0.01	0.16	(-0.32, 0.30)
			Gender: Male	-0.14	0.13	(-0.40, 0.12)	-0.16	0.13	(-0.40, 0.09)	-0.15	0.14	(-0.43, 0.13)
			Age in years	0.04	0.07	(-0.09, 0.17)	0.05	0.06	(-0.07, 0.17)	0.05	0.07	(-0.09, 0.19)
SNP	Pos	Allele	Gene	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
AX-15319183	6:167352075	C/G	*RNASET2*	0.01	0.19	(-0.36, 0.39)	-0.03	0.18	(-0.39, 0.33)	0.02	0.20	(-0.38, 0.42)
AX-41911849	6:167360724	A/G	*RNASET2*	0.43	0.22	(0.00, 0.86)	0.44	0.20	(0.06, 0.83)	0.49	0.24	(0.03, 0.96)
AX-42391645	8:27319769	G/C	*CHRNA2*	0.01	0.16	(-0.29, 0.32)	-0.01	0.14	(-0.28, 0.26)	0.01	0.16	(-0.31, 0.34)
AX-38419741	8:27319847	T/A	*CHRNA2*	**0.11**	0.35	(-0.59, 0.80)	**-0.14**	0.31	(-0.75, 0.48)	0.13	0.37	(-0.60, 0.86)
AX-15934253	8:27334098	T/C	*CHRNA2*	**-0.15**	0.44	(-1.02, 0.71)	**0.06**	0.39	(-0.70, 0.82)	-0.19	0.47	(-1.12, 0.74)
AX-12672764	13:32927894	T/C	*BRCA2*	-0.07	0.19	(-0.44, 0.31)	-0.10	0.16	(-0.40, 0.21)	-0.07	0.20	(-0.47, 0.33)
AX-31620127	15:48016563	C/T	*SEMA6D*	0.79	0.26	(0.28, 1.31)	0.79	0.25	(0.30, 1.28)	0.96	0.30	(0.37, 1.55)
AX-88891100	15:78857896	A/T	*CHRNA5*	0.87	0.36	(0.17, 1.57)	0.79	0.32	(0.16, 1.41)	0.98	0.39	(0.22, 1.74)
AX-39952685	15:78867042	G/C	*CHRNA5*	0.99	0.47	(0.07, 1.91)	0.82	0.38	(0.09, 1.56)	1.11	0.50	(0.13, 2.08)
AX-62479186	15:78878565	T/C	*CHRNA5*	0.41	0.41	(-0.40, 1.22)	0.46	0.39	(-0.31, 1.23)	0.46	0.45	(-0.41, 1.33)
AX-88907114	19:41353727	T/C	*CYP2A6*	0.52	0.34	(-0.16, 1.19)	0.49	0.33	(-0.15, 1.13)	0.58	0.38	(-0.15, 1.32)
⋮												

(b) Non-smokers

				REF-DS			*ORIG-DS*			*MLE*		
			Demographic variable	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
			Education: No high school	-0.84	0.93	(-2.67, 0.99)	-0.58	0.78	(-2.11, 0.95)	-10.53	3.82	(-18.01, -3.04)
			Education: High school graduate	-1.68	0.52	(-2.69, -0.66)	-1.56	0.43	(-2.39, -0.72)	-11.23	3.75	(-18.58, -3.88)
			Gender: Male	-0.30	0.41	(-1.10, 0.51)	-0.16	0.32	(-0.78, 0.46)	-1.94	1.03	(-3.96, 0.09)
			Age in years	-0.52	0.20	(-0.91, -0.13)	-0.56	0.16	(-0.87, -0.26)	-2.59	0.98	(-4.51, -0.67)
SNP	Pos	Allele	Gene	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
AX-15319183	6:167352075	C/G	*RNASET2*	-0.71	0.55	(-1.78, 0.36)	0.01	0.40	(-0.79, 0.80)	-4.32	1.84	(-7.92, -0.71)
AX-41911849	6:167360724	A/G	*RNASET2*	0.69	0.65	(-0.59, 1.97)	0.37	0.47	(-0.55, 1.29)	4.46	2.00	(0.54, 8.39)
AX-42391645	8:27319769	G/C	*CHRNA2*	**-0.11**	0.49	(-1.07, 0.85)	**0.18**	0.30	(-0.41, 0.78)	-1.90	2.00	(-5.81, 2.02)
AX-38419741	8:27319847	T/A	*CHRNA2*	0.50	1.04	(-1.54, 2.53)	0.23	0.61	(-0.97, 1.42)	3.37	3.00	(-2.51, 9.26)
AX-15934253	8:27334098	T/C	*CHRNA2*	0.11	1.40	(-2.64, 2.86)	0.38	0.82	(-1.23, 1.98)	5.37	4.21	(-2.88, 13.62)
AX-12672764	13:32927894	T/C	*BRCA2*	-0.83	0.62	(-2.04, 0.37)	-0.57	0.38	(-1.32, 0.18)	-8.25	2.64	(-13.42, -3.08)
AX-31620127	15:48016563	C/T	*SEMA6D*	1.77	0.75	**(0.30, 3.24)**	0.43	0.46	**(-0.48, 1.34)**	9.23	3.27	(2.81, 15.64)
AX-88891100	15:78857896	A/T	*CHRNA5*	0.78	1.18	(-1.54, 3.10)	1.15	0.87	(-0.56, 2.85)	1.54	3.17	(-4.68, 7.75)
AX-39952685	15:78867042	G/C	*CHRNA5*	-0.54	1.30	(-3.09, 2.01)	-0.99	0.73	(-2.41, 0.44)	-2.85	3.98	(-10.65, 4.96)
AX-62479186	15:78878565	T/C	*CHRNA5*	-1.28	1.34	(-3.92, 1.35)	-1.33	1.10	(-3.49, 0.82)	**-19.64**	**3410.98**	(-6705.04, 6665.75)
AX-88907114	19:41353727	T/C	*CYP2A6*	0.86	0.88	**(-0.86, 2.59)**	1.40	0.68	**(0.06, 2.74)**	3.52	2.18	(-0.75, 7.78)
⋮												

Pos: physical location of a SNP on a chromosome (Assembly GRCh37/hg19). Est: estimated coefficient in the logistic regression models for the overall risk of lung cancer. SE: estimated standard error. CI: confidence interval.

2.5 Discussion

Our work has produced several intriguing results that can be impactful in both theory and practical implementation. From extensive simulations we have discovered the unsatisfactory performance of de-biased lasso in drawing inference with high-dimensional GLMs. We have further pinpointed an essential assumption that hardly holds for GLMs in general, i.e. the sparsity of the high-dimensional inverse information matrix $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ (*van de Geer et al.*, 2014), making de-biased lasso fail to deliver reliable inference in practice. This type of ℓ_0 sparsity conditions on matrices is not uncommon in the literature of high-dimensional inference. A related ℓ_0 sparsity condition on $\mathbf{w}^* = \mathbf{I}^{*-1}{}_{\gamma\gamma}\mathbf{I}^*_{\gamma\theta}$ can be found in *Ning and Liu* (2017), where $\mathbf{I}^*$ is the information matrix under the truth, but is not well justified in a general GLM setting. When testing a global null hypothesis ($\boldsymbol{\beta}^0 = \mathbf{0}$), however, the sparsity of $\mathbf{\Theta}_{\boldsymbol{\xi}^0}$ reduces to the sparsity of the covariate precision matrix, which becomes less of an issue (see *Cai et al.* 2019).

Our detailed work leads to practical guidelines as to how to use de-biased lasso for proper statistical inference with high-dimensional GLMs. Our work summarily suggests that, when $p > n$, de-biased lasso may not be applicable in general; when $p < n$ with diverging p, it is preferred to use the refined de-biased lasso, which directly inverts the Hessian matrix and provides improved confidence interval coverage probabilities for a wide range of p; when p is rather small relative to n (often viewed as a fixed p problem), the refined de-biased lasso yields results nearly identical to MLE and the original de-biased lasso.

2.6 Technical proofs

We provide three lemmas that are useful for proving Theorem II.1. And the proof for Theorem II.1 is provided at the end of this section. Without loss of generality,

we denote the dimension of the parameter $\boldsymbol{\xi}$ by p instead of $(p+1)$ to simplify the notation in the proofs. Consequently, the matrices such as $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\xi}}$ are considered as $p \times p$ matrices. The simplification of notation does not affect derivations.

Lemma II.5. *Under (C1) - (C4) in Section 2.4.1, we have* $\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\|_1 = \mathcal{O}_P(s_0\lambda)$ *and* $\|\mathbf{X}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0)\|_2^2/n = \mathcal{O}_P(s_0\lambda^2)$.

Proof. Because $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) > 0$ in (C2), the compatibility condition holds for all index sets $S \subset \{1, \cdots, p\}$ by Lemma 6.23 of *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 $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}^0} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\|_\infty = \mathcal{O}_P(\sqrt{log(p)/n})$. Then by Lemma 6.17 of *Bühlmann and van de Geer* (2011), we have the $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}^0}$-compatibility condition. Finally, the first part of Lemma II.5 follows from Theorem 6.4 in *Bühlmann and van de Geer* (2011).

For the second claim, *Ning and Liu* (2017) showed that $(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0)^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}^0} (\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0)^T = \mathcal{O}_P(s_0\lambda^2)$, then under (C4), we obtain the desired result. □

Lemma II.6. *Under (C1) - (C5) in Section 2.4.1, if we further assume that* $s_0\lambda \to 0$ *and* $L_p^2\sqrt{\frac{p}{n}} \to 0$, *then* $\widetilde{\boldsymbol{\Theta}}$ *converges with the following rate*

$$||\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0}|| = \mathcal{O}_P\left(L_p^2\sqrt{\frac{p}{n}} + s_0\lambda\right).$$

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

$$\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-1}\| \leq \|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}\| \cdot \|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\| \cdot \|\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-1}\|. \tag{2.8}$$

By (C2), $\|\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-1}\|$ is bounded. We obtain the convergence rate of $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-1}\|$ by calculating the rate of $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\|$ and showing that $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}\|$ is bounded with probability going to 1.

Note that $\|\widehat{\mathbf{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| \leq \|\widehat{\mathbf{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0}\| + \|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\|$. When the rows of $\mathbf{X}$ are sub-Gaussian, so are the rows of $\mathbf{X}_{\boldsymbol{\xi}^0}$ due to the boundedness of the weights w_i in (C3). First, for $\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\|$, *Vershynin* (2010) shows that for every $t > 0$, it holds with probability at least $1 - 2\exp(-c_L' t^2)$ that

$$\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| \leq \|\mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| \max(\delta, \delta^2) \leq c_{\max} \max(\delta, \delta^2), \tag{2.9}$$

where $\delta = C_L\sqrt{\dfrac{p}{n}} + \dfrac{t}{\sqrt{n}}$. Here C_L, $c_L' > 0$ depend only on $L_p = \|\mathbf{\Sigma}_{\boldsymbol{\xi}^0}^{-\frac{1}{2}} \boldsymbol{x}_1 \omega_1(\boldsymbol{\xi}^0)\|_{\psi_2}$. In fact $c_L' = c_1/L_p^4$ and $C_L = L_p^2\sqrt{\log 9/c_1}$, where c_1 is an absolute constant. For $s > 0$ and $t = sC_L\sqrt{p}$, the probability becomes $1 - 2\exp(-c_2 s^2 p)$, $c_2 > 0$ being some absolute constant, and $\delta = (s+1)C_L\sqrt{\dfrac{p}{n}}$. Thus $\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| = \mathcal{O}_p\left(L_p^2\sqrt{\dfrac{p}{n}}\right)$.

Note that

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

By (C1) and (C3),

$$\begin{aligned}
\|\boldsymbol{W}_{\widehat{\boldsymbol{\xi}}}^2 - \boldsymbol{W}_{\boldsymbol{\xi}^0}^2\| &= \max_i |\ddot{\rho}(y_i, \boldsymbol{x}_i^T\widehat{\boldsymbol{\xi}}) - \ddot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)| \\
&\leq c_{Lip} \cdot \max_i |\boldsymbol{x}_i^T(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0)| \\
&\leq c_{Lip} K \cdot \|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\|_1.
\end{aligned} \tag{2.10}$$

By Lemma II.5, we have $\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\|_1 = \mathcal{O}_P(s_0\lambda)$. In this case, $\|\boldsymbol{W}_{\widehat{\boldsymbol{\xi}}}^2 - \boldsymbol{W}_{\boldsymbol{\xi}^0}^2\| = \mathcal{O}_P(s_0\lambda)$. By (C5) and *Vershynin* (2010), $\lambda_{\max}(\mathbf{X}^T\mathbf{X}/n) = \mathcal{O}_P(1)$. Thus $\|\widehat{\mathbf{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}^0}\| = \mathcal{O}_P(s_0\lambda)$. Therefore, after combining the two parts, we have $\|\widehat{\mathbf{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| = \mathcal{O}_P\left(L_p^2\sqrt{\dfrac{p}{n}} + s_0\lambda\right)$. Under $L_p^2\sqrt{\dfrac{p}{n}} = o(1)$ and $s_0\lambda = o(1)$, we have $\|\widehat{\mathbf{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \mathbf{\Sigma}_{\boldsymbol{\xi}^0}\| = o_P(1)$.

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

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

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

$$\begin{aligned} P\left(\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}\| \geq \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) - \epsilon}\right) &= P(\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}) \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) - \epsilon) \\ &\leq P(|\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}) - \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0})| \geq \epsilon) \\ &\leq P(\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\| \geq \epsilon). \end{aligned}$$

Since $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\| = o_P(1)$, we have $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1}\| = \mathcal{O}_P(1)$. Finally, by (2.8), $\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}^{-1}\| = \mathcal{O}_P(\|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}\|) = \mathcal{O}_P\left(L_p^2\sqrt{\frac{p}{n}} + s_0\lambda\right)$. □

Lemma II.7. *Under (C1)-(C3) in Section 2.4.1, when* $\frac{p}{n} \to 0$*, it holds that for any vector* $\boldsymbol{\alpha}_n \in \mathbb{R}^p$ *with* $\|\boldsymbol{\alpha}_n\|_2 = 1$*,*

$$\frac{\sqrt{n}\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}}{\sqrt{\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n}} \xrightarrow{d} N(0,1).$$

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

$$Z_{ni} = \frac{1}{\sqrt{n}}\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}(y_i, \boldsymbol{x}_i) = \frac{1}{\sqrt{n}}\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{x}_i\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0),$$

and $s_n^2 = Var\left(\sum_{i=1}^n Z_{ni}\right)$. Note that $\mathbb{E}[\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)|\boldsymbol{x}_i] = 0$ and consequently $\mathbb{E}(Z_{ni}) = 0$. Because $\{(y_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^n$ are *i.i.d.*, we can show that $s_n^2 = \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n$. To show $\frac{\sum_{i=1}^n Z_{ni}}{s_n} \xrightarrow{d} N(0,1)$, 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 $n \to \infty$,

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

Due to the boundedness of the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}$, $\boldsymbol{\alpha}_n^T \boldsymbol{\Theta}_{\boldsymbol{\xi}^0} \boldsymbol{\alpha}_n \geq \lambda_{\min}(\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}) = 1/\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}^0}) \geq c_{\max}^{-1}$. On the other hand, by the Cauchy-Schwarz inequality, it holds almost surely that

$$\left(\boldsymbol{\alpha}_n^T \boldsymbol{\Theta}_{\boldsymbol{\xi}^0} \boldsymbol{x}_i\right)^2 \leq \|\boldsymbol{\alpha}_n\|_2^2 \cdot \|\boldsymbol{\Theta}_{\boldsymbol{\xi}^0} \boldsymbol{x}_i\|_2^2 \leq \left[\|\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| \cdot \|\boldsymbol{x}_i\|_2\right]^2 \leq c_{\min}^{-2} \cdot \mathcal{O}(pK^2).$$

Inside the indicator, it holds almost surely that

$$\begin{aligned} \frac{Z_{ni}^2}{s_n^2} &= \frac{[\dot{\rho}(y_i, \boldsymbol{x}_i^T \boldsymbol{\xi}_0)]^2 \left(\boldsymbol{\alpha}_n^T \boldsymbol{\Theta}_{\boldsymbol{\xi}^0} \boldsymbol{x}_i\right)^2}{n \boldsymbol{\alpha}_n^T \boldsymbol{\Theta}_{\boldsymbol{\xi}^0} \boldsymbol{\alpha}_n} \\ &\leq [\dot{\rho}(y_i, \boldsymbol{x}_i^T \boldsymbol{\xi}_0)]^2 \cdot c_{\min}^{-2} c_{\max} \cdot \mathcal{O}(K^2 \frac{p}{n}) \\ &\leq K_1^2 c_{\min}^{-2} c_{\max} \cdot \mathcal{O}(K^2 \frac{p}{n}), \end{aligned}$$

where the last inequality follows from the boundedness of $\dot{\rho}(y_i, \boldsymbol{x}_i^T \boldsymbol{\xi}_0)$ in condition (C3). 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, by the Dominated Convergence Theorem, the Lindeberg condition holds and the Lindeber-Feller Central Limit Theorem guarantees the asymptotic normality. □

Finally, we provide the theoretical proof for the main result in Section 2.4.1, Theorem II.1.

Proof of Theorem II.1. Recall that from (2.7),

$$\sqrt{n} \boldsymbol{\alpha}_n^T (\widetilde{\boldsymbol{b}} - \boldsymbol{\xi}^0) - \sqrt{n} \boldsymbol{\alpha}_n^T \widetilde{\boldsymbol{\Theta}} \boldsymbol{\Delta} = -\sqrt{n} \boldsymbol{\alpha}_n^T \widetilde{\boldsymbol{\Theta}} \mathbb{P}_n \dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}.$$

First, we show that $\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n = o_{\mathbb{P}}(1)$ and that $\dfrac{\sqrt{n}\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}}{\sqrt{\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n}} = \dfrac{\sqrt{n}\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}}{\sqrt{\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n}} + o_{\mathbb{P}}(1)$. Then by Slutsky's Theorem, the asymptotic distribution of the target $\dfrac{\sqrt{n}\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}}{\sqrt{\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n}}$ can be derived by using the asymptotic distribution of $\dfrac{\sqrt{n}\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}}{\sqrt{\boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n}}$, which has been proved in Lemma II.7. In the final step, as long as $\sqrt{n}\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta} = o_P(1)$, the asymptotic distribution of $\dfrac{\sqrt{n}\boldsymbol{\alpha}_n^T(\widetilde{\boldsymbol{b}} - \boldsymbol{\xi}^0)}{\sqrt{\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n}}$ follows immediately.

According to Lemma II.6, it follows that

$$|\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\boldsymbol{\alpha}_n| = |\boldsymbol{\alpha}_n^T(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0})\boldsymbol{\alpha}_n| \leq \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| \cdot \|\boldsymbol{\alpha}_n\|_2^2 = o_P(1).$$

By the Cauchy-Schwartz inequality,

$$\sqrt{n}|\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0} - \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}| \leq \sqrt{n}\|\boldsymbol{\alpha}_n\|_2 \cdot \|(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0})\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_2.$$

Since

$$\begin{aligned}\|(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0})\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_2 &\leq \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| \cdot \|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_2 \\ &\leq \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| \cdot \sqrt{p}\|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_\infty,\end{aligned}$$

we have

$$\begin{aligned}\sqrt{n}\left|\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0} - \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\right| &\leq \sqrt{np} \cdot \|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_\infty \cdot \mathcal{O}_P\left(L_p^2\sqrt{\frac{p}{n}} + s_0\lambda\right) \\ &= \|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_\infty \cdot \mathcal{O}_P\left(L_p^2 p + \sqrt{np}s_0\lambda\right).\end{aligned}$$

By definition,

$$\|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_\infty = \max_j\left|\frac{1}{n}\sum_{i=1}^n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}(y_i, \boldsymbol{x}_i)\right| = \max_j\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)\right|.$$

Assume $|\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)| \leq K_1$ for all i and the constant $K_1 > 0$ in condition (C3). As $|x_{ij}\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)| \leq KK_1$ almost surely holds for all i and j, we apply Lemma 14.15 in

Bühlmann and van de Geer (2011), for all $t > 0$,

$$\mathbb{P}\left(\max_j \left|\frac{1}{n}\sum_{i=1}^n x_{ij}\dot{\rho}(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}^0)\right| \geq KK_1\sqrt{2\left(t^2 + \frac{\log(2p)}{n}\right)}\right) \leq \exp[-nt^2].$$

For $t^2 = \dfrac{\log(2p)}{n}$, we know that $\|\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\|_\infty = \mathcal{O}_P\left(\sqrt{\dfrac{\log(p)}{n}}\right)$. Then we have

$$\sqrt{n}\left|\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0} - \boldsymbol{\alpha}_n^T\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\mathbb{P}_n\dot{\boldsymbol{\rho}}_{\boldsymbol{\xi}^0}\right| \leq \mathcal{O}_P\left(L_p^2 p\sqrt{\frac{\log(p)}{n}} + s_0\lambda\sqrt{p\log(p)}\right),$$

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

Finally, we prove $|\sqrt{n}\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}| = o_P(1)$. By the Cauchy-Schwartz inequality, $|\sqrt{n}\boldsymbol{\alpha}_n^T\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}| \leq \sqrt{n}\|\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}\|_2$, we only need that $\sqrt{n}\|\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Delta}\|_2 = o_P(1)$. In equation (2.3),

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

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

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

where the last equality holds by Lemma II.5. Since $\|\boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| = \mathcal{O}(1)$ and $\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_{\boldsymbol{\xi}^0}\| =$

$o_P(1)$, then $\|\widetilde{\mathbf{\Theta}}\| = \mathcal{O}_P(1)$, and we have

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

By the assumption of $\sqrt{np}s_0\lambda^2 = o(1)$, $\sqrt{n}\|\widetilde{\mathbf{\Theta}}\mathbf{\Delta}\|_2 = o_P(1)$. Applying Slutsky's Theorem and Lemma II.7 gives the results. □

CHAPTER III

Statistical Inference for Cox Proportional Hazards Model with A Diverging Number of Covariates

3.1 Introduction

The Cox proportional hazards model (*Cox*, 1972) has been widely used for analysis of censored time-to-event data. This model is semi-parametric without specifying the baseline hazard function, and *Cox* (1972) proposed the maximum partial likelihood estimator to infer the unknown finite-dimensional parameter in the posited hazard function. *Andersen and Gill* (1982) proved the asymptotic distribution for the maximum partial likelihood estimator using martingale theory under the fixed dimension setting.

In the big data era, it is now possible to collect a large amount of information in biomedical studies such as genomics and imaging studies. For example, the Boston Lung Cancer Study provides rich resources of clinical, gene expression, methylation and genomics data, which enables innovative investigations into the molecular mechanisms underlying lung cancer patient survival and promotes precision medicine for lung cancer patients (*McKay et al.*, 2017). High-dimensionality of the covariates collected has brought new challenges to parameter estimation and uncertainty quantification in the Cox model. In high-dimensional settings, where the number of covariates

p increases with the sample size n or even greater than n, the conventional maximum partial likelihood estimation is usually ill-conditioned. Penalized estimators have emerged as a useful tool for simultaneous variable selection and estimation (*Tibshirani*, 1997; *Fan and Li*, 2002; *Gui and Li*, 2005; *Antoniadis et al.*, 2010). *Huang et al.* (2013) and *Kong and Nan* (2014) have studied the non-asymptotic oracle inequalities of the lasso estimator in the Cox model, which entail additional difficulties since the negative log partial likelihood loss function is not a sum of independent and identically distributed terms nor Lipschitz.

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 de-biasing the lasso estimator via low-dimensional projection or inverting the Karush–Kuhn–Tucker condition. In the same paper, *van de Geer et al.* (2014) extended the de-biasing lasso idea to generalized linear models, using the nodewise lasso regression to approximate the large inverse information matrix. *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.

Literature in high-dimensional Cox model inference is limited and unsatisfactory. *Fang et al.* (2017) developed decorrelated tests for hypothesis testing of low-dimensional components, similar to *Ning and Liu* (2017) but in the high-dimensional Cox model. *Kong et al.* (2018) extended the de-biased lasso approach in *van de Geer et al.* (2014) to potentially misspecified Cox model, using the nodewise lasso regression to estimate the inverse information matrix. *Yu et al.* (2018) proposed a de-biased lasso approach by estimating the inverse information matrix with CLIME, adapted from *Cai et al.* (2011) that was originally designed for precision matrix estimation. However, these existing methods for statistical inference in the high-dimensional Cox

model have certain limitations. Due to the properties of the nodewise lasso regression and CLIME, both *Kong et al.* (2018) and *Yu et al.* (2018) restricted the number of non-zero elements of each row in the inverse Fisher information matrix to be small, i.e. ℓ_0 sparsity. In Chapter II, it has been argued that such a sparsity assumption on the high-dimensional inverse information matrix does not hold in general settings of generalized linear models. This is because, in generalized linear models, the information matrix takes the form of $\mathbb{E}[X^T W_{\beta^0} X/n]$, where W_{β^0} is a diagonal matrix with the response variances on the diagonal, distorting the interpretability and the validity of sparsity on its inverse matrix in general settings. The same argument is also applicable in the Cox model. *Fang et al.* (2017) imposed an ℓ_0 sparsity assumption on $w^* = H^{*-1}_{\theta\theta} H^*_{\theta\alpha}$, where, in their notation, H^* is the Fisher information matrix, α is the low-dimensional component of interest, θ is the high-dimensional nuisance parameter, and w^* is approximated using the Dantzig selector. To summarize, imposing these sparsity conditions plays an important role in the development of the theoretical properties of the aforementioned estimators under the "large p, small n" scenario, yet has no practical interpretation and does not usually hold in the Cox model. Our extensive simulations also show that these methods perform unsatisfactorily in correcting the biases from penalized estimators and delivering honest confidence intervals.

Without imposing structural assumptions on the inverse information matrix, such as the aforementioned ℓ_0 sparsity, it is hard to estimate such a high-dimensional matrix with guaranteed precision when the number of covariates p exceeds the sample size n. In this paper, we consider the problem of drawing inference for regression coefficients in the Cox model without sparse estimation for the inverse information matrix, under a promising "large n, diverging p" scenario where $p < n$ but p is allowed to increase with n to infinity. Our primary focus is on improving the bias correction and delivering more reliable confidence intervals and inference results. Chapter II also suggested directly inverting the information matrix after variable selection in

generalized linear models, when the covariate dimension p grows with the sample size n satisfying $p^2 \log(p)/n \to 0$ as $n \to 0$ under certain conditions. Our numerical exploration indicates that directly applying this approach in the Cox model is problematic in bias correction and that more careful tuning of the parameters in matrix estimation procedures is warranted.

Inspired by *Javanmard and Montanari* (2014), we propose a de-biased lasso approach via solving quadratic programming problems to estimate the inverse information matrix, which can be viewed as a trade-off between estimation bias and variance and does not rely on unrealistic ℓ_0 sparsity assumptions on the large inverse information matrix. The contribution of this chapter is in both theoretical and practical aspects. First, unlike *Javanmard and Montanari* (2014) focusing on linear regression, this work entails careful treatment of the sum of non independently nor identically distributed terms in the loss function. We consider random rather than deterministic designs. Second, we have found that tuning parameter selection is crucial for the inverse information matrix estimation and consequently sufficient bias correction. A carefully designed adaptive procedure is proposed to select the important tuning parameter in the quadratic programming problems and is shown with satisfactory numerical performance. From the practical side, this also distinguishes our work from *Yu et al.* (2018).

The rest of this chapter is organized as follows. Section 3.2 introduces the proposed de-biasing approach approach via quadratic programming for matrix estimation, with a novel procedure for selection of the tuning parameter. Section 3.3 provides the theoretical justification that lays the foundation for reliable inference on linear combinations of the resulting de-biased lasso estimator. We demonstrate the superior performance of our proposed method via simulation studies in Section 3.4, and an application to the Boston Lung Cancer Study is shown in Section 3.5. Following the conclusing remarks in Section 3.6, technical proofs are presented in Section 3.7.

3.2 Method

3.2.1 Background and set-up

We first introduce some notation that will be used throughout this chapter. 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 $\|x\|_q = (\sum_{j=1}^r |x_j|^q)^{1/q}$, $q \geq 1$. 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 $\|A\|_\infty = \max_{i,j} |a_{ij}|$.

The Cox model assumes that the true hazard function for the underlying failure time T, conditional on a p-dimensional vector of covariates $X \in \mathbb{R}^p$, is $\lambda(t|X) = \lambda_0(t) \exp\{X^T \beta^0\}$, where $\lambda_0(t)$ 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 true regression coefficients. The observed survival time is denoted as $Y = \min(T, C)$, where the censoring time C is independent of T given the covariates X. Let $\delta = 1(T \leq C)$ denote the event indicator. We have n independent and identically distributed observations $\{Y_i, \delta_i, X_i\}_{i=1}^n$. The primary goal is to simultaneously estimate and draw inference on the regression coefficients β^0.

3.2.2 Quadratic programming for matrix estimation in the de-biased lasso

We write the negative log partial likelihood function based on the Cox model as

$$\ell_n(\beta) = -\frac{1}{n} \sum_{i=1}^n \left[X_i^T \beta - \log \left\{ \frac{1}{n} \sum_{j=1}^n 1(Y_j \geq Y_i) \exp(X_j^T \beta) \right\} \right] \delta_i. \tag{3.1}$$

The first and second order derivatives of (3.1) with respect to β are denoted as

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

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

$$\widehat{\eta}_n(t;\beta) = \frac{\widehat{\mu}_1(t;\beta)}{\widehat{\mu}_0(t;\beta)} = \frac{\sum_{j=1}^{n} 1(Y_j \geq t)\exp\{X_j^T\beta\}X_j}{\sum_{j=1}^{n} 1(Y_j \geq t)\exp\{X_j^T\beta\}}.$$

The lasso estimator $\widehat{\beta}$ minimizes the penalized negative log partial likelihood, i.e.

$$\widehat{\beta} \in \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\ell_n(\beta) + \lambda||\beta||_1\right\}, \tag{3.2}$$

for some tuning parameter $\lambda > 0$.

Motivated by *Javanmard and Montanari* (2014), we consider the one-step estimator updated from $\widehat{\beta}$ by first solving the following quadratic programming problem for each $j = 1, \ldots, p$,

$$\min\{m^T\widehat{\Sigma}m : m \in \mathbb{R}^p, \|\widehat{\Sigma}m - e_j\|_\infty \leq \gamma_n\}, \tag{3.3}$$

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

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n} \delta_i\{X_i - \widehat{\eta}_n(Y_i;\widehat{\beta})\}^{\otimes 2}. \tag{3.4}$$

$\widehat{\Sigma}$ in (3.3) is an alternative for the Hessian matrix and can be replaced by the second order derivative $\ddot{\ell}_n(\widehat{\beta})$. Choosing $\widehat{\Sigma}$ over $\ddot{\ell}_n(\widehat{\beta})$ in (3.3) is due to theoretical convenience, and the numerical difference in the resulting de-biased lasso estimators is negligible. Let the column vector $m^{(j)} \in \mathbb{R}^p$ be a solution to (3.3), and define the $p \times p$ matrix $\widehat{\Theta} = (m^{(1)}, \ldots, m^{(p)})^T$. Then the de-biased lasso estimator for β^0 is

$$\widehat{b} = (\widehat{b}_1, \ldots, \widehat{b}_p)^T = \widehat{\beta} - \widehat{\Theta}\dot{\ell}_n(\widehat{\beta}), \tag{3.5}$$

that is, $\widehat{b}_j = \widehat{\beta}_j - m^{(j)^T}\dot{\ell}_n(\widehat{\beta})$. In Section 3.3, we will provide the asymptotic theory for $\widehat{b}$ under commonly used regularity conditions.

The quadratic programming problem (3.3) is the same as in *Javanmard and Montanari* (2014) *per se*, except that the latter has a different definition $\sum_{i=1}^n X_i X_i^T/n$ for the matrix $\widehat{\Sigma}$ in linear models. Since $\widehat{\eta}_n(Y_i;\widehat{\beta})$ involves data from all subjects, $\widehat{\Sigma}$ as shown in (3.4) is no longer a sum of independently and identically distributed terms, which poses additional theoretical difficulties. This de-biasing lasso procedure through solving (3.3) is also related to *Yu et al.* (2018). As mentioned in the introduction, *Yu et al.* (2018) employed CLIME for the inverse information matrix estimation, which can be obtained in a column-wise fashion by solving linear programming problems

$$\min\{\|m\|_1 : m \in \mathbb{R}^p, \|\widehat{\Sigma}m - e_j\|_\infty \le \gamma_n\},\ j = 1, \ldots, p.$$

Unlike *Yu et al.* (2018), our proposed method shares the property with *Javanmard and Montanari* (2014) that the resulting $\widehat{\Theta}$ is not a sparse matrix due to the quadratic loss. Computationally, (3.3) can be easily implemented using existing softwares such as the R function `solve.QP`, which is usually fast and can be programmed in parallel for a very large dimension.

3.2.3 Selection of the tuning parameter

We have found that selecting a proper tuning parameter γ_n is very crucial for sufficient bias correction in $\widehat{b}$, which is demonstrated with a simple example. We simulate $n = 500$ subjects and $p = 100$ covariates independently from $N(0, 1)$, and only two coefficients of β^0 in the Cox model are non-zero, taking values of 1 and 0.3. The underlying survival time Y follows an exponential distribution with rate $\exp(X^T\beta^0)$, and the censoring time is simulated from exponential distribution with rate $0.2\exp(X^T\beta^0)$, which results in 15% - 20% censoring. Figure 3.1 illustrates how the estimation bias and the empirical coverage probability from the de-biased lasso

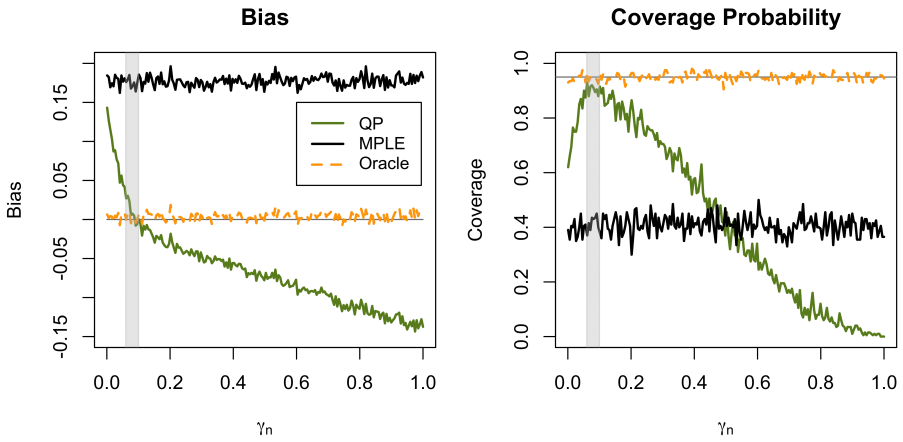

Figure 3.1: Estimation bias and 95% confidence interval coverage probability for $\beta_1^0 = 1$ with the tuning parameter $\gamma_n \in [0, 1]$ in a simulated example with $n = 500$ observations and $p = 100$ independent covariates. The methods in comparison include the proposed de-biased lasso with quadratic programming (QP), the maximum partial likelihood estimation (MPLE) and the oracle estimator (ORACLE).

approach change as γ_n ranges from 0 to 1. Intuitively, one should select γ_n within the shaded range to achieve desirable inference results.

The cross-validation criterion in *Cai et al.* (2011) for estimating the sparse inverse covariance matrix is inappropriate for our purpose, since it originates from the idea of maximizing the log likelihood for independent and identically distributed Gaussian random vectors. *Yu et al.* (2018) used 10-fold cross-validation to choose the γ_n that minimizes the criterion $\text{tr}(\text{diag}(\widehat{\Sigma}\widehat{\Theta} - I_p)^2)$, which is an alternative option given in the R package `clime` but still leaves large biases in the true signals in their simulation studies. *van Houwelingen et al.* (2006) proposed a cross-validated partial log-likelihood criterion base on leave-one-out estimates, which has been modified for K-fold cross-validation and implemented in the R package `glmnet` for Cox model. However, our simulation shows that using this criterion for de-biased lasso tends to select the largest possible γ_n and makes no bias correction, which is intuitive since the same criterion is adopted by `glmnet` to obtain the lasso estimator and facilitates relatively stable prediction rather than bias correction and inference.

To achieve sufficient bias correction and reliable inference, a desirable tuning pa-

rameter γ_n should be close to 0 and results in a de-biased estimator fitting the data well. Evaluation of a cross-validation criterion value by directly plugging in the de-biased estimator is highly discouraged, as estimation error from every component of the de-biased estimator can accumulate and mounts to severe inaccuracy issues. Thus, we propose to select γ_n following Algorithm 3.1, based on an idea of "active de-biased lasso estimator". Step 2.2 represents a de-noising step for the plug-in estimator, which is effective in practice. One may also repeat the cross-validation for multiple times and minimize over the average cv_g in Step 3, to reduce instability due to random data splitting.

Algorithm 3.1 Selection of the tuning parameter γ_n using cross-validation

Step 1 Pre-determine a grid of points for γ_n in [0,1], denoted as $\gamma_n^{(g)}, g = 1, \cdots, G$, and set each $cv_g = 0$.

Step 2 Randomly split the dataset into K folds of the same size, and at each time leave one part for testing and the others for training. For $k = 1, \cdots, K$:

Step 2.1 Use the kth training data to compute the de-biased lasso estimator with $\gamma_n^{(g)}, g = 1, \cdots, G$, denoted as $\widehat{b}^{(gk)}$.

Step 2.2 Define the active de-biased lasso estimator $\widehat{b}^{(gk)}_{active} = \widehat{b}^{(gk)} \cdot 1(j \in \widehat{A})$, i.e. setting components of $\widehat{b}^{(gk)}$ that are not in the set $\widehat{A}$ to 0. The active set $\widehat{A}$ can be obtained by retaining the variables that pass the multiple testing (e.g. Bonferroni correction) thresholds based on Theorem III.1.

Step 2.3 Compute the negative log-likelihood on the kth testing dataset with $\widehat{b}^{(gk)}_{active}$, $\ell^{(k)}(\widehat{b}^{(gk)}_{active})$.

Step 2.4 Set $cv_g \leftarrow cv_g + \ell^{(k)}(\widehat{b}^{(gk)}_{active})$, for $g = 1, \cdots, G$.

Step 3 Let $\hat{g} = \arg\min_g cv_g$. The final output tuning parameter value is $\gamma_n^{(\hat{g})}$.

3.3 Theoretical results

In this section we study the asymptotic properties of the de-biased estimator $\widehat{b}$. Let $\mu_r(t;\beta) = E[1(Y \geq t)X^{\otimes r}\exp\{X^T\beta\}]$ be the expectation of $\widehat{\mu}_r(t;\beta)$. For theoretical purpose, we define population-level counterparts for the weighted average covariates as

$$\eta_0(t;\beta) = \frac{\mu_1(t;\beta)}{\mu_0(t;\beta)} = \frac{E[1(Y \geq t)\exp\{X^T\beta\}X]}{E[1(Y \geq t)\exp\{X^T\beta\}]},$$

and for the random matrix $\widehat{\Sigma}$ as

$$\Sigma_{\beta^0} = E\left[\{X - \eta_0(Y;\beta^0)\}^{\otimes 2}\delta\right].$$

Take the inverse matrix $\Theta_{\beta^0} = \Sigma_{\beta^0}^{-1}$. The first order Taylor expansion of $\dot{\ell}_{nj}(\widehat{\beta})$, the jth component in $\dot{\ell}_n(\widehat{\beta})$, at β^0, is

$$\dot{\ell}_{nj}(\widehat{\beta}) = \dot{\ell}_{nj}(\beta^0) + [\ddot{\ell}_{nj}(\widetilde{\beta}^{(j)})]^T(\widehat{\beta} - \beta^0), \tag{3.6}$$

where $\widetilde{\beta}^{(j)}$ lies between $\widehat{\beta}$ 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 $B_n = (\ddot{\ell}_{n1}(\widetilde{\beta}^{(1)}), \ldots, \ddot{\ell}_{np}(\widetilde{\beta}^{(p)}))^T$. Suppose $c \in \mathbb{R}^p$ is a p-dimensional vector, and the parameter of interest is $c^T\beta^0$. Plugging (3.6) in (3.5), we have

$$\begin{aligned} c^T(\widehat{b} - \beta^0) = &-c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0) - c^T(\widehat{\Theta} - \Theta_{\beta^0})\dot{\ell}_n(\beta^0) \\ &- c^T(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0) + c^T\widehat{\Theta}(\widehat{\Sigma} - B_n)(\widehat{\beta} - \beta^0). \end{aligned} \tag{3.7}$$

The first term in (3.7) is the leading part, and the others will be proved to be asymptotically negligible.

We make several assumptions to establish the theoretical properties of the debiased lasso estimator.

(A1) Covariates are almost surely uniformly bounded, i.e. $\|X_i\|_\infty \leq K$ for some constant $K > 0$ for $i = 1, 2, \ldots, n$.

(A2) $|X_i^T\beta^0| \leq K_1$ uniformly for all $i = 1, \cdots, n$ with some constant $K_1 > 0$, almost surely.

(A3) The follow-up time stops at a finite time point $\tau > 0$, with probability $\pi_0 = \mathbb{P}(Y \geq \tau) > 0$.

(A4) Let

$$\widetilde{\Sigma}_{\beta^0}(t) = \int_0^t \left\{ \mu_2(u;\beta^0) - \frac{\mu_1(u;\beta^0)\mu_1^T(u;\beta^0)}{\mu_0(u;\beta^0)} \right\} d\Lambda_0(u).$$

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

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

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

(A5) The eigenvalues of Σ_{β^0} are bounded, i.e. there exist two constants $\lambda_{\min}$ and $\lambda_{\max}$ such that $0 < \lambda_{\min} \leq \lambda_{\min}(\Sigma_{\beta^0}) \leq \lambda_{\max}(\Sigma_{\beta^0}) \leq \lambda_{\max} < \infty$, where $\lambda_{\min}(\Sigma_{\beta^0})$ and $\lambda_{\max}(\Sigma_{\beta^0})$ are 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 (A1). *Fang et al.* (2017) and *Kong et al.* (2018) also posed (A2) for the Cox model, i.e. uniform boundedness on the multiplicative hazard. Under (A1), (A2) can be implied by bounded overall signal $\|\beta^0\|_1$. (A3) is usually used for survival models with censored data (*Andersen and Gill*, 1982). (A4) ensures the covergence of a predictable variation process in the martingale central limit theorem and thus the asymptotic normality of the de-biased lasso estimator. $\widetilde{\Sigma}_{\beta^0}(t)$ can be viewed as the information matrix up to time t. It is easy to see that $\widetilde{\Sigma}_{\beta^0}(\tau) = \Sigma_{\beta^0}$ and $v(\tau) = 1$. The limiting function $v(t)$ also depends on $c \in \mathbb{R}^p$, the linear combination vector of interest. (A4) is an alternative assumption to the stringent boundedness condition on $\|\Theta_{\beta^0} X_i\|_\infty$, which was essential in *van de Geer et al.* (2014) for statistical inference in high-dimensional generalized linear models and in *Fang et al.* (2017) for Cox model. The bounded eigenvalue condition on Σ_{β^0}, (A5), is standard in inference for high-dimensional models. Since we focus on random designs, unlike *Huang et al.* (2013) and *Yu et al.* (2018), we do not directly assume the compatability condition on $\ddot{\ell}_n(\beta^0)$.

The following theorem establishes the asymptotic distribution for linear combinations of the resulting de-biased lasso estimator $\widehat{b}$.

Theorem III.1. *Assume that* $\lambda \asymp \sqrt{\log(p)/n}$ *and* $\|\Theta_{\beta^0}\|_{1,1}^2 p s_0 \log(p)/\sqrt{n} \to 0$ *as* $n \to \infty$. *Under assumptions (A1) – (A5), for any* $c \in \mathbb{R}^p$ *such that* $\|c\|_2 = 1$ *and* $\|c\|_1 \leq a_*$ *with some absolute constant* $a_* > 0$, *we have*

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

Theorem III.1 provides the foundation for drawing inference on the regression parameters. Suppose one is interested in testing $H_0 : c^T \beta^0 = a_0$ versus the alternative $H_1 : c^T \beta^0 \neq a_0$ for some known $c \in \mathbb{R}^p$ and constant a_0. Based on the test statistic $T = \sqrt{n}(c^T \widehat{b} - a_0)/(c^T \widehat{\Theta} c)^{1/2}$, 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 the standard normal distribution. Corollary III.2 discusses the type I error and the power of the test $\phi(T)$, and Corollary III.3 ensures that the confidence interval constructed based on Theorem III.1 achieves the nominal coverage probability asymptotically.

Corollary III.2. *Suppose that the conditions in Theorem III.1 hold. Then, for the test* $\phi(T)$, *the type I error rate* $\mathrm{pr}(\phi(T) = 1 | H_0) \to \alpha$ *as* $n \to \infty$, *and, under the truth* β^0, *the power function* $\mathrm{pr}(\phi(T) = 1) \to 1 - \Phi(z_{\alpha/2} + (a_0 - c^T \beta^0)/(c^T \Theta_{\beta^0} c)^{1/2}) + \Phi(-z_{\alpha/2} + (a_0 - c^T \beta^0)/(c^T \Theta_{\beta^0} c)^{1/2})$, *where* $\Phi(\cdot)$ *is the cumulative density function for standard normal distribution.*

Corollary III.3. *Suppose that the conditions in Theorem III.1 hold. The level* α *confidence interval for* $c^T \beta^0$ *is constructed as* $J(\alpha) = [c^T \widehat{b} - z_{\alpha/2}(c^T \widehat{\Theta} c/n)^{1/2},\ c^T \widehat{b} + z_{\alpha/2}(c^T \widehat{\Theta} c/n)^{1/2}]$. *Then* $\mathrm{pr}(c^T \beta^0 \in J(\alpha)) \to 1 - \alpha$, *as* $n \to \infty$.

Theorem III.1 lays the foundation for inference on a single linear combination, $c^T\beta^0$. In fact, based on the results in Theorem III.1 and the Cramér-Wold device, we can also justify simultaneous inference on multiple linear combinations, that is $A\beta^0$ for some $l \times p$ matrix A, as summarized in Theorem III.4. Parallel to the corollaries above, Corollary III.5 provides the theoretical results for hypothesis testing and confidence region in this setting. Since the assumption (A4) is dependent on the combination vector $c \in \mathbb{R}^p$, we need a variation of (A4) for Theorem III.4 to hold.

(A4)′ $\widetilde{\Sigma}_{\beta^0}(t)$ is the same as defined in (A4). For the combination matrix of interest $A \in \mathbb{R}^{l\times p}$ and any vector $\omega \in \mathbb{R}^l$, it holds that

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

for any $t \in [0, \tau]$ and some fixed function $v'(\cdot) > 0$.

Note that in the above alternative assumption (A4)′, $v'(\cdot)$ is specific to A and ω.

Theorem III.4. *Suppose that the $l \times p$ matrix A has full row rank, 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. Also, assume the conditions in Theorem III.1 hold with (A4) replaced by (A4)′. Take $\lambda \asymp \sqrt{\log(p)/n}$ for the lasso estimator. Then we have*

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

Corollary III.5. *Suppose the conditions in Theorem III.4 hold. For the $l \times p$ matrix A in Theorem III.4, under the null hypothesis $H_0 : A\beta^0 = a_0$ ($a_0 \in \mathbb{R}^l$), the statistic $T' = n(A\widehat{b} - a^0)^T\widehat{F}^{-1}(A\widehat{b} - a^0) \xrightarrow{\mathcal{D}} \chi^2_l$, where $\widehat{F} = A\widehat{\Theta}A^T$. For $\alpha \in (0, 1)$, let the confidence region for $A\beta^0$ be $J'(\alpha) = \{a \in \mathbb{R}^l : n(A\widehat{b}-a)^T\widehat{F}^{-1}(A\widehat{b}-a) \le \chi^2_{l,\alpha}\}$, where $\chi^2_{l,\alpha}$ is the upper α-th quantile from χ^2_l. Then $\mathrm{pr}(A\beta^0 \in J'(\alpha)) \to 1 - \alpha$, as $n \to \infty$.*

3.4 Numerical experiments

We simulate $n = 500$ observations with $p = 20, 100, 200$ covariates. The covariates X are first simulated from two settings, $N(0, I_p)$ and AR(1) with correlation 0.5, and then truncated at $|X^{(j)}| \leq 2.5,\ j = 1, \ldots, p$. In the true regression coefficients β^0, the first element β_1^0 varies from 0 to 2 by increment 0.2, four of the rest are arbitrarily chosen to take values of 1, 1, 0.5 and 0.5, and all others are zero. The underlying survival time T follows an exponential distribution with hazard $\lambda(t|X) = \exp\{X^T\beta^0\}$, and the censoring time C, independent of T, follows Uniform$(1, 20)$. We monitor the estimation bias for β_1^0, its model-based standard error, coverage probability at significance level $\alpha = 0.05$ and mean squared error. The methods in comparison include: (1) our proposed de-biased lasso with quadratic programming for matrix $\widehat{\Theta}$ (QP), (2) the de-biased lasso with node-wise lasso for matrix $\widehat{\Theta}$ (NW) in *Kong et al.* (2018), (3) the de-biased lasso with CLIME for matrix $\widehat{\Theta}$ (CLIME) in *Yu et al.* (2018), (4) the decorrelated Wald test (DECOR) in *Fang et al.* (2017) and (5) the oracle estimator as if the true model were known (ORACLE). Note that we use the tuning parameter selection procedure described in Algorithm 3.1 of Section 3.2.3 for QP.

For the lasso estimator, we use 10-fold cross-validation to select the tuning parameter λ. 5-fold cross-validation is used for tuning parameter selection in CLIME, QP and NW. For the active set in Step 2.2 of Algorithm 3.1, we adopt the Bonferroni correction with the adjusted p-value threshold $0.1/p$, where p is the number of covariates.

Figures 3.2 and 3.3 show the simulation results for the independent and the AR(1) covariance structures, respectively. When the dimension $p = 20$, our proposed method QP and the decorrelated Wald test DECOR have almost identical performance to the oracle estimator ORACLE. When the dimension is relatively large compared to the sample size, i.e. $p = 100, 200$, the proposed estimator QP displays the smallest estimation biases and the confidence interval coverage probabilities closest to the nominal

level 95% in both cases. CLIME and NW both suffer from insufficient bias correction due to penalized estimation for the matrix Θ_{β^0}, and thus have severe under-coverage problems with the confidence intervals. Compared to the independent covariance case, the proposed method QP performs worse in the AR(1) covariance case on bias correction and confidence interval coverage, but is still the best in all methods considered, especially when p is large.

We also recorded the average computational time spent on computing $\widehat{\Theta}$ only (Table 3.1), comparing the `R` functions `solve.QP` in the package `quadprog` for the proposed quadratic programing procedure, `clime` in the package `clime` and `sugm` in the package `flare` for CLIME. All data were included without cross-validation, and three candidate values of γ_n, which were 0.3, 1 and 2 times of $\sqrt{\log(p)/n}$, were used for demonstration. We fixed $\beta_1^0 = 1$ and simulated $n = 500$ observations. The covariates had AR(1) covariance structure. The other settings were the same as those introduced above. The time columns in Table 3.1 were generated on a MacBook with 2.7GHz Intel Core i5 processor and 8GB memory, and averaged over 10 replications. Time ratio of each implementation compared to `solve.QP` was also reported in each setup. Our proposed implementation with `solve.QP` is the most computationally efficient among the three. For a large dimension, `clime` took the longest time in general.

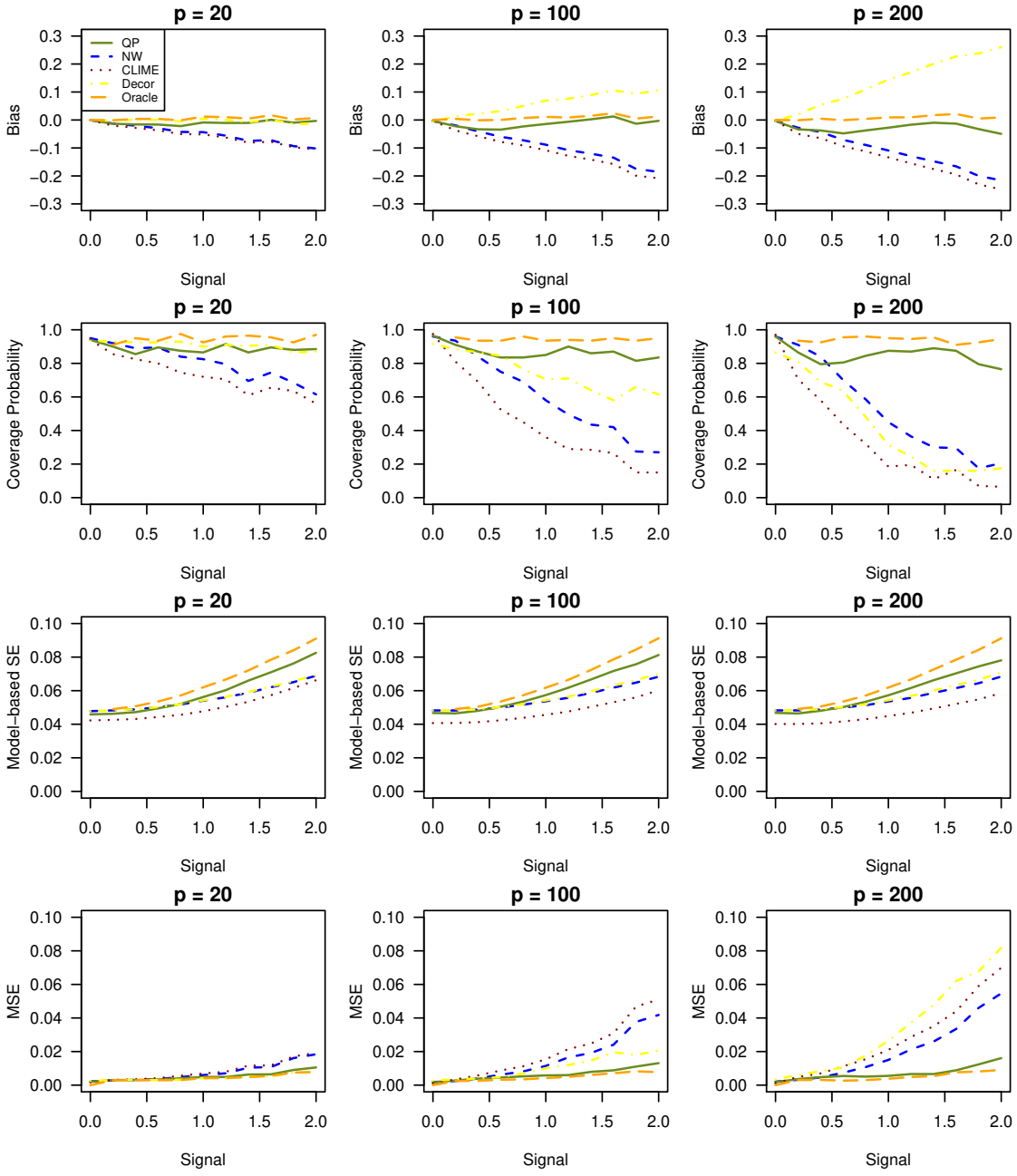

Figure 3.2: Estimation bias, coverage probability, model-based standard error and mean squared error for the five estimators under comparison, QP, NW, CLIME, DECOR and ORACLE, in the simulation with $n = 500$ observations and independent covariance structure for covariates.

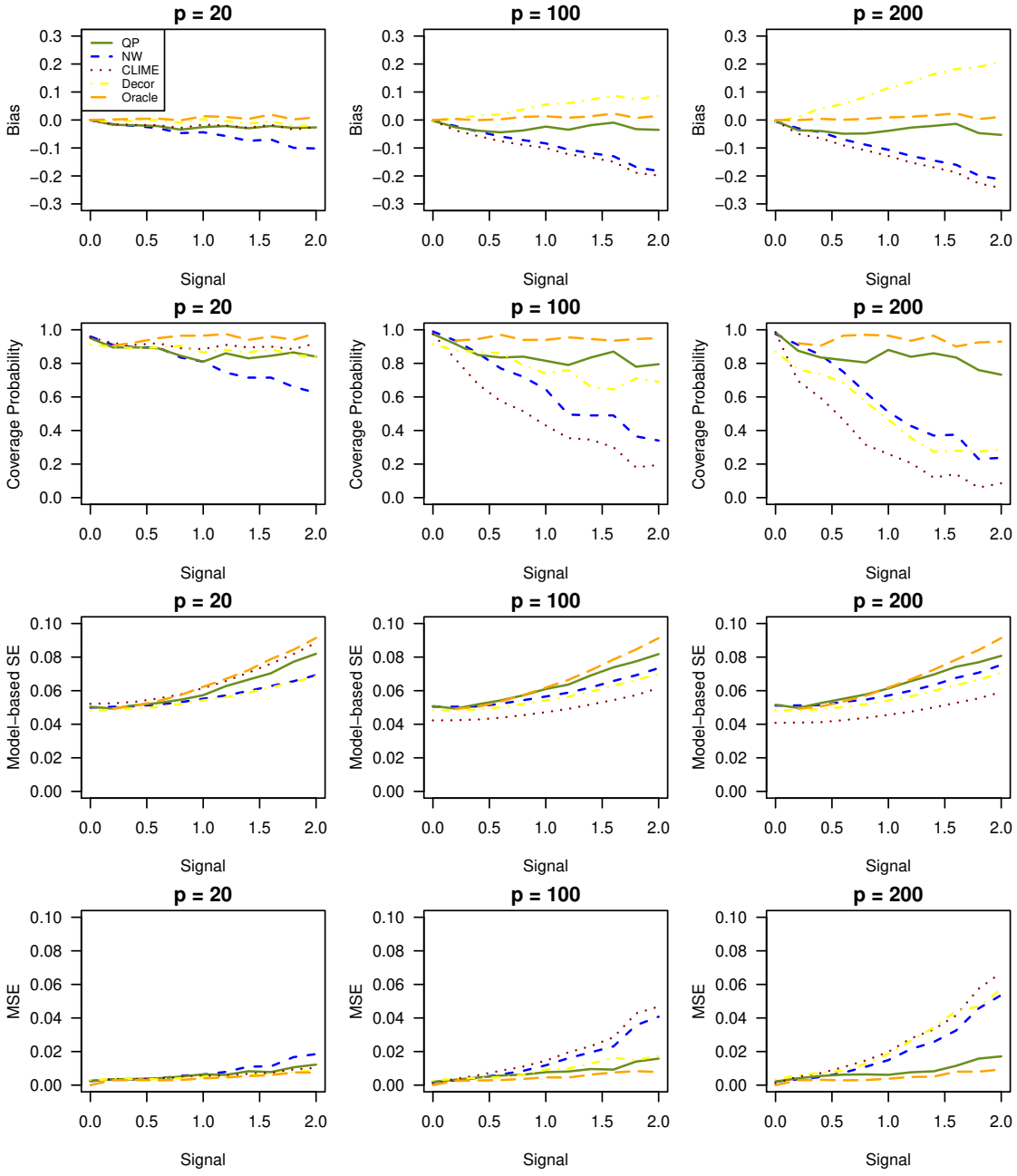

Figure 3.3: Estimation bias, coverage probability, model-based standard error and mean squared error for the five estimators under comparison, QP, NW, CLIME, DECOR and ORACLE, in the simulation with $n = 500$ observations and AR(1) covariance structure for covariates ($\rho = 0.5$).

Table 3.1: Comparison of the computational time spent on computing $\widehat{\Theta}$. Time (in seconds) is averaged over 10 replications under each setting. Time ratio is with respect to the proposed method implemented using `solve.QP`.

$p = 20$	solve.QP		clime		flare	
	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
$p = 100$	solve.QP		clime		flare	
	Time	Ratio	Time	Ratio	Time	Ratio
$\gamma_n = 0.3\sqrt{\log(p)/n}$	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
$p = 200$	solve.QP		clime		flare	
	Time	Ratio	Time	Ratio	Time	Ratio
$\gamma_n = 0.3\sqrt{\log(p)/n}$	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

3.5 Application to the Boston Lung Cancer Study

Lung cancer is the leading cause of cancer deaths in the United States, for both men and women. Non-small cell lung cancer (NSCLC) accounts for approximately 80% to 85% among all the lung cancer cases and is the most common histological type of lung cancer (*Houston et al.*, 2018). Identification of genetic variants associated with lung cancer patient survival is of great interest in modern translational cancer research, which has the potential to refine prognosis and promote individualized decision making on treatment and clinical care. Despite a large number of 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 and Amos*, 2018). The Boston Lung Cancer Study (BLCS) is a large epidemiology cohort for investigating the molecular cause underlying lung cancer, where lung cancer cases

have been enrolled at Massachusetts General Hospital and the Dana-Farber Cancer Institute from 1992 to present. We applied the proposed de-biased lasso method to a subset of the BLCS data and simultaneously investigated the joint effects of certain genotyped SNPs on NSCLC patient overall survival.

The subset of data in this analysis consisted of $n = 561$ NSCLC patients with their diagnosis dates, follow-up times were available and genotypes on Axiom array available. Among all these patients, 437 (77.9%) were observed deaths and 124 (22.1%) were censored. The longest observed survival time was 8584 days and the shortest was 6 days. The restricted mean survival and censoring times at $\tau = 8584$ days were 2124 (SE: 105) and 4397 (SE: 187) days, respectively. Patient characteristics adjusted in the Cox proportional hazards model, including age at diagnosis, race, education level, gender, smoking status, histological type, cancer stage, treatment indicators and an indicator for missing treatment information, are summarized in Table 3.2.

With the 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 focused on 32 genes in the CARM ER pathway, which was the largest pathway *Tang et al.* (2020) considered and described in their supplementary document and contained the two reported genes *HDAC2* and *PPARGC1A*. It was also of great interest to investigate whether the susceptibility loci studied in Chapter II were associated with patient survival. We extracted 312 genotyped SNPs from the 32 genes in the CARM ER pathway and the nine target genes in Section 2.4.3 in the BLCS data (minor allele frequency > 0.01, genotype call rate $> 95\%$). After we implemented a pruning step to avoid including many SNPs in high linkage disequilibrium using PLINK (*Purcell et al.*, 2007) (window size 50, step size 5, and $r^2 > 0.7$), the number of SNPs was reduced to 217. SNPs were coded by the number of copies of the minor allele, i.e. 0, 1 or 2, and were assumed to have additive effects on the log hazard

Table 3.2: Characteristics of $n = 561$ patients in the Boston Lung Cancer Study for survival analysis

Variable	Category / Unit	Count (%) / Mean (SD)
Age	Years old	60.0 (10.9)
Race	Caucasian	528 (94.1%)
	Others	33 (5.9%)
Education	No high school	79 (14.1%)
	High school	141 (25.1%)
	At least 1-2 years of college	341 (60.8%)
Gender	Male	215 (38.3%)
	Female	346 (61.7%)
Smoker	Current or recently quit	508 (90.6%)
	Never	53 (9.4%)
Histology	Adenocarcinoma	360 (64.2%)
	Squamous cell carcinoma	115 (20.5%)
	Large cell carcinoma	45 (8.0%)
	Unspecified	41 (7.3%)
Stage[a]	Early	243 (43.3%)
	Late	318 (56.7%)
Surgery	No	177 (31.6%)
	Yes	361 (64.3%)
Chemotherapy	No	300 (53.5%)
	Yes	238 (42.4%)
Radiation	No	332 (59.2%)
	Yes	206 (36.7%)
Treatment record	Missing[b]	23 (4.1%)

[a] Stages I and II classified as early stage, and stages III and IV as late stage.
[b] No treatment information on surgery, chemotherapy or radiation available for these patients.

ratio. Therefore, the subset of the BLCS data we analyzed included $n = 561$ NSCLC patients and $p = 231$ covariates.

Table 3.3 summarizes the coefficient estimates in the Cox proportional hazards model, for all patient characteristics and the top ten SNPs ranked by QP p-values. In general, QP results in 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 than MPLE. As a result, the de-biased lasso estimator is also numerically more stable, and has narrower confidence intervals since the standard errors are not estimated using the inverted Hessian matrix. In fact, *van de Geer et al.* (2014) proved in that the de-biased lasso in linear regression is semi-parametrically efficient. When the dimension p is very small, the difference between the two methods would be negligible.

Among patient characteristics, QP found that the adenocarcinoma subtype is associated with better patient survival compared to large cell carcinoma, consistent with previous findings (*Janssen-Heijnen and Coebergh*, 2001), and MPLE has failed to identify such an association. AX-11672686 in *CHRNA2*, AX-11673610 in *GRIP2* and AX-11264571 in *BRCA2* are the three most significant SNPs associated with NSCLC patient survival identified by QP. Interestingly, significant associations were found between AX-11672686 and nicotine dependence (*Wang et al.*, 2014). *GRIP1* has not been reported in the literature review (*Bossé and Amos*, 2018). *Hershberger et al.* (2005) showed that "NSCLC cells express proteins necessary to generate a transcriptional response to estrogen and suggest that ERβ and GRIP1 are likely mediators of this response". While AX-11264571 has been found to be associated with breast cancer (*Qiu et al.*, 2010), it may also be associated with lung cancer susceptibility although not achieving genome-wide significance in *Yu et al.* (2011). Our proposed method QP also identified four other SNPs with non-adjusted 95% CIs excluding zero, two of which MPLE did not identify. Two SNPs signaled by QP at level 0.05 are located in *CREBBP*, which is one of the most frequently mutated genes in small cell lung cancer (*Jia et al.*, 2018).

We also tested for the association between education and patient survival. Let $A_{2\times p} = (e_2, e_3)^T$ indicate the two parameters of interest, corresponding to the effects of high school graduate and at least 1–2 years of college compared to the reference level of no high school. Then, the test statistic for a hypothesis of no education effect

is 0.259. The resulting p-value is $\mathbb{P}(\chi^2_2 > 0.259) = 0.879$. Thus, based on the data, we did not have significant evidence to claim the association between education and NSCLC patient survival after adjusting all other existing characteristics and SNPs. These results show that our method can be useful in providing reliable inference for scientific discovery, validation and interpretation, even though the actual functions of genetic variants would need further biological investigations.

Table 3.3: Coefficient estimates in the Cox proportional hazards model for the Boston Lung Cancer Study data

			QP				MPLE			
Variable	Note		Est	SE	P-value	95% CI	Est	SE	P-value	95% CI
Race	Others vs Caucasian		-0.163	0.201	0.416	(-0.557, 0.231)	0.065	0.561	0.908	(-1.034, 1.163)
Education	HS vs No HS		-0.018	0.091	0.840	(-0.198, 0.161)	-0.142	0.253	0.574	(-0.637, 0.353)
	College vs No HS		-0.037	0.076	0.625	(-0.185, 0.111)	-0.085	0.218	0.698	(-0.513, 0.343)
Gender	Male vs 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 vs No		0.103	0.142	0.470	(-0.176, 0.381)	0.066	0.299	0.825	(-0.519, 0.651)
Histology	AD vs LCC		-0.259	0.076	0.001	(-0.409, -0.11)	-0.467	0.294	0.112	(-1.043, 0.109)
	SCC vs LCC		0.065	0.094	0.488	(-0.120, 0.251)	-0.030	0.314	0.923	(-0.646, 0.585)
	Unspecified vs LCC		0.046	0.132	0.729	(-0.213, 0.304)	-0.119	0.384	0.756	(-0.871, 0.633)
Stage	Late vs Early		0.352	0.081	< 0.001	(0.193, 0.510)	0.553	0.190	0.004	(0.180, 0.926)
Surgery	Yes vs No		-1.102	0.085	< 0.001	(-1.269, -0.936)	-2.115	0.226	< 0.001	(-2.557, -1.672)
Chemotherapy	Yes vs No		0.025	0.078	0.753	(-0.128, 0.177)	-0.239	0.220	0.278	(-0.671, 0.193)
Radiation	Yes vs No		0.047	0.077	0.548	(-0.105, 0.198)	0.248	0.198	0.211	(-0.140, 0.636)
Treatment record	Missing vs 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)
⋮										

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

3.6 Concluding remarks

Motivated by the work of *Javanmard and Montanari* (2014), we have proposed a de-biased lasso approach for reliable estimation and inference in Cox proportional hazards model when the number of covariates p is allowed to diverge with the sample size n. The proposed de-biased lasso estimator has been proven asymptotically unbiased and normally distributed under certain mild regularity conditions. Unlike existing methods in *Fang et al.* (2017); *Yu et al.* (2018); *Kong et al.* (2018), we do not require a sparse estimation for the inverse information matrix Θ_{β^0} under unrealistic sparsity assumption about Θ_{β^0}, by exploiting the quadratic programming procedure. We have shown that the proposed de-biased lasso estimator performs better than its competitors in terms of bias correction and reliable confidence interval coverage.

Although we have only considered the "large n, diverging p" scenario where $p < n$ in this chapter, the same approach can be used for reliable inference in the more challenging "large p, small n" scenario where p can be much larger than n. With a slight variation in the implementation, one may replace $\widehat{\Sigma}$ with $\widehat{\Sigma} + H$ for some diagonal matrix $H = \text{diag}(h_1, \cdots, h_p)$, $h_j > 0$, in the quadratic programming problems to stablize the estimates. It would make an interesting future direction to prove the theoretical validity in this challenging scenario. From the simulation, we see that the proposed method requires careful treatment of the tuning parameter selection. Even though it outperforms all competitors in the simulation, other procedures for selecting a proper tuning parameter γ_n may be investigated in order to further improve its performance.

Sometimes, it may be of interest not to penalize and select certain regression parameters, e.g. for treatment or exposure effects. Lasso, whether putting penalties on these variables or not, can be viewed as a useful tool to provide a more stable initial estimator than the maximum (partial) likelihood estimation. From the theoretical perspective, there will be no difference in the convergence rates of the lasso

Table 3.4: Comparison between penalizing and not penalizing β_1, for estimating $\beta_1^0 = 0.5$

	$n = 250$		$n = 500$		$n = 1000$	
β_1 Penalized	Yes	No	Yes	No	Yes	No
Lasso	0.378	0.524	0.409	0.507	0.440	0.509
QP Est	0.468	0.516	0.477	0.499	0.492	0.503
QP SE	0.071	0.071	0.052	0.053	0.038	0.038
QP Cov	0.805	0.870	0.875	0.915	0.885	0.910
QP EmpSE	0.094	0.093	0.059	0.059	0.044	0.044
QP MSE	0.010	0.009	0.004	0.003	0.002	0.002

estimator itself and the excess loss, if we do not penalize a fixed number of coefficients (*Bühlmann and van de Geer*, 2011). The de-biased lasso provides inference for all regression parameters, and no parameters are estimated exactly at zero. The theoretical results for the de-biased lasso will remain the same even when these variables are left in the model unpenalized. Computationally, having some coefficients unpenalized in the lasso can be easily implemented as a special case of the weighted lasso, where the weights for the corresponding coefficients are specified as zero. Not penalizing a subset of coefficients will result in numerical differences in the estimates. We herein provide a simulated example in the Cox model, with 200 replications. The sample size n varies between 250 and 1000, and the dimension $p = 100$. The covariate vector X is simulated from a multivariate normal distribution with mean zero and AR(1) covariance ($\rho = 0.5$). Suppose we are interested in studying the difference when the first covariate X_1 is forced into the model without penalization. $\beta_1^0 = \beta_2^0 = 0.5$, and two other coefficients are fixed at 0.5 and 1, respectively. Other settings remain the same as introduced in Section 3.4. Table 3.4 and Table 3.5 summarize the results for estimating β_1^0 and β_2^0 respectively. As the sample size increases, the difference between penalizing and not penalizing β_1 diminishes. If we do not penalize β_1, the de-biased lasso estimate for β_1^0 is very close to the lasso estimator (Table 3.4). Due to the correlation between X_1 and X_2, not penalizing β_1 results in slightly more bias for estimating β_2^0 (Table 3.5).

Table 3.5: Comparison between penalizing and not penalizing β_1, for estimating $\beta_2^0 = 0.5$

	$n = 250$		$n = 500$		$n = 1000$	
β_1 Penalized	Yes	No	Yes	No	Yes	No
Lasso	0.392	0.325	0.421	0.378	0.434	0.404
QP Est	0.489	0.448	0.493	0.475	0.488	0.478
QP SE	0.077	0.076	0.057	0.057	0.041	0.041
QP Cov	0.885	0.785	0.910	0.885	0.910	0.870
QP EmpSE	0.092	0.103	0.063	0.064	0.043	0.044
QP MSE	0.009	0.013	0.004	0.005	0.002	0.002

3.7 Technical proofs

We first present the lemmas below, along with their proofs, that will be used to prove Theorem III.1 and Theorem III.4. Some of these lemmas are interesting results by themselves. Additional notation from counting processes and martingale theory is defined here for the proofs. Under the Cox model, define the counting process $N_i(t) = 1(Y_i \leq t, \delta_i = 1)$ and the intensity process $A_i(t;\beta) = \int_0^t 1(Y_i \geq s)\exp(X_i^T\beta)d\Lambda_0(s)$, where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the baseline cumulative 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 $\mathcal{F}_i(t) = \sigma\{N_i(s), 1(Y_i \geq s), X_i : s \in (0,t]\}$. $\{X_i - \widehat{\eta}_n(t;\beta^0)\}$ is predictable with respect to the filtration $\mathcal{F}(t) = \sigma\{N_i(s), 1(Y_i \geq s), X_i : s \in (0,t], i = 1, \cdots, n\}$. Notationally, we will not distinguish between the usual expectation and the outer expectation, and all conclusions still hold regardless of measurability.

Lemma III.6 below characterizes the difference between $\widehat{\eta}_n(t;\beta^0)$ and $\eta_0(t;\beta^0)$, which is useful to prove the asymptotic distribution for the leading term $\sqrt{n}c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0)$ as well as to establish the convergence rate for $\widehat{\Sigma} - \Sigma_{\beta^0}$.

Lemma III.6. *Under assumptions (A1) – (A3), we have*

$$\sup_{t\in[0,\tau]} |\widehat{\mu}_0(t;\beta^0) - \mu_0(t;\beta^0)| = \mathcal{O}_P(\sqrt{\log(p)/n}),$$
$$\sup_{t\in[0,\tau]} \|\widehat{\mu}_1(t;\beta^0) - \mu_1(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}),$$
$$\sup_{t\in[0,\tau]} \|\widehat{\eta}_n(t;\beta^0) - \eta_0(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

Proof of Lemma III.6. The first two statements in the conclusion are similar to those in *Kong and Nan* (2014), where the setups are different. Consider a class of functions of $y \geq 0$ and $x \in \mathbb{R}^p$ indexed by t, $\mathcal{F}_0 = \{1(y \geq t)\exp(x^T\beta^0) : t \in [0,\tau]\}$. For any $0 < \epsilon < 1$, consider the cumulative distribution function for Y and take an positive integer $m < 2/\epsilon$ and a sequence of points $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = \infty$ such that $\mathbb{P}(t_i < Y \leq t_{i+1}) < \epsilon,\ i = 0, 1, \ldots, m-1$. For each $i = 1, \cdots, m$, define the bracket $[L_i, U_i]$, where $L_i(x,y) = 1(y \geq t_i)\exp(x^T\beta^0)$ and $U_i(x,y) = 1(y > t_{i-1})\exp(x^T\beta^0)$. We have $L_i(x,y) \leq 1(y \geq t)\exp(x^T\beta^0) \leq U_i(x,y)$ for $t_{i-1} < t \leq t_i$, and

$$[\mathbb{E}\{U_i(X,Y) - L_i(X,Y)\}^2]^{1/2} = [\mathbb{E}\{1(t_{i-1} < Y < t_i)\exp(2X^T\beta^0)\}]^{1/2} \leq e^{K_1}\sqrt{\epsilon},$$
$$\mathbb{E}|U_i(X,Y) - L_i(X,Y)| = \mathbb{E}\{1(t_{i-1} < Y < t_i)\exp(X^T\beta^0)\} \leq e^{K_1}\epsilon.$$

So the bracketing numbers, the definition of which can be found in *van der Vaart* (1998), satisfy

$$N_{[]}(e^{K_1}\sqrt{\epsilon}, \mathcal{F}_0, L_2(\mathbb{P})) \leq \frac{2}{\epsilon}, \quad N_{[]}(e^{K_1}\epsilon, \mathcal{F}_0, L_1(\mathbb{P})) \leq \frac{2}{\epsilon},$$

or equivalently,

$$N_{[]}(\epsilon, \mathcal{F}_0, L_2(\mathbb{P})) \leq \frac{2e^{2K_1}}{\epsilon^2}, \quad N_{[]}(\epsilon, \mathcal{F}_0, L_1(\mathbb{P})) \leq \frac{2e^{K_1}}{\epsilon} < \infty.$$

By Glivenko-Cantelli Theorem and Donsker Theorem (*van der Vaart*, 1998), the class $\mathcal{F}_0$ is $\mathbb{P}$-Glivenko-Cantelli and $\mathbb{P}$-Donsker. So $\sup_{t\in[0,\tau]} |\widehat{\mu}_0(t;\beta^0) - \mu_0(t;\beta^0)| \xrightarrow{a.s.} 0$, and moreover, by Theorem 2.14.9 of *van der Vaart and Wellner* (1996) with $V = 2$,

$$\mathbb{P}\left(\sqrt{n}\sup_{t\in[0,\tau]} |\widehat{\mu}_0(t;\beta^0) - \mu_0(t;\beta^0)| > s\right) \leq De^{-s^2},$$

for every $s > 0$ and some constant $D > 0$ only depending on K_1. Setting $s = \sqrt{2\log(p)}$ implies that

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

For the second statement, we consider classes of functions of $(x, y) = (x_1, \cdots, x_p, y)$ indexed by t,

$$\mathcal{F}_1^k = \{1(y \geq t)e^{x^T\beta^0}x_k : t \in [0,\tau]\},\ k = 1, \cdots, p.$$

Since $|e^{x^T\beta^0}x_k| \leq Ke^{K_1}$, similar to the argument above, we have

$$N_{[]}(\epsilon, \mathcal{F}_1^k, L_2(\mathbb{P})) \leq \left(\frac{\sqrt{2}e^{K_1}K}{\epsilon}\right)^2.$$

By Theorem 2.14.9 of *van der Vaart and Wellner* (1996) with $V = 2$, we have

$$\mathbb{P}\left(\sqrt{n}\sup_{t\in[0,\tau]} |\widehat{\mu}_{1k}(t;\beta^0) - \mu_{1k}(t;\beta^0)| > s\right) \leq D's^2e^{-2s^2} \leq D'e^{-1}e^{-s^2}$$

for every $s > 0$, where D' is a constant that only depends on K and K_1, and $\widehat{\mu}_{1k}$ and

μ_{1k} are the kth components of $\widehat{\mu}_1$ and μ_1, respectively. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{n}\sup_{t\in[0,\tau]}\|\widehat{\mu}_1(t;\beta^0)-\mu_1(t;\beta^0)\|_\infty > s\right)\\
&\leq \mathbb{P}\left(\bigcup_{k=1}^{p}\left\{\sqrt{n}\sup_{t\in[0,\tau]}|\widehat{\mu}_{1k}(t;\beta^0)-\mu_{1k}(t;\beta^0)| > s\right\}\right)\\
&\leq pD'e^{-s^2}.
\end{aligned}$$

For example, taking $s=\sqrt{2\log(p)}$ would complete the proof for $\sup_{t\in[0,\tau]}\|\widehat{\mu}_1(t;\beta^0)-\mu_1(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$.

Finally, we rewrite

$$\begin{aligned}
\widehat{\eta}_n(t;\beta^0)-\eta_0(t;\beta^0) &= \frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}-\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)}\\
&= \frac{\widehat{\mu}_1(t;\beta^0)}{\mu_0(t;\beta^0)}-\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)}+\frac{\widehat{\mu}_1(t;\beta^0)}{\mu_0(t;\beta^0)}\left(\frac{\mu_0(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}-1\right).
\end{aligned}$$

By assumptions (A1) – (A3), $\mu_0(t;\beta^0)\geq e^{-K_1}\pi_0 > 0$ and $\sup_{t\in[0,\tau]}\|\widehat{\mu}_1(t;\beta^0)\|_\infty = \mathcal{O}_P(1)$. Also, since

$$\inf_{t\in[0,\tau]}\widehat{\mu}_0(t;\beta^0)\geq\mu_0(t;\beta^0)-|\widehat{\mu}_0(t;\beta^0)-\mu_0(t;\beta^0)|\geq e^{-K_1}\pi_0-\sup_{t\in[0,\tau]}|\widehat{\mu}_0(t;\beta^0)-\mu_0(t;\beta^0)| > e^{-K_1}\frac{\pi_0}{2}$$

eventually almost surely, then

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left\|\frac{\widehat{\mu}_1(t;\beta^0)}{\mu_0(t;\beta^0)}\left(\frac{\mu_0(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}-1\right)\right\|_\infty\\
&\leq \sup_{t\in[0,\tau]}\left\|\frac{\widehat{\mu}_1(t;\beta^0)}{\mu_0(t;\beta^0)}\right\|_\infty\cdot\sup_{t\in[0,\tau]}\left|\frac{\mu_0(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}-1\right|\\
&\leq \mathcal{O}_P(1)\sup_{t\in[0,\tau]}\left|\mu_0(t;\beta^0)-\widehat{\mu}_0(t;\beta^0)\right| = \mathcal{O}_P(\sqrt{\log(p)/n}).
\end{aligned}$$

By the arguments above,

$$\begin{aligned}\sup_{t\in[0,\tau]}\|\widehat{\eta}_n(t;\beta^0)-\eta_0(t;\beta^0)\|_\infty &\le \sup_{t\in[0,\tau]}\left\|\frac{1}{\mu_0(t;\beta^0)}\left(\widehat{\mu}_1(t;\beta^0)-\mu_1(t;\beta^0)\right)\right\|_\infty \\ &\quad+\sup_{t\in[0,\tau]}\left\|\frac{\widehat{\mu}_1(t;\beta^0)}{\mu_0(t;\beta^0)}\left(\frac{\mu_0(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}-1\right)\right\|_\infty \\ &= \mathcal{O}_P(\sqrt{\log(p)/n}).\end{aligned}$$

□

Lemma III.7 establishes the asymptotic distribution for the leading term $-c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0)$ in the decomposition of $c^T(\widehat{b}-\beta^0)$, up to rescaling with the standard deviation.

Lemma III.7. *Assume $p^2\log(p)/n\to 0$. Under assumptions (A1) – (A5), for any $c\in\mathbb{R}^p$ such that $\|c\|_2=1$ and $\|c\|_1\le a_*$ with some absolute constant $a_*>0$,*

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

Proof of Lemma III.7. Switching notation to martingales, we rewrite

$$\begin{aligned}\frac{-\sqrt{n}c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0)}{\sqrt{c^T\Theta_{\beta^0}c}} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_i-\frac{\widehat{\mu}_1(Y_i;\beta^0)}{\widehat{\mu}_0(Y_i;\beta^0)}\right\}\delta_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_i-\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}dN_i(t) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_i-\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}dM_i(t).\end{aligned}$$

Let $Q_i(t)=\frac{1}{\sqrt{n}}\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_i-\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}$, $i=1,\dots,n$, which are predictable with respect to the filtration $\mathcal{F}$. Then

$$\frac{-\sqrt{n}c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0)}{\sqrt{c^T\Theta_{\beta^0}c}}=\sum_{i=1}^n\int_0^\tau Q_i(t)dM_i(t). \tag{3.8}$$

For any $t \in [0, \tau]$, let $U(t) = \sum_{i=1}^{n} \int_0^t Q_i(u) dM_i(u)$, whose predictable variation process is

$$\begin{aligned}
\langle U \rangle(t) &= \sum_{i=1}^{n} \int_0^t Q_i(u)^2 1(Y_i \geq u) e^{X_i^T \beta^0} d\Lambda_0(u) \\
&= \sum_{i=1}^{n} \int_0^t \frac{c^T \Theta_{\beta^0}}{c^T \Theta_{\beta^0} c} \left\{ X_i - \frac{\widehat{\mu}_1(t; \beta^0)}{\widehat{\mu}_0(t; \beta^0)} \right\}^{\otimes 2} \Theta_{\beta^0} c 1(Y_i \geq u) e^{X_i^T \beta^0} d\Lambda_0(u) \\
&= \frac{c^T \Theta_{\beta^0}}{c^T \Theta_{\beta^0} c} \left[\int_0^t \left\{ \widehat{\mu}_2(u; \beta^0) - \frac{\widehat{\mu}_1(u; \beta^0) \widehat{\mu}_1(u; \beta^0)^T}{\widehat{\mu}_0(u; \beta^0)} \right\} d\Lambda_0(u) \right] \Theta_{\beta^0} c
\end{aligned}$$

Similar to the proof in Lemma III.6, 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

$$\begin{aligned}
\left\| \int_0^t \{\mu_2(u; \beta^0) - \widehat{\mu}_2(u; \beta^0)\} \lambda_0(u) du \right\|_\infty &\leq \sup_{u \in [0,\tau]} \|\widehat{\mu}_2(u; \beta^0) - \mu_2(u; \beta^0)\|_\infty \int_0^\tau \lambda_0(u) du \\
&= \mathcal{O}_P(\sqrt{\log(p)/n}). \qquad (3.9)
\end{aligned}$$

Since

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

by (A1) and Lemma III.6,

$$\left\| \int_0^t \left\{ \frac{\widehat{\mu}_1(u; \beta^0) \widehat{\mu}_1^T(u; \beta^0)}{\widehat{\mu}_0(u; \beta^0)} - \frac{\mu_1(u; \beta^0) \mu_1^T(u; \beta^0)}{\mu_0(u; \beta^0)} \right\} \lambda_0(u) du \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}). \qquad (3.10)$$

Combining (3.9) and (3.10), we see that uniformly for all $t \in [0, \tau]$,

$$\left\| \int_0^t \left\{ \widehat{\mu}_2(u;\beta^0) - \frac{\widehat{\mu}_1(u;\beta^0)\widehat{\mu}_1(u;\beta^0)^T}{\widehat{\mu}_0(u;\beta^0)} \right\} d\Lambda_0(u) - \right.$$
$$\left. \int_0^t \left\{ \mu_2(u;\beta^0) - \frac{\mu_1(u;\beta^0)\mu_1(u;\beta^0)^T}{\mu_0(u;\beta^0)} \right\} d\Lambda_0(u) \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

Then

$$\left| \langle U \rangle(t) - \frac{c^T \Theta_{\beta^0}}{c^T \Theta_{\beta^0} c} \left[\int_0^t \left\{ \mu_2(u;\beta^0) - \frac{\mu_1(u;\beta^0)\mu_1(u;\beta^0)^T}{\mu_0(u;\beta^0)} \right\} d\Lambda_0(u) \right] \Theta_{\beta^0} c \right|$$
$$\leq \lambda_{\min}^{-1} (\|c\|_1 \|\Theta_{\beta^0}\|_{1,1})^2 \mathcal{O}_P(\sqrt{\log(p)/n})$$
$$\leq \lambda_{\min}^{-1} a_*^2 p \lambda_{\max}^2 \mathcal{O}_P(\sqrt{\log(p)/n}) \to_P 0$$

as long as $p^2 \log(p)/n \to 0$. By (A4), $\langle U(t) \rangle \to_P v(t)$.

Now we check the Lindeberg condition. For any $\epsilon > 0$, define the truncated process

$$U_\epsilon(t) = \sum_{i=1}^n \int_0^t Q_i(u) 1\{|Q_i(u)| > \epsilon\} dM_i(u),$$

whose predictable variation process is

$$\langle U_\epsilon \rangle(t) = \sum_{i=1}^n \int_0^t Q_i^2(u) 1\{|Q_i(u)| > \epsilon\} 1(Y_i \geq u) e^{X_i^T \beta^0} \lambda_0(u) du$$
$$= \sum_{i=1}^n \int_0^t Q_i^2(u) 1\{|\sqrt{n} Q_i(u)| > \sqrt{n}\epsilon\} 1(Y_i \geq u) e^{X_i^T \beta^0} \lambda_0(u) du.$$

Let $Q_{\max} = \sup_{t \in [0,\tau]} \max_{1 \leq i \leq n} |\sqrt{n} Q_i(t)|$, then $1\{|\sqrt{n} Q_i(u)| > \sqrt{n}\epsilon\} \leq 1\{Q_{\max} >$

$\sqrt{n}\epsilon\}$. By (A1),

$$\sup_{t\in[0,\tau]}\max_{1\le i\le n}\left|\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_i-\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}\right|\le\lambda_{\min}^{-1/2}\|c\|_1\|\Theta_{\beta^0}\|_{1,1}2K=\mathcal{O}(\sqrt{p}),$$

and $Q_{\max}=\mathcal{O}(\sqrt{p})$. When $p/n\to 0$, $1\{Q_{\max}>\sqrt{n}\epsilon\}=0$ eventually. Thus $\langle U_\epsilon\rangle(t)\to_P 0$. Finally, by the martingale central limit theorem, the asymptotic normality is concluded. □

Lemma III.8. *Under assumptions (A1) – (A5), for the lasso estimator $\widehat{\beta}$, we have*

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

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

Lemma III.8 provides theoretical properties of the lasso estimator in the Cox model, the proof of which is omitted. This is a direct result from Theorem 1 in *Kong and Nan* (2014).

Lemma III.9. *Under assumptions (A1) – (A5), with probability going to 1, Θ_{β^0} is a feasible solution to the constraint in the quadratic programming problem, $\|\Theta_{\beta^0}\widehat{\Sigma}-I_p\|_\infty\le\gamma_n$, for $\gamma_n\asymp\|\Theta_{\beta^0}\|_{1,1}s_0\lambda+\|\Theta_{\beta^0}\|_{1,1}\sqrt{\log(p)/n}$. If $\lambda\asymp\sqrt{\log(p)/n}$, it suffices to take $\gamma_n\asymp\|\Theta_{\beta^0}\|_{1,1}s_0\lambda$.*

Lemma III.9 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 Θ_{β^0}, because the information matrix involves the coefficient estimates.

Proof of Lemma III.9. Write $A_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{X_i - \eta_0(t;\beta^0)\right\}^{\otimes 2} dN_i(t) - \Sigma_{\beta^0}$.

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty &\le \left\| \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\left\{X_i - \widehat{\eta}_n(t;\widehat{\beta})\right\}^{\otimes 2} - \left\{X_i - \eta_0(t;\beta^0)\right\}^{\otimes 2}\right] dN_i(t) \right\|_\infty \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{X_i - \eta_0(t;\beta^0)\right\}^{\otimes 2} dN_i(t) - \Sigma_{\beta^0} \right\|_\infty \\
&\le \left\| \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{X_i - \widehat{\eta}_n(t;\widehat{\beta})\right\} \left\{\widehat{\eta}_n(t;\widehat{\beta}) - \eta_0(t;\beta^0)\right\}^T dN_i(t) \right\|_\infty \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\widehat{\eta}_n(t;\widehat{\beta}) - \eta_0(t;\beta^0)\right\} \left\{X_i - \eta_0(t;\beta^0)\right\}^T dN_i(t) \right\|_\infty + \|A_n\|_\infty .
\end{aligned}$$

Note that for all $t \in [0,\tau]$, $\|X_i - \widehat{\eta}_n(t;\hat{\beta})\|_\infty \le 2K$ and $\|X_i - \eta_0(t;\beta^0)\|_\infty \le 2K$. Then

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty &\le \frac{4K}{n}\sum_{i=1}^{n}\int_0^\tau \|\widehat{\eta}_n(t;\hat{\beta}) - \eta_0(t;\beta^0)\|_\infty dN_i(t) + \|A_n\|_\infty \\
&\le \frac{4K}{n}\sum_{i=1}^{n}\int_0^\tau \|\widehat{\eta}_n(t;\hat{\beta}) - \widehat{\eta}_n(t;\beta^0)\|_\infty dN_i(t) \\
&\quad + \frac{4K}{n}\sum_{i=1}^{n}\int_0^\tau \|\widehat{\eta}_n(t;\beta^0) - \eta_0(t;\beta^0)\|_\infty dN_i(t) + \|A_n\|_\infty . \qquad (3.11)
\end{aligned}$$

By the mean value theorem, for the jth component in $\widehat{\eta}_n$ (denoted by $\widehat{\eta}_{nj}$), there exists some $\bar{\beta}^{(j)}$ lying between $\widehat{\beta}$ and β^0 such that

$$\widehat{\eta}_{nj}(t;\widehat{\beta}) = \widehat{\eta}_{nj}(t;\beta^0) + \left[\left.\frac{\partial \widehat{\eta}_{nj}(t;\beta)}{\partial \beta}\right|_{\beta=\bar{\beta}^{(j)}}\right]^T (\widehat{\beta} - \beta^0).$$

Consider β in a neighborhood of β^0, i.e. when $\|\beta - \beta^0\|_1 \le \delta'$ for some $\delta' > 0$, $e^{X_i^T\beta} \le e^{|X_i^T\beta|} \le e^{|X_i^T\beta^0| + K\delta'} \le e^{K_1 + K\delta'}$, and $e^{X_i^T\beta} \ge e^{-|X_i^T\beta|} \ge e^{-K_1 - K\delta'}$. Since $\{1(Y \ge t) : t \in [0,\tau]\}$ is $\mathbb{P}$-Glivenko-Cantelli, $\sup_{t\in[0,\tau]} |\frac{1}{n}\sum_{i=1}^{n} 1(Y \ge t) - \mathbb{P}(Y \ge$

$t)| \xrightarrow{a.s.} 0$, and then uniformly for $t \in [0,\tau]$ and $\beta \in \{\beta : \|\beta - \beta^0\|_1 \le \delta'\}$,

$$\widehat{\mu}_0(t;\beta) \ge \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \ge t) e^{-K_1 - K\delta'} \xrightarrow{a.s.} \mathbb{P}(Y \ge t) e^{-K_1-K\delta'} \ge \frac{\pi_0}{2} e^{-K_1-K\delta'}.$$

In this case, uniformly for $t \in [0,\tau]$ and $\beta \in \{\beta : \|\beta - \beta^0\|_1 \le \delta'\}$,

$$\begin{aligned}
\left\| \frac{\partial \widehat{\eta}_n(t;\beta)}{\partial \beta^T} \right\|_\infty &= \left\| \frac{\widehat{\mu}_2(t;\beta)\widehat{\mu}_0(t;\beta) - \widehat{\mu}_1(t;\beta)\widehat{\mu}_1(t;\beta)^T}{\widehat{\mu}_0^2(t;\beta)} \right\|_\infty \\
&\le_{a.s.} \left(\frac{\pi_0}{2} e^{-K_1-K\delta'}\right)^{-2} \left\{ e^{K_1+K\delta'} K^2 \cdot e^{K_1+K\delta'} + e^{2(K_1+K\delta')} K^2 \right\} \\
&= \frac{8}{\pi_0^2} e^{4(K_1+K\delta')} K^2 < \infty,
\end{aligned}$$

i.e. $\left\| \frac{\partial \widehat{\eta}_n(t;\beta)}{\partial \beta^T} \right\|_\infty$ is uniformly bounded almost surely. When $s_0\lambda \to 0$, we have $\|\widehat{\eta}_n(t;\widehat{\beta}) - \widehat{\eta}_n(t;\beta^0)\|_\infty \le \mathcal{O}_P(\|\widehat{\beta} - \beta^0\|_1) = \mathcal{O}_P(s_0\lambda)$ and the first term in (3.11) is $\frac{4K}{n}\sum_{i=1}^n \int_0^\tau \|\widehat{\eta}_n(t;\widehat{\beta}) - \widehat{\eta}_n(t;\beta^0)\|_\infty dN_i(t) = \mathcal{O}_P(s_0\lambda)$.

For the second term in (3.11), we use an argument from Lemma III.6 that $\sup_{t\in[0,\tau]} \|\widehat{\eta}_n(t;\beta^0) - \eta_0(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$ and then have

$$\begin{aligned}
&\frac{4K}{n}\sum_{i=1}^{n} \int_0^\tau \|\widehat{\eta}_n(t;\beta^0) - \eta_0(t;\beta^0)\|_\infty dN_i(t) \\
\le &\frac{4K}{n}\sum_{i=1}^{n} \int_0^\tau \sup_{t\in[0,\tau]} \|\widehat{\eta}_n(t;\beta^0) - \eta_0(t;\beta^0)\|_\infty dN_i(t) \\
= &\mathcal{O}_P(\sqrt{\log(p)/n}).
\end{aligned}$$

For the last term A_n, by Hoeffding's concentration inequality, we have for every $t > 0$ and $j,k = 1,\cdots,p$,

$$\mathbb{P}\left(|A_n(j,k)| \ge t\right) \le 2\exp\{-nt^2/C'\},$$

where C' is a constant only depending on K^4. Since A_n is a symmetric matrix,

$$\begin{aligned}\mathbb{P}\left(\|A_n\|_\infty \geq t\right) &= \mathbb{P}\left(\bigcup_{1\leq j\leq p, j\leq k\leq p} |A_n(j,k)| \geq t\right) \\ &\leq \sum_{j=1}^{p}\sum_{k=j}^{p} \mathbb{P}\left(|A_n(j,k)| \geq t\right) \\ &\leq p(p+1)\exp\{-nt^2/C'\}.\end{aligned}$$

So $\|A_n\|_\infty = \mathcal{O}_P\left(\sqrt{\log(p)/n}\right)$. Combining the three terms in (3.11), we have $\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty \leq \mathcal{O}_P(s_0\lambda + \sqrt{\log(p)/n})$. Finally, we conclude that

$$\begin{aligned}\|\Theta_{\beta^0}\widehat{\Sigma} - I_p\|_\infty &\leq \|\Theta_{\beta^0}\|_{1,1}\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty \\ &= \mathcal{O}_P\left(\|\Theta_{\beta^0}\|_{1,1}s_0\lambda + \|\Theta_{\beta^0}\|_{1,1}\sqrt{\log(p)/n}\right).\end{aligned}$$

□

Lemma III.10. *Assume* $\limsup_{n\to\infty} p\gamma_n \leq 1-\epsilon'$ *for some* $\epsilon' \in (0,1)$. *Then, under assumptions (A1) – (A5),* $\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty = \mathcal{O}_P(\gamma_n\|\Theta_{\beta^0}\|_{1,1})$.

Proof of Lemma III.10. Note that $\widehat{\Theta} - \Theta_{\beta^0} = \widehat{\Theta}(I_p - \widehat{\Sigma}\Theta_{\beta^0}) + (\widehat{\Theta}\widehat{\Sigma} - I_p)\Theta_{\beta^0}$, then on the event $\{\|\widehat{\Sigma}\Theta_{\beta^0} - I_p\|_\infty \leq \gamma_n\}$, we have

$$\begin{aligned}\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty &\leq \|\widehat{\Theta}\|_{\infty,\infty}\|I_p - \widehat{\Sigma}\Theta_{\beta^0}\|_\infty + \|\widehat{\Theta}\widehat{\Sigma} - I_p\|_\infty\|\Theta_{\beta^0}\|_{1,1} \\ &\leq \gamma_n\|\widehat{\Theta}\|_{\infty,\infty} + \gamma_n\|\Theta_{\beta^0}\|_{1,1}.\end{aligned}$$

Since $\|\widehat{\Theta}\|_{\infty,\infty} \leq \|\widehat{\Theta} - \Theta_{\beta^0}\|_{\infty,\infty} + \|\Theta_{\beta^0}\|_{\infty,\infty} \leq p\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty + \|\Theta_{\beta^0}\|_{1,1}$, we can obtain

$$\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty \leq \gamma_n\left(p\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty + \|\Theta_{\beta^0}\|_{1,1}\right) + \gamma_n\|\Theta_{\beta^0}\|_{1,1}.$$

When $\limsup_{n\to\infty} \gamma_n p \leq 1-\epsilon' < 1$, then for n large enough,

$$\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty \leq 2\gamma_n \|\Theta_{\beta^0}\|_{1,1}/(1-\gamma_n p) \asymp \gamma_n \|\Theta_{\beta^0}\|_{1,1}.$$

Therefore, by Lemma III.9, $\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty = \mathcal{O}_P(\gamma_n \|\Theta_{\beta^0}\|_{1,1})$. □

Lemma III.11. *By assumption (A1), for each $t > 0$,*

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

Proof of Lemma III.11. Noting that $\|X_i - \widehat{\eta}_n(t;\beta^0)\|_\infty \leq 2K$ uniformly for all i, Lemma III.11 is a direct result of Lemma 3.3(ii) in *Huang et al.* (2013). □

With Lemmas III.6, we complete the proof of Theorem III.1.

Proof of Theorem III.1. Recall that

$$\begin{aligned} c^T(\widehat{b} - \beta^0) = &-c^T\Theta_{\beta^0}\dot{\ell}_n(\beta^0) - c^T(\widehat{\Theta} - \Theta_{\beta^0})\dot{\ell}_n(\beta^0) \\ &- c^T(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0) + c^T\widehat{\Theta}(\widehat{\Sigma} - B_n)(\widehat{\beta} - \beta^0), \end{aligned}$$

where $B_n = \left(\ddot{\ell}_{n1}(\widetilde{\beta}^{(1)})^T, \cdots, \ddot{\ell}_{np}(\widetilde{\beta}^{(p)})^T\right)^T$.

First, we show that $\sqrt{n}c^T(\widehat{\Theta} - \Theta_{\beta^0})\dot{\ell}_n(\beta^0) = o_P(1)$. By Lemma III.10 and Lemma III.11,

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

Second, we show that $\sqrt{n}c^T(\widehat{\Theta}\widehat{\Sigma} - I_p)(\widehat{\beta} - \beta^0) = o_P(1)$. By Lemma III.8,

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

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

$$\widehat{\Sigma} - B_n = (\widehat{\Sigma} - \Sigma_{\beta^0}) + (\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0)) + (\ddot{\ell}_n(\beta^0) - B_n). \tag{3.12}$$

By the proof of Lemma III.9, we see that with $\lambda \asymp \sqrt{\log(p)/n}$, $\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty = \mathcal{O}_P(s_0\lambda)$. We rewrite

$$\begin{aligned}
\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0) &= \mathbb{E}\int_0^\tau \{X_i - \eta_0(t;\beta^0)\}^{\otimes 2} e^{X_i^T\beta^0} 1(Y_i \geq t)\lambda_0(t)dt \\
&\quad - \int_0^\tau \left\{\widehat{\mu}_2(t;\beta^0) - \frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right\}\lambda_0(t)dt \\
&\quad - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2}\right\} dM_i(t) \\
&= \int_0^\tau \{\mu_2(t;\beta^0) - \widehat{\mu}_2(t;\beta^0)\}\lambda_0(t)dt \\
&\quad + \int_0^\tau \left\{\frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \frac{\mu_1(t;\beta^0)\mu_1^T(t;\beta^0)}{\mu_0(t;\beta^0)}\right\}\lambda_0(t)dt \\
&\quad - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2}\right\} dM_i(t).
\end{aligned} \tag{3.13}$$

Similar to the proof in Lemma III.6, 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 $\| \int_0^\tau \{\mu_2(t;\beta^0) - \widehat{\mu}_2(t;\beta^0)\}\lambda_0(t)dt\|_\infty \leq \sup_{t\in[0,\tau]} \|\widehat{\mu}_2(t;\beta^0) - \mu_2(t;\beta^0)\|_\infty \int_0^\tau \lambda_0(t)dt = \mathcal{O}_P(\sqrt{\log(p)/n})$. 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 (3.13), by (A1) and Lemma III.6,

$$\left\| \int_0^\tau \left\{ \frac{\widehat{\mu}_1(t;\beta^0)\widehat{\mu}_1^T(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \frac{\mu_1(t;\beta^0)\mu_1^T(t;\beta^0)}{\mu_0(t;\beta^0)} \right\} \lambda_0(t)dt \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)}\right]^{\otimes 2} \right\} dM_i(t)$ is a sum of n independent and identically distributed mean zero terms, and each term $\left\| \int_0^\tau \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)}\right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty$ is bounded by $2K^2(1+e^{K_1}\Lambda_0(\tau))$ 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 III.9, 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

$$\begin{aligned} &\left\| \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2} \right\} dM_i(t) \right. \\ &\left. -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\mu_2(t;\beta^0)}{\mu_0(t;\beta^0)} - \left[\frac{\mu_1(t;\beta^0)}{\mu_0(t;\beta^0)}\right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty = \mathcal{O}_P\left(\sqrt{\frac{\log(p)}{n}}\right), \end{aligned}$$

and thus for the third term in (3.13), $\left\| \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ \frac{\widehat{\mu}_2(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)} - \left[\frac{\widehat{\mu}_1(t;\beta^0)}{\widehat{\mu}_0(t;\beta^0)}\right]^{\otimes 2} \right\} dM_i(t) \right\|_\infty =$

$\mathcal{O}_P(\sqrt{\log(p)/n})$. Therefore, by (3.13), $\|\Sigma_{\beta^0} - \ddot{\ell}_n(\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n})$.

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

$$\ddot{\ell}_{njk}(\widetilde{\beta}^{(j)}) - \ddot{\ell}_{njk}(\beta^0) = (\widetilde{\beta}^{(j)} - \beta^0)^T \left.\frac{\partial \ddot{\ell}_{njk}(\beta)}{\partial \beta}\right|_{\beta=\overline{\beta}^{(jk)}},$$

where $\overline{\beta}^{(jk)}$ lies in the segment between $\widetilde{\beta}^{(j)}$ and β^0. Under assumptions (A1) – (A3), when $\|\beta - \beta^0\|_1 \leq \delta'$ for $\delta' > 0$ small enough, $\left\|\frac{\partial \ddot{\ell}_{njk}(\beta)}{\partial \beta}\right\|_\infty$ is bounded by some constant related to δ' uniformly for all (j,k). Since $s_0\lambda = o(1)$, we have $\|B_n - \ddot{\ell}_n(\beta^0)\|_\infty \leq \mathcal{O}_P(\|\widehat{\beta} - \beta^0\|_1) = \mathcal{O}_P(s_0\lambda)$.

Combining the three parts in (3.12), we have that for $\lambda \asymp \sqrt{\log(p)/n}$, $\|\widehat{\Sigma} - B_n\|_\infty = \mathcal{O}_P(s_0\lambda)$. Then

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

Then, we show that the variance estimator is consistent, i.e. $c^T(\widehat{\Theta} - \Theta_{\beta^0})c \to_P 0$ 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 $\sqrt{n}c^T(\widehat{b} - \beta^0)/(c^T\widehat{\Theta}c)^{1/2} \xrightarrow{\mathcal{D}} N(0,1)$. □

CHAPTER IV

Confidence Intervals for Stratified Cox Model with Many Covariates: With Applications to Kidney Transplant Data

4.1 Introduction

In 2016, nearly 125,000 people in the United States started treatment for end stage renal disease (ESRD), and more than 726,000 were on dialysis or were living with a kidney transplant, according to a 2019 Centers for Disease Control and Prevention report[1]. Successful renal transplantation improves the quality of life and increases survival for patients with ESRD, as compared with long-term dialysis (*Wolfe et al.*, 1999). There are still ongoing challenges to optimize access to kidney transplant and graft survival. It is thus crucial to study the potential risk factors of renal transplant failure to provide evidence-based explanations and improve prediction and prognosis. Recipient's and donor's age affect graft survival, although donor's age may have much stronger effects than those of recipient's age (*Kasiske and Snyder*, 2002). Other important factors may include, but limited to, immunosuppressive therapy, cardiac or respiratory disease, obesity, chronic infection such as HIV or hepatitis and

[1]See the web content `https://www.cdc.gov/kidneydisease/pdf/2019_National-Chronic-Kidney-Disease-Fact-Sheet.pdf`

diabetes (*Rodger*, 2012; *Legendre et al.*, 2014).

United Network for Organ Sharing (UNOS) is the non-profit organization that manages the organ transplant system under contract with the federal government in the United States (`https://unos.org/`). The Scientific Registry of Transplant Recipients (SRTR) system has been keeping records of kidney transplant information from waitlisted candidates, recipients and donors in the United States, submitted by members of the Organ Procurement and Transplantation Network (OPTN). Post-transplant outcomes such as graft survival and patient survival are closely monitored at UNOS. Therefore, the SRTR database is a rich resource for studying kidney transplantation. While the SRTR website[2] has only reported the regression coefficients in a Cox proportional hazards model for a selective subset of factors that are considered predictive of the post-transplant outcomes, our primary goal is to study the joint associations of as many potential risk factors as possible on patient graft survival after kidney transplantation.

Many of these risks factors are very complex and the number of covariates can easily increase beyond the level where the conventional survival models can provide reliable statistical inference on their associations. In addition, the kidney transplant center where a transplant occurs also plays an important role, for example, through the quality of care delivered to patients. Although there are statistical methods developed to incorporate the center effects, e.g. *He et al.* (2019) with a lognormal frailty, they would usually bring additional computational complexity and are not our main focus.

We consider the stratified Cox proportional hazards model (*Kalbfleisch and Prentice*, 2002) in this chapter with potentially a diverging number of covariates, to which we extend the inferential method of de-biasing lasso proposed in Chapter III. The stratified Cox model has many applications in biomedical studies and allows for dif-

[2] `https://www.srtr.org/reports-tools/posttransplant-outcomes/`

ferent baseline hazards in different strata, which is particularly useful when some covariate effects do not satisfy the proportional hazards assumption or data are stratified based on some factors not of primary interest. A stratification factor can be gender, age groups or geographic areas, and in the example of kidney transplants, stratification can be naturally based on transplant centers. When the number of covariates is large compared to, though not necessarily larger than, the sample size, the conventional methods, e.g. the maximum partial likelihood estimation, may give rise to very biased estimates and unreliable inference results. *Morris et al.* (2018) implemented the gradient boosting algorithm for variable selection in the stratified Cox model with high-dimensional covariates. However, to the best of our knowledge, there lacks work on the empirical and theoretical studies of statistical inference in the stratified Cox model with a diverging number of covariates.

The rest of this chapter is organized as follows. Section 4.2 introduces the setup of the stratified Cox proportional hazards model and the de-biasing lasso approach for inference on the regression parameters. Section 4.3 provides the theoretical results for the de-biased lasso estimator under certain regularity conditions. Simulation studies and an application to the national kidney transplant data are presented in Section 4.4 and Section 4.5, respectively. Finally, technical proofs of the main theorem and useful lemmas are provided in Section 4.6.

4.2 Method

4.2.1 Stratified Cox proportional hazards model

We first introduce some notation. Let T denote the underlying failure time, and C the censoring time, which is assumed independent of T given the p-dimensional covariates of interest, $X \in \mathbb{R}^p$. $\delta = 1(T \leq C)$ is the event indicator and $Y = \min(T, C)$ denotes the observed survival time. Suppose that in the kth stratum,

we have n_k observations, $k = 1, \cdots, K$, and $N = \sum_{k=1}^{K} n_k$ is the total number of observations in the data. Observations within the kth stratum are indexed by i, $i = 1, \cdots, n_k$.

The stratified Cox model assumes that the true hazard function for the underlying failure time T_{ki}, conditional on X_{ki}, is

$$\lambda_{ki}(t|X_{ki}) = \lambda_{0k}(t) \exp\{X_{ki}^T \beta^0\},$$

where $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)^T \in \mathbb{R}^p$ is an unknown vector of common regression coefficients across strata, and $\lambda_{0k}(t)$ is the unknown baseline hazard function in stratum k. The problem of interest is the estimation and reliable inference on the regression coefficients β^0.

We assume independence between strata and among observations in each stratum. The negative log partial likelihood is written as

$$\ell(\beta) = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left[\beta^T X_{ki} - \log \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} 1(Y_{kj} \geq Y_{ki}) \exp(\beta^T X_{kj}) \right\} \right] \delta_{ki}. \quad (4.1)$$

The negative log partial likelihood (4.1) can be rewritten as

$$\begin{aligned} \ell(\beta) &= -\sum_{k=1}^{K} \frac{n_k}{N} \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\beta^T X_{ki} - \log \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} 1(Y_{kj} \geq Y_{ki}) \exp(\beta^T X_{kj}) \right\} \right] \delta_{ki} \\ &= \sum_{k=1}^{K} \frac{n_k}{N} \ell_k(\beta), \end{aligned}$$

where $\ell_k(\beta)$ denotes the usual negative log likelihood for the kth stratum. Let $\dot{\ell}(\beta)$

and $\ddot{\ell}(\beta)$ be the first and the second order derivatives with respect to β. Here,

$$\dot{\ell}(\beta) = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\{X_{ki} - \frac{\widehat{\mu}_{1k}(Y_{ki};\beta)}{\widehat{\mu}_{0k}(Y_{ki};\beta)}\right\}\delta_{ki},$$
$$\ddot{\ell}(\beta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\{\frac{\widehat{\mu}_{2k}(Y_{ki};\beta)}{\widehat{\mu}_{0k}(Y_{ki};\beta)} - \left[\frac{\widehat{\mu}_{1k}(Y_{ki};\beta)}{\widehat{\mu}_{0k}(Y_{ki};\beta)}\right]^{\otimes 2}\right\}\delta_{ki},$$

where $\widehat{\mu}_{rk}(t;\beta) = {n_k}^{-1}\sum_{j=1}^{n_k} 1(Y_{kj} \geq t)X_{kj}^{\otimes r}\exp\{X_{kj}^T\beta\},\ r = 0, 1, 2.$

4.2.2 De-biasing the lasso estimator

Similar to the Cox model, the lasso estimator in the stratified Cox model minimizes the penalized negative log partial likelihood, i.e.

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\{\ell(\beta) + \lambda\|\beta\|_1\} \tag{4.2}$$

for some tuning parameter $\lambda > 0$.

The de-biased lasso estimator is defined as

$$\widehat{b} = \widehat{\beta} - \widehat{\Theta}\dot{\ell}(\widehat{\beta}), \tag{4.3}$$

where $-\widehat{\Theta}\dot{\ell}(\widehat{\beta})$ acts as a bias correction term from $\widehat{\beta}$. In (4.3), $\widehat{\Theta}$ is obtained by solving quadratic programming problems, which is parallel to the single stratum case in Chapter III. Let

$$\widehat{\Sigma} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\delta_{ki}\left[X_{ki} - \widehat{\eta}_k(Y_{ki};\widehat{\beta})\right]^{\otimes 2},$$

where $\widehat{\eta}_k(t;\beta) = \widehat{\mu}_{1k}(t;\beta)/\widehat{\mu}_{0k}(t;\beta)$ is the vector of weighted average covariates. For each $j = 1, \cdots, p$, let $m^{(j)}$ be the solution to the following quadratic programming

problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ m^T \widehat{\Sigma} m : \|\widehat{\Sigma} m - e_j\|_\infty \leq \gamma \right\}, \tag{4.4}$$

where $\gamma > 0$ is a tuning parameter that controls the bias correction. Then we define the $p \times p$ matrix $\widehat{\Theta} = (m^{(1)}, \cdots, m^{(p)})^T$, i.e. with ${m^{(j)}}^T$ as the jth row.

As has been shown in Chapter III, it is very important to select a proper tuning parameter γ in (4.4) to achieve the desirable amount of bias correction and generate honest confidence intervals. Through extensive simulations, we have found that Algorithm 3.1 in Chapter III is still feasible for selecting the tuning parameter γ in the case of stratification, except that we need to modify the random data splitting in the cross-validation. Instead of randomly splitting the whole dataset into multiple equally sized folds, one should split the data within each stratum into different folds simultaneously as before, so that each of the final cross-validation folds has data aggregated from all the strata. We will omit the outline of detailed algorithm due to overwhelming overlaps with Algorithm 3.1 in Chapter III.

4.3 Theoretical results

We provide theoretical justification for the proposed de-biased lasso estimator and the corresponding inference procedure in Section 4.2. Throughout this section, the number of strata K is considered fixed and we proceed by allowing the minimum number of observations $n_{min} = \min_{1 \leq k \leq K} n_k$ to increase. In addition, it is assumed that the proportion of observations in each stratum $\dfrac{n_k}{N} \to r_k$ as $n_{min} \to \infty$, for some $r_k \in (0, 1),\ k = 1, \cdots, K$.

For convenience, we define $\mu_{hk}(t; \beta) = \mathbb{E}[\widehat{\mu}_{hk}(t; \beta)]$, the expectation of $\widehat{\mu}_{hk}(t; \beta)$ in Section 4.2.1, $h = 0, 1, 2,\ k = 1, \cdots, K$. The population-level weighted covariate

process for $\widehat{\eta}_k(t;\beta) = \widehat{\mu}_{1k}(t;\beta)/\widehat{\mu}_{0k}(t;\beta)$ is $\eta_{k0}(t;\beta) = \mu_{1k}(t;\beta)/\mu_{0k}(t;\beta)$. Let

$$\Sigma_{\beta^0,k} = \mathbb{E}[\{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2}\delta_{ki}]$$

be the population-level information matrix for the kth stratum, $k = 1, \cdots, K$. The total information matrix is then defined as a weighted average of the stratum-specific information matrices,

$$\Sigma_{\beta^0} = \sum_{k=1}^{K} r_k \Sigma_{\beta^0,k}.$$

The inverse information matrix is $\Theta_{\beta^0} = \Sigma_{\beta^0}^{-1}$, which the matrix $\widehat{\Theta}$ obtained from solving (4.4) can be viewed as an approximation for.

To prove the main theoretical result of this chapter, we assume the following regularity conditions similar to those in Chapter III.

(A1) Covariates are almost surely uniformly bounded, i.e. $\|X_{ki}\|_\infty \leq M$ for some constant $M > 0$.

(A2) $|X_{ki}^T\beta^0| \leq M_1$ uniformly for all k and i with some constant $M_1 > 0$, almost surely.

(A3) The follow-up time stops at a finite time point $\tau > 0$, with probability $\pi_0 = \min_{k,i} P(Y_{ki} \geq \tau) > 0$.

(A4) For any $t \in [0, \tau]$,

$$\frac{c^T\Theta_{\beta^0}}{c^T\Theta_{\beta^0}c}\left[\int_0^t \left\{\mu_{2k}(u;\beta^0) - \frac{\mu_{1k}(u;\beta^0)\mu_{1k}(u;\beta^0)^T}{\mu_{0k}(u;\beta^0)}\right\} d\Lambda_{0k}(u)\right]\Theta_{\beta^0}c \to v_k(t)$$

as $n \to \infty$, for some function $v_k(\cdot) > 0$, $k = 1, \cdots, K$.

(A5) $0 < \epsilon_0 \leq \min_k \lambda_{\min}(\Sigma_{\beta^0,k}) \leq \max_k \lambda_{\max}(\Sigma_{\beta^0,k}) \leq 1/\epsilon_0 < \infty$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and the largest eigenvalues of a matrix, respectively.

Similar to Theorem III.4, if we are interested in the inference on $J\beta^0$ for some $J \in \mathbb{R}^{m\times p}$, the following assumption (A4)$'$ in required in replacement of (A4).

(A4)$'$ For any $\omega \in \mathbb{R}^m$,

$$\frac{\omega^T J\Theta_{\beta^0}}{\omega^T J\Theta_{\beta^0} J^T\omega}\left[\int_0^t \left\{\mu_{2k}(u;\beta^0) - \frac{\mu_{1k}(u;\beta^0)\mu_{1k}(u;\beta^0)^T}{\mu_{0k}(u;\beta^0)}\right\} d\Lambda_{0k}(u)\right]\Theta_{\beta^0} J^T\omega \to v_k'(t)$$

as $n \to \infty$, for any $t \in [0,\tau]$ and some function $v_k'(\cdot) > 0$, $k = 1, \cdots, K$.

Theorem IV.1. *Assume that* $\|\Theta_{\beta^0}\|_{1,1}^2\{\max_k |n_k/N - r_k| + s_0\lambda\}p\sqrt{\log(p)} \to 0$ *as* $n_{min} \to \infty$, *with the tuning parameter* $\lambda \asymp \sqrt{\log(p)/n_{min}}$ *for* $\widehat{\beta}$ *in (4.2). Under assumptions (A1) – (A5), for any* $c \in \mathbb{R}^p$ *such that* $\|c\|_2 = 1$ *and* $\|c\|_1 \le a_*$, *where* $a_* > 0$ *is some absolute constant, we have*

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

Furthermore, suppose that the $m \times p$ *matrix* J *has a fixed number of rows* m *and full row rank,* $\|J\|_{\infty,\infty} = \mathcal{O}(1)$ *and* $J\Theta_{\beta^0}J^T \to F$ *for some constant* $m \times m$ *positive definite matrix* F. *Then, with the additional assumption (A4)$'$,*

$$\sqrt{N}J(\widehat{b} - \beta^0) \xrightarrow{\mathcal{D}} \mathcal{N}(0,F).$$

Theorem IV.1 shows the asymptotic normality for linear combinations of the debiased lasso estimator (4.3). Based on Theorem IV.1, one may construct the asymptotic level α confidence interval for $c^T\beta^0$ as

$$\left[c^T\widehat{b} - z_{\alpha/2}(c^T\widehat{\Theta}c/N)^{1/2},\ c^T\widehat{b} + z_{\alpha/2}(c^T\widehat{\Theta}c/N)^{1/2}\right].$$

The proof of Theorem IV.1, almost parallel to that of Theorem III.1 and Theorem III.4, is provided in Section 4.6 along with some useful lemmas.

4.4 Simulation studies

We simulate $K = 10$ strata and, for simplicity, an equal number of subjects $n_k \equiv n = 60$ per stratum, with $p = 100$ covariates. The covariates X are first simulated independently from multivariate Gaussian distribution, with mean zero and two covariance structures, independence and AR(1) with correlation $\rho = 0.5$, in two different settings. The simulated covariates are then truncated at ± 2.5 if they exceed the bound. The first element β_1^0 in the true regression coefficient vector β^0 varies from 0 to 2 by increment 0.2, four of the rest coefficients are arbitrarily chosen to take values of 1, 1, 0.5 and 0.5, and all others are zero. Constant baseline hazards for different strata are simulated as $\lambda_{0k}(t) = \lambda_{0k} \sim \text{Uniform}(0.1, 0.5)$. The underlying survival time T in the kth stratum, given covariates X, follows an exponential distribution with constant hazard $\lambda_{0k} \exp\{X^T \beta^0\}$. The censoring time C, independent of T, is simulated from Uniform$(1, 30)$ and truncated at the maximum observation time $\tau = 20$. And the observed survival time $Y = \min(T, C)$. We monitor the estimation bias for β_1^0 over a range of [0, 2], as well as its model-based standard error, coverage probability of 95% confidence interval and empirical standard error. Besides the proposed method (QP), we also include the oracle estimator as if the true model were known (ORACLE) and the usual maximum partial likelihood estimator (MPLE). We use 5-fold cross-validation to select tuning parameters, both for the lasso estimator in `cv.glmnet()` and for γ in QP. For each value of β_1^0, the simulation is repeated 200 times.

Figure 4.1 and Figure 4.2 show the simulation results under the independence and the AR(1) covariance structures for the covariates. The conventional MPLE yields estimates for β_1^0 with the largest biases and variation. When $\beta_1^0 = 0$, the coverage probability of the 95% confidence interval from MPLE is only 91% in Figure 4.2, and the corresponding type 1 error of 9% is beyond the target level of 5%. The coverage probability from MPLE continues to drop as β_1^0 increases, to 28% when $\beta_1^0 = 2$ in

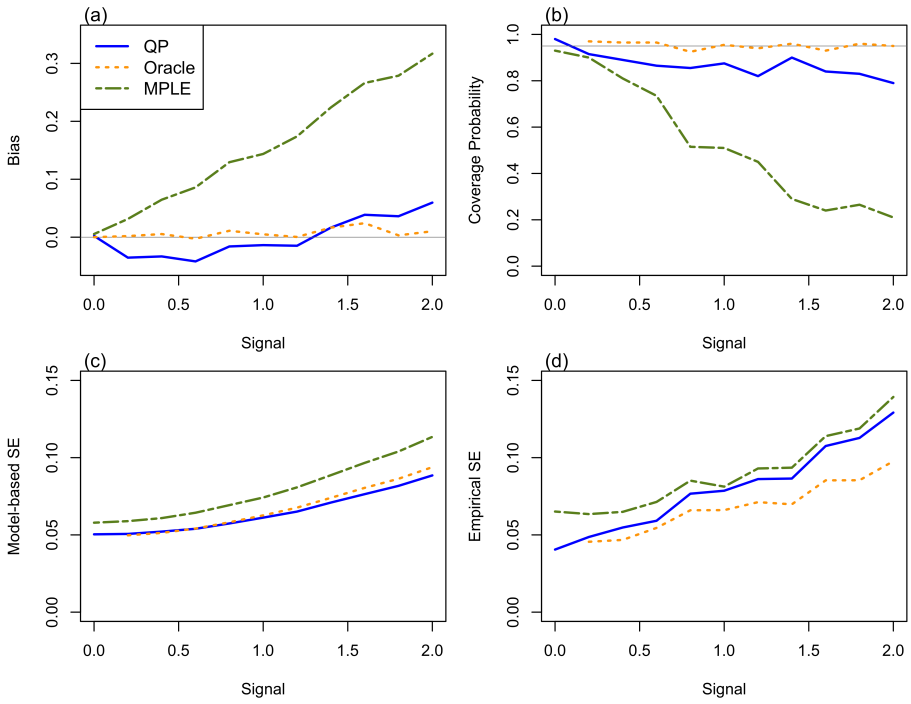

Figure 4.1: Simulation results when $K = 10, n = 60, p = 100$ and covariates follow $N(0, I_p)$. (a) Estimation bias for β_1^0, (b) Empirical coverage probability of 95% confidence interval for β_1^0, (c) Model-based standard error and (d) Empirical standard error. The x-axis represents the true value of the first regression coefficient β_1^0.

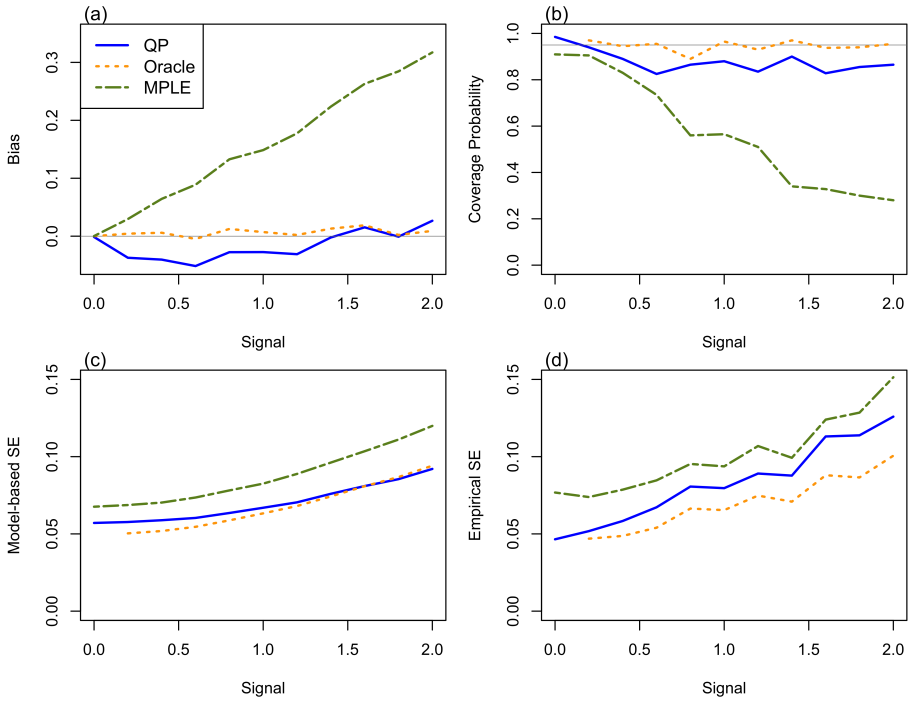

Figure 4.2: Simulation results when $K = 10, n = 60, p = 100$ and covariates have AR(1) covariance structure ($\rho = 0.5$). (a) Estimation bias for β_1^0, (b) Empirical coverage probability of 95% confidence interval for β_1^0, (c) Model-based standard error and (d) Empirical standard error. The x-axis represents the true value of the first regression coefficient β_1^0.

Figure 4.2. On the contrary, our proposed method QP preserves the type 1 error rates in both cases and the biases are very close to those from the oracle estimator. Its 95% confidence interval coverage probability lingers beyond the 85% level for most of the signal values. Comparing the standard errors, we find that the model-based equation has underestimated the true variability of the proposed de-biased lasso estimator for larger signals, which has an adverse effect on the coverage probability.

4.5 Application to the national kidney transplant data

In this section, we apply the proposed de-biasing lasso approach to the U.S. kidney transplant data, collected from the SRTR system on all donors, waitlisted candidates and transplant recipients in the United States.[3] In this analysis, we have included the patients who were greater than 25 years old at the time of receiving transplant from cadaveric donors during the year of 2000 to 2001. Donor age is a very important factor for kidney graft survival, whose effects are usually stronger than recipient age, and there has been many discussions over the matching between recipient and donor age (*Kasiske and Snyder*, 2002; *Veroux et al.*, 2012). We first separated the data into three groups for further analysis, based on whether recipient age is in the range of $[25, 45], (45, 60]$ or greater than 60 years old. Within each group, transplant centers with less than 20 transplants are excluded. The final total sample size is 4432, 5564 and 1643, in these age groups respectively. Table 4.1 summarizes the sample size information, age and gender characteristics in the study population by recipient age group. The 45-60 and > 60 years old groups have the largest and the smallest sample size, respectively. The number of patients within each center varies significantly, and the distribution is very skewed with mostly small centers in all three groups (Figure 4.3).

[3]The interpretation of the presented results does not reflect those of the SRTR or the U.S. government.

Table 4.1: Study population characteristics by recipient age group

Recipient age group	$[25, 45]$	$(45, 60]$	> 60
Variable	Mean (SD) / Count (%)		
# Centers	106 (–)	126 (–)	47 (–)
# Patients	4432 (100%)	5564 (100%)	1643 (100%)
# Events	2101 (47.4%)	3043 (54.7%)	1156 (70.4%)
Recipient age	36.9 (5.8)	52.9 (4.2)	66.4 (4.3)
Donor age (years)			
≤ 10	301 (6.8%)	263 (4.7%)	67 (4.1%)
$(10, 20]$	749 (16.9%)	818 (14.7%)	166 (10.1%)
$(20, 30]$	813 (18.3%)	906 (16.3%)	207 (12.6%)
$(30, 40]$	673 (15.2%)	784 (14.1%)	201 (12.2%)
$(40, 50]$	1003 (22.6%)	1204 (21.6%)	298 (18.1%)
$(50, 60]$	697 (15.7%)	1096 (19.7%)	366 (22.3%)
> 60	196 (4.4%)	493 (8.9%)	338 (20.1%)
Recipient gender			
Male	2609 (58.9%)	3420 (61.5%)	1056 (64.3%)
Female	1823 (41.1%)	2144 (38.5%)	587 (35.7%)
Donor gender			
Male	2679 (60.4%)	3262 (58.6%)	926 (56.4%)
Female	1753 (39.6%)	2302 (41.4%)	717 (43.6%)

The failure time of interest is defined as the time from a patient's receiving kidney transplantation to graft failure, which is when a transplanted kidney ceases to function properly, or death, whichever occurred first. Figure 4.4 plots the group-specific Kaplan-Meier curves and the complementary log-log curves for overall survival probabilities. Pooling all data without considering center effects, the test statistic for proportional hazards assumption (*Grambsch and Therneau*, 1994) is $\chi^2 = 22.86$ (df=2, p-value=1.09×10^{-5}). Thus, to proceed, we fit a separate model to each of the three recipient age groups. One of our analytic goals is to estimate the effects of donor age (categorized by an increment of 10 years), as well as recipient and donor gender, in each recipient age group and observe if there is any difference across recipient groups. The other primary goal is to compare the major significant factors and their effects across recipient groups. A total number of 132 variables are involved in the analysis, including but limited to recipient age (linear), indicators for donor age, candidate

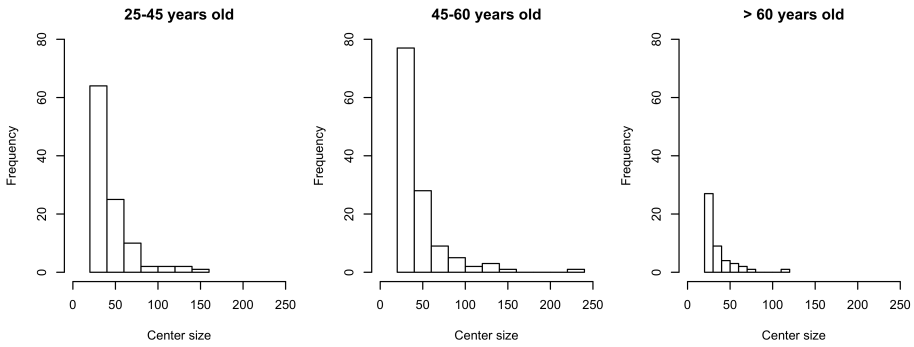

Figure 4.3: Histograms of transplant center size in the three recipient age groups

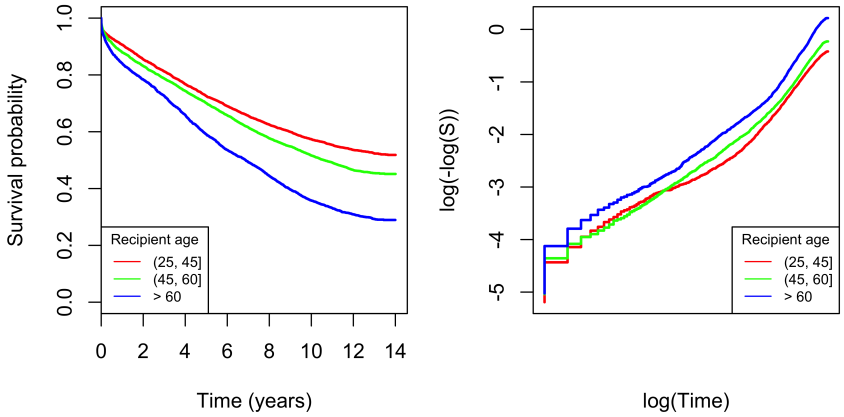

Figure 4.4: Kaplan-Meier curves (left) and complementary log-log (right) of the survival probabilities in the three recipient age groups

and donor's gender, race, history of diabetes and hypertension, blood type and HLA matching, recipient's primary kidney diagnosis. Since the transplant center effects are not of primary interest yet may affect graft survival through, e.g. quality of patient care, each group-specific Cox model is stratified by recipient transplant center. For the de-biasing lasso approach via quadratic programming (QP), we still use the 5-fold cross-validation for selecting tuning parameters for both λ and γ, as in Section 4.4. For comparison, we also include the results from the maximum partial likelihood estimation (MPLE).

Figure 4.5 shows that the point estimates from the de-biased lasso estimator are generally smaller in magnitudes compared to MPLE in the older than 60 age group, which has only 1643 patients, while are close to those from the MPLE for the vast

majority of variables in the largest $(45, 60]$ age group. This indicates that the number of variables is relatively large compared to the sample size in the older than 60 age group. The model-based standard error estimates from de-biasing the lasso are consistently smaller than those of the MPLE, leading to narrower confidence intervals.

Due to the space limit, for each recipient age group, we report the 95% confidence intervals for regression coefficients of a subset of variables whose p-values from de-biasing the lasso are less than 0.05, that is, the corresponding confidence intervals for log hazard ratio exclude zero; see Figure 4.6 for the $(25, 45]$ age group, Figure 4.7 for the $(45, 60]$ age group and Figure 4.8 for the older than 60 age group. In all three age groups, whether any medications have been given to recipients for maintenance or anti-rejection ("1: REC_IMMUNO_MAINT_MEDS_Y", Yes versus No) has the largest effects. It is consistent with long-standing studies on the need of immunosuppression for prevention of graft rejection (*Opelz*, 1994). But the de-biased lasso estimates are more consistent across groups, while the MPLE is prone to inflated estimates with smaller sample sizes. The de-biased lasso is more powerful to detect the significant associations that the MPLE does not, such as candidate being Hispanic ("19: CAN_Hisp" in Figure 4.6) and recipient having primary diagnosis of type 1 diabetes with insulin dependency and juvenile onset ("21: REC_DGN_3011" in Figure 4.6) in the $(25, 45]$ age group, recipient having primary diagnosis of hypertensive nephrosclerosis compared to others ("20: REC_DGN_3040" in Figure 4.7), of type 1 diabetes with insulin dependency and juvenile onset compared to others ("26: REC_DGN_3011" in Figure 4.6) in the $(45, 60]$ age group, and donor being white compared to black ("13: DON_RACE_SRTR_white" in Figure 4.8) and candidate having symptomatic peripheral vascular disease ("14: CAN_PERIPH_VASC_Y" in Figure 4.8) in the older than 60 age group.

While recipient age has no significant effect on death censored graft survival in the younger two groups, an older age is associated with increased hazard in the older than

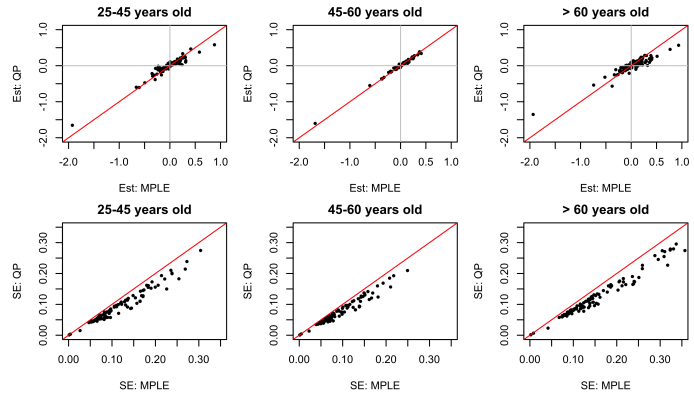

Figure 4.5: Comparison between the de-biased lasso estimator via quadratic programming (y-axis) and the MPLE (x-axis) in their point estimates and model-based standard error estimates. The red lines are the 45 degree lines.

60 years old group (for one year older, the de-biased lasso estimate of hazard ratio = 1.03, 95% CI 1.01 - 1.04, p-value = 2.5×10^{-4}). Figure 4.9 shows the estimated hazard ratios of different donor age categories compared to the reference of donor age between 20-30 years old. Again, with smaller sample sizes, the de-biased lasso estimates are less inflated and more stable than the MPLE. Among recipients between 25 and 60 years old, we observe increased hazards of receiving kidneys from donors aged 50-60 and over 60, while this effect starts to manifest earlier from donors aged 40-50 among recipients over 60 years old. For kidney transplantation, studies have conferred conflicting results on donor and recipient age combinations (*Dayoub et al.*, 2018). Our results from this database can contribute more insights to the existing literature. Table 4.2 reports the estimated effects of recipient and donor gender in each recipient age group. Jointly testing whether the recipient and donor gender have effects on graft survival does not show any significance, with test statistics 3.08 (p-value = 0.21), 3.06 (p-value = 0.22) and 0.39 (p-value = 0.82) in the three groups respectively.

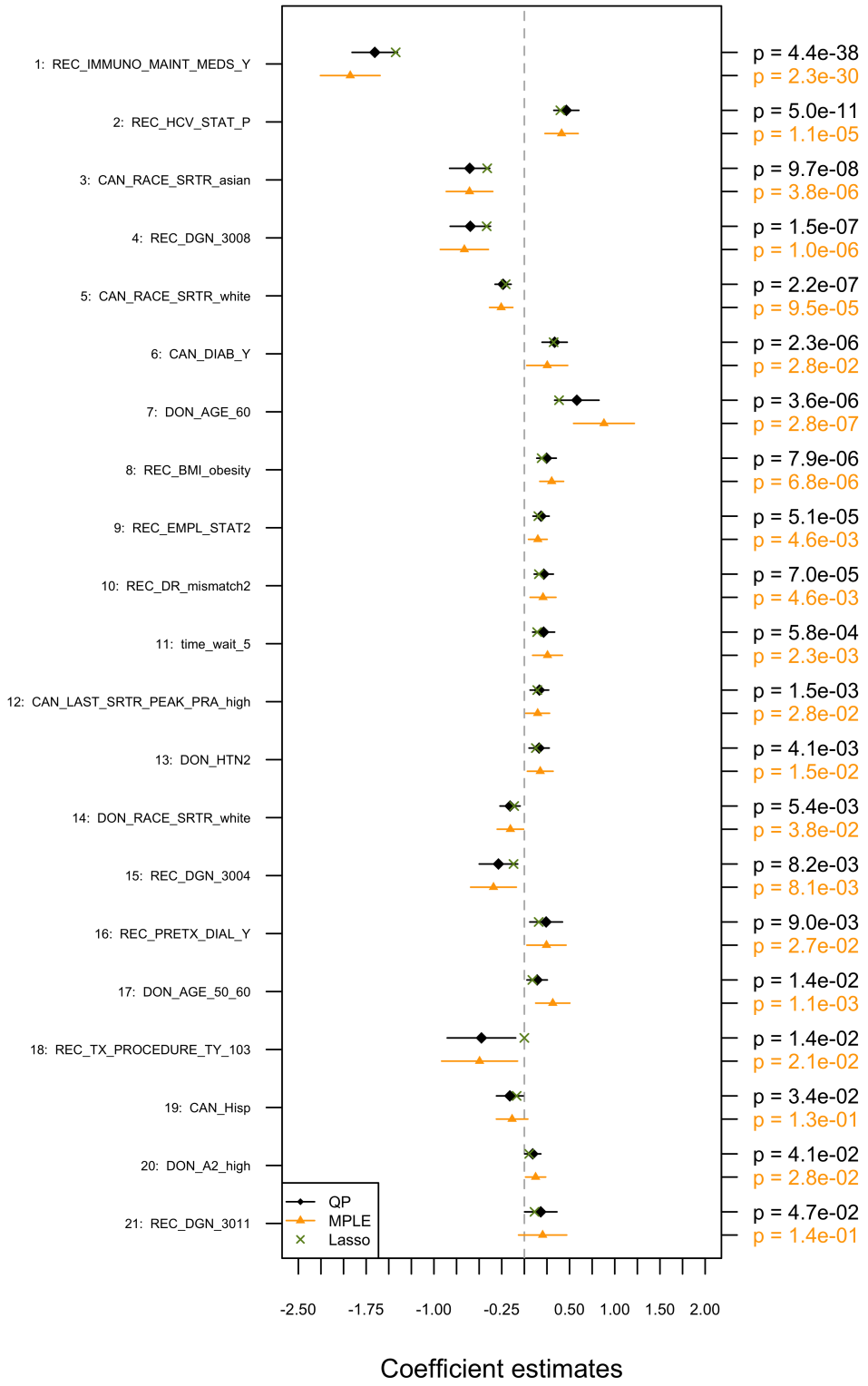

Figure 4.6: Point estimates and 95% confidence intervals for the de-biased lasso estimator via quadratic programming and the MPLE in the $(25, 45]$ age group. Only the variables whose p-values from the de-biasing lasso are less than 0.05 are reported, in ascending order of their p-values.

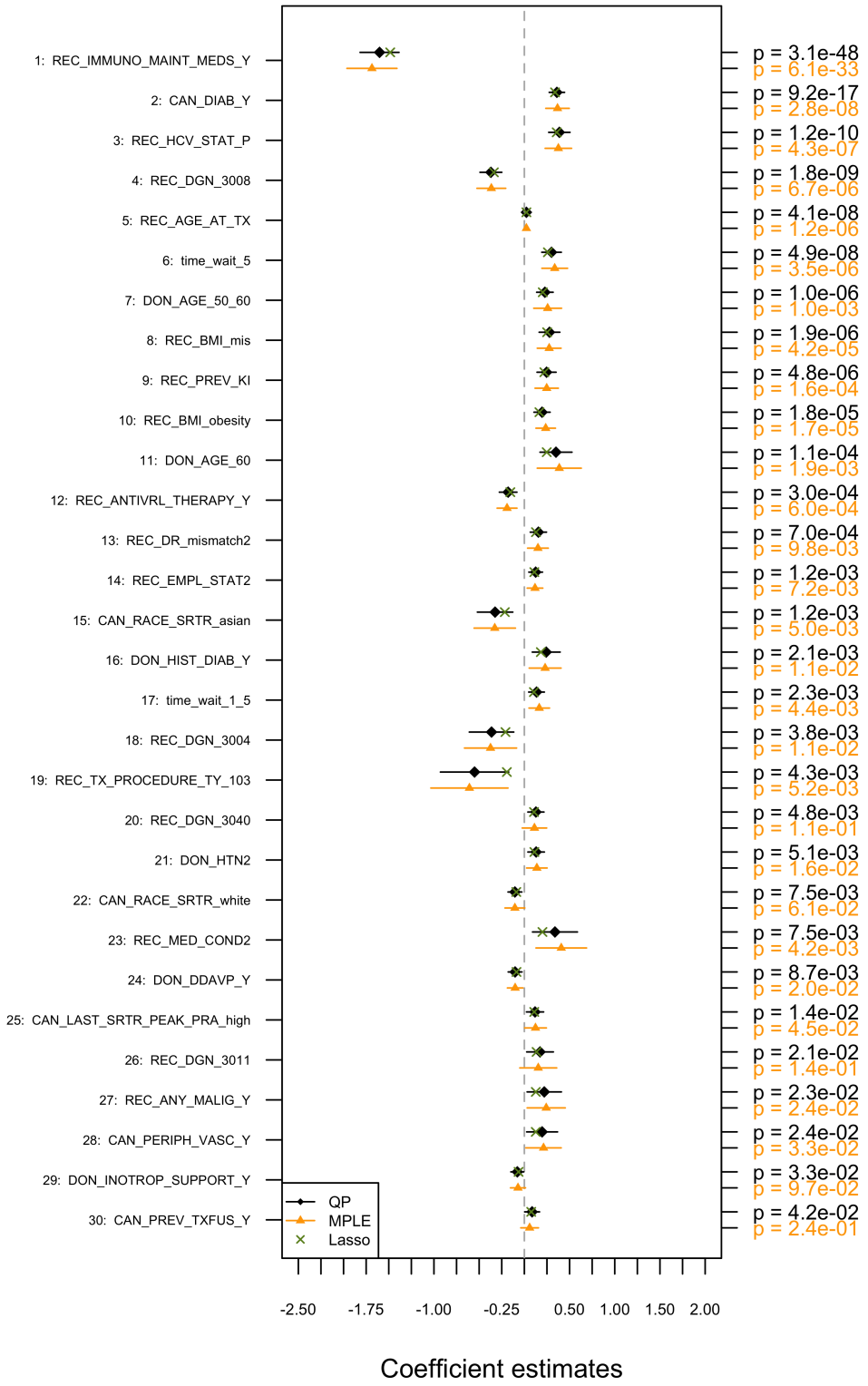

Figure 4.7: Point estimates and 95% confidence intervals for the de-biased lasso estimator via quadratic programming and the MPLE in the $(45, 60]$ age group. Only the variables whose p-values from the de-biasing lasso are less than 0.05 are reported, in ascending order of their p-values.

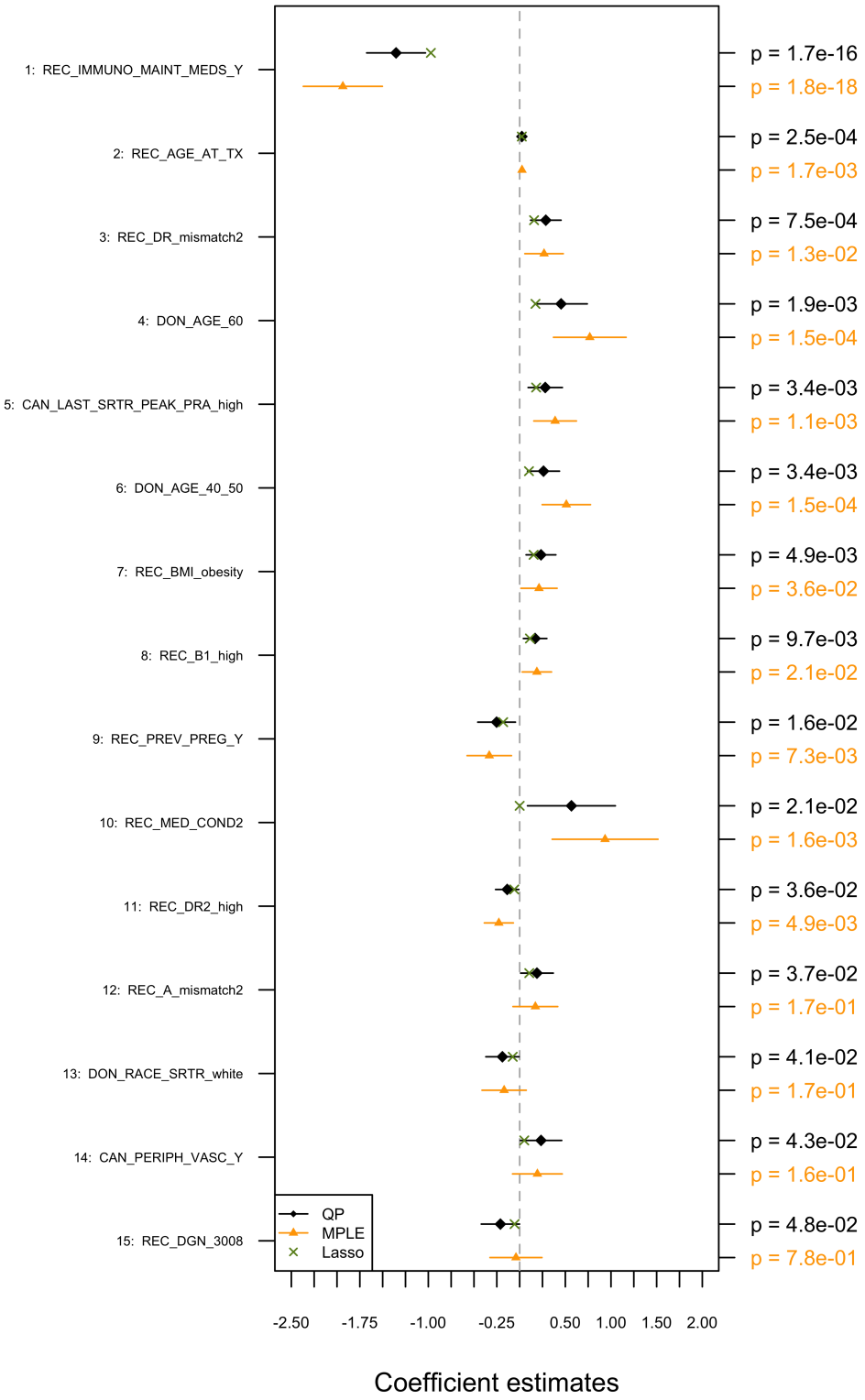

Figure 4.8: Point estimates and 95% confidence intervals for the de-biased lasso estimator via quadratic programming and the MPLE in the older than 60 age group. Only the variables whose p-values from the de-biasing lasso are less than 0.05 are reported, in ascending order of their p-values.

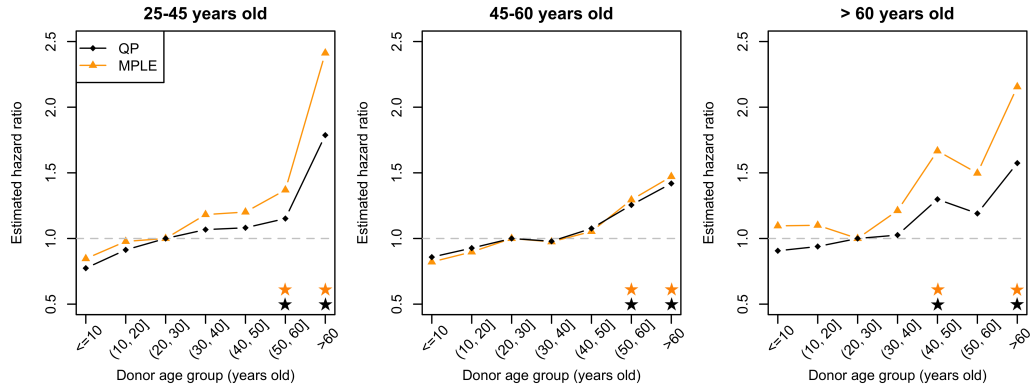

Figure 4.9: Estimated hazard ratios of donor age, compared to the reference donor age $(20, 30]$, across three recipient age groups. Color stars at the bottom indicate significant difference from the reference donor age group at level 0.05 by their corresponding methods.

Table 4.2: Estimated recipient and donor gender effects on graft survival across three recipient age groups, comparing male to the reference level of female.

		MPLE			QP		
Recipient age	Gender effect	Est	SE	P-value	Est	SE	P-value
$(25, 45]$	Recipient	0.006	0.064	0.930	3×10^{-4}	0.051	0.996
	Donor	-0.025	0.057	0.660	-0.078	0.044	0.079
$(45, 60]$	Recipient	0.050	0.060	0.405	0.047	0.049	0.339
	Donor	-0.005	0.050	0.928	-0.055	0.037	0.143
> 60	Recipient	-0.026	0.107	0.812	0.014	0.089	0.870
	Donor	0.087	0.085	0.305	0.037	0.062	0.546

4.6 Technical proofs

The proofs presented in this section are similar to those for the unstratified Cox proportional hazards model in Chapter III, but with the additional complexity of stratification. For completeness of notation, we define the counting process $N_{ki}(t) = 1(Y_{ki} \leq t, \delta_{ki} = 1)$ and the intensity process $A_{ki}(t; \beta) = \int_0^t 1(Y_{ki} \geq s) \exp(X_{ki}^T \beta) d\Lambda_{0k}(s)$, where $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s) ds$ is the baseline cumulative hazard function, $k = 1, \cdots, K,\ i = 1, \cdots, n_k$. Let $M_{ki}(t; \beta) = N_{ki}(t) - A_{ki}(t; \beta)$. $M_{ki}(t; \beta^0)$ is a martingale with respect to the filtration $\mathcal{F}_{ki}(t) = \sigma\{N_{ki}(s), 1(Y_{ki} \geq s), X_{ki} : s \in (0, t]\}$. $\{X_{ki} - \widehat{\eta}_k(t; \beta^0)\}$ is predictable with respect to the filtration $\mathcal{F}(t) = \sigma\{N_{ki}(s), 1(Y_{ki} \geq s), X_{ki} : s \in (0, t], k = 1, \cdots, K,\ i = 1, \cdots, n_k\}$.

Lemma IV.2. *Under Assumptions (A1) – (A3), for $k = 1, \cdots, K$, we have*

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

$$\sup_{t\in[0,\tau]} \|\widehat{\mu}_{1k}(t;\beta^0) - \mu_{1k}(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}),$$

$$\sup_{t\in[0,\tau]} \|\widehat{\eta}_{k}(t;\beta^0) - \eta_{k0}(t;\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n}).$$

Lemma IV.2 is simply the result of Lemma III.6 applied to each of the K strata. We omit its proof here.

Lemma IV.3. *Assume $p^2\log(p)/n_{min} \to 0$. Under Assumptions (A1) – (A5), for any $c \in \mathbb{R}^p$ such that $\|c\|_2 = 1$ and $\|c\|_1 \le a_*$ with some absolute constant $a_* > 0$,*

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

Proof of Lemma IV.3. This proof is similar to Lemma III.7, and thus we will omit some details for simplicity. We rewrite

$$\begin{aligned}
\frac{-\sqrt{N}c^T\Theta_{\beta^0}\dot{\ell}(\beta^0)}{\sqrt{c^T\Theta_{\beta^0}c}} &= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_{ki} - \frac{\widehat{\mu}_{1k}(Y_{ki};\beta^0)}{\widehat{\mu}_{0k}(Y_{ki};\beta^0)}\right\}\delta_{ki} \\
&= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^\tau \frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_{ki} - \frac{\widehat{\mu}_{1k}(t;\beta^0)}{\widehat{\mu}_{0k}(t;\beta^0)}\right\} dN_{ki}(t) \\
&= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^\tau \frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_{ki} - \frac{\widehat{\mu}_{1k}(t;\beta^0)}{\widehat{\mu}_{0k}(t;\beta^0)}\right\} dM_{ki}(t). \qquad (4.5)
\end{aligned}$$

Denote $U(t) = \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^t \frac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_{ki} - \frac{\widehat{\mu}_{1k}(s;\beta^0)}{\widehat{\mu}_{0k}(s;\beta^0)}\right\} dM_{ki}(s)$. Then the vari-

ation process for $U(t)$ is

$$\begin{aligned}\langle U\rangle(t) &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{1}{N}\int_0^t \frac{c^T\Theta_{\beta^0}}{c^T\Theta_{\beta^0}c}\left\{X_{ki}-\widehat{\eta}_k(u;\beta^0)\right\}^{\otimes 2}1(Y_{ki}\geq u)e^{X_{ki}^T\beta^0}d\Lambda_{0k}(u)\Theta_{\beta^0}c \\ &= \frac{c^T\Theta_{\beta^0}}{c^T\Theta_{\beta^0}c}\left[\sum_{k=1}^{K}\frac{n_k}{N}\int_0^t\left\{\widehat{\mu}_{2k}(u;\beta^0)-\frac{\widehat{\mu}_{1k}(u;\beta^0)\widehat{\mu}_{1k}^T(u;\beta^0)}{\widehat{\mu}_{0k}(u;\beta^0)}\right\}d\Lambda_{0k}(u)\right]\Theta_{\beta^0}c.\end{aligned} \tag{4.6}$$

By (A4), similar to the proof of Lemma III.7, we have

$$\begin{aligned}&\frac{c^T\Theta_{\beta^0}}{c^T\Theta_{\beta^0}c}\left[\int_0^t\left\{\widehat{\mu}_{2k}(u;\beta^0)-\frac{\widehat{\mu}_{1k}(u;\beta^0)\widehat{\mu}_{1k}^T(u;\beta^0)}{\widehat{\mu}_{0k}(u;\beta^0)}\right\}d\Lambda_{0k}(u)\right]\Theta_{\beta^0}c \\ =&\frac{c^T\Theta_{\beta^0}}{c^T\Theta_{\beta^0}c}\left[\int_0^t\left\{\mu_{2k}(u;\beta^0)-\frac{\mu_{1k}(u;\beta^0)\mu_{1k}^T(u;\beta^0)}{\mu_{0k}(u;\beta^0)}\right\}d\Lambda_{0k}(u)\right]\Theta_{\beta^0}c+o_P(1) \\ \to& v_k(t).\end{aligned}$$

Since $n_k/N\to r_k$, then $\langle U\rangle(t)\to_P\sum_{k=1}^K r_kv_k(t)$.

For any $\epsilon>0$, define $G_{ki}(u)=\dfrac{1}{\sqrt{N}}\dfrac{c^T\Theta_{\beta^0}}{\sqrt{c^T\Theta_{\beta^0}c}}\left\{X_{ki}-\dfrac{\widehat{\mu}_{1k}(u;\beta^0)}{\widehat{\mu}_{0k}(u;\beta^0)}\right\}$ and the truncated process $U_\epsilon(t)=\displaystyle\sum_{k=1}^K\sum_{i=1}^{n_k}\int_0^t G_{ki}(u)1(|G_{ki}(u)|>\epsilon)dM_{ki}(u)$. The variation process of $U_\epsilon(t)$ is

$$\langle U_\epsilon\rangle(t)=\sum_{k=1}^K\sum_{i=1}^{n_k}\int_0^t G_{ki}^2(u)1(|G_{ki}(u)|>\epsilon)dA_{ki}(u),$$

where $dA_{ki}(u)=1(Y_{ki}\geq u)e^{X_{ki}^T\beta^0}d\Lambda_{0k}(u)$. Since

$$|\sqrt{N}G_{ki}(u)|\leq a_*\|\Theta_{\beta^0}\|_{1,1}2M\lambda_{\min}^{-1/2}(\Theta_{\beta^0})=\mathcal{O}(\sqrt{p}),$$

$1(|G_{ki}(u)|>\epsilon)=0$ eventually as $p/N\to 0$. So $\langle U_\epsilon\rangle(t)\to_P 0$. By the martingale central limit theorem, the conclusion holds.

□

Lemma IV.4. *Under Assumptions (A1) – (A4), for* $\lambda \asymp \sqrt{\log(p)/n_{min}}$*, the lasso estimator* $\widehat{\beta}$ *satisfies*

$$\|\widehat{\beta} - \beta^0\|_1 = \mathcal{O}_P(s_0\lambda),\ \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|X_{ki}^T(\widehat{\beta} - \beta^0)|^2 = \mathcal{O}_P(s_0\lambda^2).$$

Proof of Lemma IV.4. This results from the proof in *Kong and Nan* (2014), with minor modifications. An intermediate replacement for $\ell_k(\beta)$ can be defined as

$$\widetilde{\ell}_k(\beta) = -\frac{1}{n_k}\sum_{j=1}^{n_k}\left[\beta^T X_{kj} - \log\mu_{0k}(Y_{kj};\beta)\right]\delta_{ki}.$$

The target parameter is $\bar{\beta} = \arg\min_\beta \mathbb{E}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\widetilde{\ell}_k(\beta)\right\}$. Then the excess risk for any given β is

$$\mathcal{E}(\beta) = \mathbb{E}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\widetilde{\ell}_k(\beta)\right\} - \mathbb{E}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\widetilde{\ell}_k(\bar{\beta})\right\}.$$

□

Lemma IV.5. *Under Assumptions (A1) – (A4), it holds with probability going to 1 that* $\|\Theta_{\beta^0}\widehat{\Sigma} - I_p\|_\infty \le \gamma$*, with* $\gamma \asymp \|\Theta_{\beta^0}\|_{1,1}\{\max_{1\le k\le K}|n_k/N - r_k| + s_0\lambda\}$.

Proof of Lemma IV.5. We first derive the rate for $\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty$. Note that

$$\begin{aligned}
&\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty \\
\le &\left\|\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^\tau\left[\{X_{ki} - \widehat{\eta}_k(t;\widehat{\beta})\}^{\otimes 2} - \{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2}\right]dN_{ki}(t)\right\|_\infty \\
&+ \left\|\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^\tau\{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2}dN_{ki}(t) - \Sigma_{\beta^0}\right\|_\infty \equiv a_{N1} + a_{N2}.
\end{aligned}$$

Due to the boundness condition (A1),

$$
\begin{aligned}
a_{N1} \leq & \left\| \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^\tau \{X_{ki} - \widehat{\eta}_k(t;\widehat{\beta})\}\{\eta_{k0}(t;\beta^0) - \widehat{\eta}_k(t;\widehat{\beta})\}^T dN_{ki}(t) \right\|_\infty \\
& + \left\| \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^\tau \{\eta_{k0}(t;\beta^0) - \widehat{\eta}_k(t;\widehat{\beta})\}\{X_{ki} - \eta_{k0}(t;\beta^0)\}^T dN_{ki}(t) \right\|_\infty \\
\leq & \frac{4M}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^\tau \|\eta_{k0}(t;\beta^0) - \widehat{\eta}_k(t;\widehat{\beta})\|_\infty dN_{ki}(t) \\
\leq & \frac{4M}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^\tau \|\eta_{k0}(t;\beta^0) - \widehat{\eta}_k(t;\beta^0)\|_\infty dN_{ki}(t) \\
& + \frac{4M}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^\tau \|\widehat{\eta}_k(t;\beta^0) - \widehat{\eta}_k(t;\widehat{\beta})\|_\infty dN_{ki}(t) \\
\leq & 4M\mathcal{O}_P(\sqrt{\log(p)/n_{min}}) + 4M\mathcal{O}_P(s_0\lambda) = \mathcal{O}_P(s_0\lambda),
\end{aligned}
$$

where the last inequality is a result of Lemma IV.2 and the fact that $\sup_{t\in[0,\tau]} \|\widehat{\eta}_k(t;\beta^0) - \widehat{\eta}_k(t;\widehat{\beta})\|_\infty = \mathcal{O}_P(\|\widehat{\beta} - \beta^0\|_1) = \mathcal{O}_P(s_0\lambda)$ (see the proof of Lemma III.9 in Chapter III). Since $\Sigma_{\beta^0} = \sum_{k=1}^K r_k \Sigma_{\beta^0,k}$,

$$
\begin{aligned}
a_{N2} \leq & \left\| \sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{n_k} \sum_{i=1}^{n_k} \int_0^\tau \{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2} dN_{ki}(t) - \Sigma_{\beta^0,k} \right] \right\|_\infty \\
& + \left\| \sum_{k=1}^{K} \left(\frac{n_k}{N} - r_k \right) \Sigma_{\beta^0,k} \right\|_\infty \\
\leq & \sum_{k=1}^{K} \frac{n_k}{N} \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \int_0^\tau \{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2} dN_{ki}(t) - \Sigma_{\beta^0,k} \right\|_\infty + \left\| \sum_{k=1}^{K} \left(\frac{n_k}{N} - r_k \right) \Sigma_{\beta^0,k} \right\|_\infty .
\end{aligned}
$$

The proof of Lemma III.9 in Chapter III shows that, for $k = 1, \cdots, K$,

$$
\left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \int_0^\tau \{X_{ki} - \eta_{k0}(t;\beta^0)\}^{\otimes 2} dN_{ki}(t) - \Sigma_{\beta^0,k} \right\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n_k}).
$$

So $a_{N2} = \mathcal{O}_P(\sqrt{\log(p)/n_{min}}) + \mathcal{O}(\max_k |n_k/N - r_k|)$. Therefore, $\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty = \mathcal{O}_P(s_0\lambda + \max_k |n_k/N - r_k|)$.

Finally, it is easy to see that

$$\|\Theta_{\beta^0}\widehat{\Sigma} - I_p\|_\infty = \|\Theta_{\beta^0}(\widehat{\Sigma} - \Sigma_{\beta^0})\|_\infty \leq \|\Theta_{\beta^0}\|_{1,1}\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty,$$

and $\|\Theta_{\beta^0}\widehat{\Sigma} - I_p\|_\infty = \mathcal{O}_P(\|\Theta_{\beta^0}\|_{1,1}\{s_0\lambda + \max_k |n_k/N - r_k|\})$. □

Lemma IV.6. *Assume* $\limsup_{n_{min}\to\infty} p\gamma \leq 1 - \epsilon'$ *for some* $\epsilon' \in (0, 1)$. *Then, under assumptions (A1) – (A5),* $\|\widehat{\Theta} - \Theta_{\beta^0}\|_\infty = \mathcal{O}_P(\gamma\|\Theta_{\beta^0}\|_{1,1})$.

The proof of Lemma IV.6 is identical to that of Lemma III.10 in Chapter III, and thus will be omitted.

Lemma IV.7. *Under Assumptions (A1) – (A4), for each* $t > 0$,

$$P(\|\dot{\ell}(\beta^0)\|_\infty > t) \leq 2Kpe^{-n_{min}t^2/(8M^2)}.$$

Proof of Lemma IV.7. Since $\dot{\ell}(\beta^0) = \sum_{k=1}^K \frac{n_k}{N}\dot{\ell}_k(\beta^0)$, we have

$$\begin{aligned} Pr\left(\|\dot{\ell}(\beta^0)\|_\infty > t\right) &\leq Pr\left(\sum_{k=1}^K \frac{n_k}{N}\|\dot{\ell}_k(\beta^0)\|_\infty > t\right) \\ &\leq \sum_{k=1}^K Pr\left(\|\dot{\ell}_k(\beta^0)\|_\infty > t\right) \\ &\leq \sum_{k=1}^K 2pe^{-n_k t^2/(8M^2)}. \end{aligned}$$

The last inequality holds when we apply Lemma III.11 in Chapter III for each $k = 1, \cdots, K$. □

Proof of Theorem IV.1. Let $\dot{\ell}_j(\beta)$ be the jth element of the derivative $\dot{\ell}(\beta)$. By the mean value theorem, there exists $\widetilde{\beta}^{(j)}$ between $\widehat{\beta}$ and β^0 such that $\dot{\ell}_j(\widehat{\beta}) - \dot{\ell}_j(\beta^0) =$

$\frac{\partial \dot{\ell}_j(\beta)}{\partial \beta^T}\Big|_{\beta=\widetilde{\beta}^{(j)}}(\widehat{\beta}-\beta^0)$. Denote the $p\times p$ matrix $D = \left(\frac{\partial \dot{\ell}_j(\beta)}{\partial \beta}\Big|_{\beta=\widetilde{\beta}^{(1)}}, \cdots, \frac{\partial \dot{\ell}_j(\beta)}{\partial \beta}\Big|_{\beta=\widetilde{\beta}^{(p)}}\right)^T$. By the definition of the de-biased estimator $\widehat{b}$, $c^T(\widehat{b}-\beta^0)$ can be decomposed as

$$\begin{aligned} c^T(\widehat{b}-\beta^0) &= -c^T\Theta_{\beta^0}\dot{\ell}(\beta^0) - c^T(\widehat{\Theta}-\Theta_{\beta^0})\dot{\ell}(\beta^0) \\ &\quad - c^T(\widehat{\Theta}\widehat{\Sigma}-I_p)(\widehat{\beta}-\beta^0) + c^T\widehat{\Theta}(\widehat{\Sigma}-D)(\widehat{\beta}-\beta^0) \\ &= -c^T\Theta_{\beta^0}\dot{\ell}(\beta^0) + (i) + (ii) + (iii), \end{aligned}$$

where $(i) = -c^T(\widehat{\Theta}-\Theta_{\beta^0})\dot{\ell}(\beta^0), (ii) = -c^T(\widehat{\Theta}\widehat{\Sigma}-I_p)(\widehat{\beta}-\beta^0)$ and $(iii) = c^T\widehat{\Theta}(\widehat{\Sigma}-D)(\widehat{\beta}-\beta^0)$.

We first show $\sqrt{N}(i) = o_P(1)$ and $\sqrt{N}(ii) = o_P(1)$. By Lemma IV.6 and Lemma IV.7,

$$\begin{aligned} \sqrt{N}(i) &\le \sqrt{N}\|c\|_1 \cdot \|\widehat{\Theta}-\Theta_{\beta^0}\|_{\infty,\infty} \cdot \|\dot{\ell}(\beta^0)\|_\infty \\ &\le \sqrt{N}a_*\mathcal{O}_P(p\gamma\|\Theta_{\beta^0}\|_{1,1})\mathcal{O}_P(\sqrt{\log(p)/n_{min}}) \\ &= \mathcal{O}_P(\|\Theta_{\beta^0}\|_{1,1}p\gamma\sqrt{\log(p)}) \\ &= o_P(1). \end{aligned}$$

By Lemma IV.4,

$$\begin{aligned} \sqrt{N}(ii) &\le \sqrt{N}\|c\|_1\|(\widehat{\Theta}\widehat{\Sigma}-I_p)(\widehat{\beta}-\beta^0)\|_\infty \\ &\le \sqrt{N}a_*\|\widehat{\Theta}\widehat{\Sigma}-I_p\|_\infty\|\widehat{\beta}-\beta^0\|_1 \\ &\le \sqrt{N}a_*\gamma\|\widehat{\beta}-\beta^0\|_1 \\ &= \mathcal{O}_P(\sqrt{N}\gamma s_0\lambda) \\ &= o_P(1). \end{aligned}$$

We then show that $\sqrt{N}(iii) = o_P(1)$. Note that $\widehat{\Sigma}-D = (\widehat{\Sigma}-\Sigma_{\beta^0}) + (\Sigma_{\beta^0} - \ddot{\ell}(\beta^0)) + (\ddot{\ell}(\beta^0)-D)$. By the proof of Lemma IV.5, we see that with $\lambda \asymp \sqrt{\log(p)/n_{min}}$,

$\|\widehat{\Sigma} - \Sigma_{\beta^0}\|_\infty = \mathcal{O}_P(s_0\lambda + \max_k |n_k/N - r_k|)$. Based on the proof of Theorem III.1 in Chapter III, for each stratum, $\|\ddot{\ell}^{(k)}(\beta^0) - D^{(k)}\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n_k})$, where $D^{(k)} = \left(\frac{\partial \dot{\ell}_j^{(k)}(\beta)}{\partial \beta}\Big|_{\beta=\widetilde{\beta}^{(1)}}, \cdots, \frac{\partial \dot{\ell}_j^{(k)}(\beta)}{\partial \beta}\Big|_{\beta=\widetilde{\beta}^{(p)}} \right)^T$. Since the overall negative log partial likelihood $\ell(\beta) = \sum_{k=1}^K \frac{n_k}{N}\ell^{(k)}(\beta)$, $\|\ddot{\ell}(\beta^0) - D\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n_{min}})$. Also, $\|\Sigma_{\beta^0,k} - \ddot{\ell}^{(k)}(\beta^0)\|_\infty = \mathcal{O}_P(\sqrt{\log(p)/n_k})$. Then

$$\begin{aligned}
\|\Sigma_{\beta^0} - \ddot{\ell}(\beta^0)\|_\infty &\le \left\| \sum_{k=1}^K r_k \Sigma_{\beta^0,k} - \sum_{k=1}^K \frac{n_k}{N}\Sigma_{\beta^0,k} \right\|_\infty + \left\| \sum_{k=1}^K \frac{n_k}{N}\Sigma_{\beta^0,k} - \sum_{k=1}^K \frac{n_k}{N}\ddot{\ell}^{(k)}(\beta^0) \right\|_\infty \\
&\le K \max_k (|n_k/N - r_k| \|\Sigma_{\beta^0,k}\|_\infty) + K\mathcal{O}_P(\sqrt{\log(p)/n_{min}}) \\
&= \mathcal{O}_P(\max_k |n_k/N - r_k| + \sqrt{\log(p)/n_{min}}).
\end{aligned}$$

Therefore, for $\lambda \asymp \sqrt{\log(p)/n_{min}}$, $\|\widehat{\Theta} - D\|_\infty = \mathcal{O}_P(s_0\lambda + \max_k |n_k/N - r_k|)$, and

$$\begin{aligned}
|\sqrt{N}(iii)| &\le \sqrt{N}\|c\|_1 \|\widehat{\Theta}\|_{\infty,\infty} \|\widehat{\Sigma} - D\|_\infty \|\widehat{\beta} - \beta^0\|_1 \\
&\le \mathcal{O}_P(\sqrt{N}\|\Theta_{\beta^0}\|_{1,1}(s_0^2\lambda^2 + s_0\lambda \max_k |n_k/N - r_k|)) = o_P(1).
\end{aligned}$$

Finally, for the variance,

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

By Slutsky's theorem and Lemma IV.3, $\sqrt{n}c^T(\widehat{b} - \beta^0)/(c^T\widehat{\Theta}c)^{1/2} \xrightarrow{\mathcal{D}} N(0,1)$. □

CHAPTER V

Summary and Future Work

This dissertation has focused on the development of practically useful and theoretically sound statistical methods, based on the idea of de-biasing the lasso estimator, for drawing reliable inference on the challenging cases of regression models with diverging numbers of covariates – the generalized linear models in Chapter II, the unstratified and stratified Cox proportional hazards models in Chapter III and Chapter IV, respectively. As we have pointed out, many existing methods that can handle the "large p, small n" scenario require sparse matrix estimation in their implementations and sparsity assumptions in the corresponding inverse information matrices in the theoretical developments, the latter of which lack practical interpretations and can hardly hold in general settings. The resulting estimators often have large residual biases in practice, leading to poor confidence interval coverage and hypothesis testing results. Although the theories we developed for our proposed methods reside in the "large n, diverging p" scenario, they do not require the unrealistic sparsity assumptions on the inverse information matrices and are shown to outperform their competitors. This dissertation would make a great addition to the literature of high-dimensional inference and careful reflections on the trade-off between high-dimensionality and practicality. It would also be worthwhile to pursue the completion of theories with possible modifications to the proposed de-biasing lasso methods, under the "large p,

small n" scenario but without the matrix sparsity assumptions.

The methods developed in this dissertation are based on the assumption of correct model specification, i.e. the relationship between outcomes of interest and covariates is correctly specified. In fact, estimating equations provide a generalization to many classical estimation methods in that correct specification is only needed for a few moments instead of the entire distribution (*Godambe*, 1991). One promising future direction is to extend the de-biasing lasso approach with quadratic programming discussed in Chapters III and IV to high-dimensional inference for more general estimating equations.

Often, intrinsic group structures are present among covariates of interest, for example, within biological pathways or indicators of different levels of a categorical variable. Even though our theories justify making inference on multiple linear combinations simultaneously, including a group of variables as a special case, the number of linear combinations allowed is fixed. Group lasso (*Yuan and Lin*, 2006) and its variations enable grouped variable selection in an "all-in-all-out" fashion. *Mitra and Zhang* (2016) has discussed the benefit of utilizing group sparsity in group inference via a scaled group lasso in linear regression. Another future direction is to incorporate the prior knowledge of group structures in the covariates beyond linear regression, and explore whether adding group penalties in the variable selection stage could improve the reliability of inference results on potentially a larger group of variables.

Efficient computation is a key component in the era of big data. Even though we have demonstrated that our proposed method using quadratic programming is more computationally efficient than one closely related competitor, CLIME, fast computation still remains as a challenging problem in the scale of biobank or electronic health records data, usually with up to a few hundred thousands of records and millions of biomarkers. Especially, survival models pose additional difficulties due to the computation of at-risk sets. It would make a great contribution to the world of

data applications to modify the proposed algorithms in this dissertation and develop computationally efficient softwares that can deliver reliable inference in the analysis of such large scale data.

BIBLIOGRAPHY

BIBLIOGRAPHY

Amos, C. I., et al. (2008), Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1, *Nature Genetics*, *40*(5), 616–622.

Andersen, P. K., and R. D. Gill (1982), Cox's regression model for counting processes: A large sample study, *The Annals of Statistics*, *10*(4), 1100–1120.

Antoniadis, A., P. Fryzlewicz, and F. Letué (2010), The Dantzig selector in Cox's proportional hazards model, *Scandinavian Journal of Statistics*, *37*(4), 531–552.

Bossé, Y., and C. I. Amos (2018), A decade of GWAS results in lung cancer, *Cancer Epidemiology, Biomarkers & Prevention*, *27*(4), 363–379.

Bühlmann, P., and S. van de Geer (2011), *Statistics for high-dimensional data: methods, theory and applications*, Heidelberg: Springer.

Cai, T., W. Liu, and X. Luo (2011), A constrained ℓ_1 minimization approach to sparse precision matrix estimation, *Journal of the American Statistical Association*, *106*(494), 594–607.

Cai, T. T., A. Zhang, and Y. Zhou (2019), Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference, *arXiv preprint arXiv:1909.09851*.

Candes, E., and T. Tao (2007), The dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, *35*(6), 2313–2351.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Dayoub, J. C., F. Cortese, A. Anžič, T. Grum, and J. P. de Magalhães (2018), The effects of donor age on organ transplants: A review and implications for aging research, *Experimental Gerontology*, *110*, 230–240.

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

Doyle, G. A., M.-J. Wang, A. D. Chou, J. U. Oleynick, S. E. Arnold, R. J. Buono, T. N. Ferraro, and W. H. Berrettini (2011), *In Vitro* and *Ex Vivo* analysis of *CHRNA3* and *CHRNA5* haplotype expression, *PloS One*, *6*(8), e23,373.

Eftekhari, H., M. Banerjee, and Y. Ritov (2019), Inference in general single-index models under high-dimensional symmetric designs, *arXiv preprint arXiv:1909.03540*.

Evans, W. E., and M. V. Relling (2004), Moving towards individualized medicine with pharmacogenomics, *Nature*, *429*(6990), 464–468.

Fan, J., and R. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., and R. Li (2002), Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, *30*(1), 74–99.

Fan, J., and H. Peng (2004), Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, *32*(3), 928–961.

Fang, E. X., Y. Ning, and H. Liu (2017), Testing and confidence intervals for high dimensional proportional hazards models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(5), 1415–1437.

Friedman, J., T. Hastie, and R. Tibshirani (2010), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, *33*(1), 1–22.

Gabrielsen, M. E., P. Romundstad, A. Langhammer, H. E. Krokan, and F. Skorpen (2013), Association between a 15q25 gene variant, nicotine-related habits, lung cancer and COPD among 56307 individuals from the HUNT study in Norway, *European Journal of Human Genetics*, *21*(11), 1293–1299.

Godambe, V. P. (1991), *Estimating functions*, New York: Oxford University Press.

Grambsch, P. M., and T. M. Therneau (1994), Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, *81*(3), 515–526.

Guan, Y., and M. Stephens (2011), Bayesian variable selection regression for genome-wide association studies and other large-scale problems, *The Annals of Applied Statistics*, *5*(3), 1780–1815.

Gui, J., and H. Li (2005), Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, *Bioinformatics*, *21*(13), 3001–3008.

Halldén, S., M. Sjögren, B. Hedblad, G. Engström, V. Hamrefors, J. Manjer, and O. Melander (2016), Gene variance in the nicotinic receptor cluster (CHRNA5-CHRNA3-CHRNB4) predicts death from cardiopulmonary disease and cancer in smokers, *Journal of Internal Medicine*, *279*(4), 388–398.

He, K., V. B. Ashby, and D. E. Schaubel (2019), Evaluating center-specific long-term outcomes through differences in mean survival time: Analysis of national kidney transplant data, *Statistics in Medicine*, *38*(11), 1957–1967.

He, Q., and D.-Y. Lin (2010), A variable selection method for genome-wide association studies, *Bioinformatics*, *27*(1), 1–8.

Hershberger, P. A., A. C. Vasquez, B. Kanterewicz, S. Land, J. M. Siegfried, and M. Nichols (2005), Regulation of endogenous gene expression in human non-small cell lung cancer cells by estrogen receptor ligands, *Cancer Research*, *65*(4), 1598–1605.

Houston, K. A., K. A. Mitchell, J. King, A. White, and B. M. Ryan (2018), Histologic lung cancer incidence rates and trends vary by race/ethnicity and residential county, *Journal of Thoracic Oncology*, *13*(4), 497–509.

Huang, J., T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang (2013), Oracle inequalities for the lasso in the Cox model, *Annals of Statistics*, *41*(3), 1142–1165.

Hung, R. J., et al. (2008), A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25, *Nature*, *452*(7187), 633–637.

Janssen-Heijnen, M. L., and J.-W. W. Coebergh (2001), Trends in incidence and prognosis of the histological subtypes of lung cancer in North America, Australia, New Zealand and Europe, *Lung Cancer*, *31*(2-3), 123–137.

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

Jia, D., et al. (2018), Crebbp loss drives small cell lung cancer and increases sensitivity to HDAC inhibition, *Cancer Discovery*, *8*(11), 1422–1437.

Kalbfleisch, J. D., and R. L. Prentice (2002), *The statistical analysis of failure time data*, Hoboken: John Wiley & Sons.

Kasiske, B. L., and J. Snyder (2002), Matching older kidneys with older patients does not improve allograft survival, *Journal of the American Society of Nephrology*, *13*(4), 1067–1072.

Kong, S., and B. Nan (2014), Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso, *Statistica Sinica*, *24*(1), 25–42.

Kong, S., Z. Yu, X. Zhang, and G. Cheng (2018), High dimensional robust inference for Cox regression models, *arXiv preprint arXiv:1811.00535*.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016), Exact post-selection inference, with application to the lasso, *The Annals of Statistics*, *44*(3), 907–927.

Legendre, C., G. Canaud, and F. Martinez (2014), Factors influencing long-term outcome after kidney transplantation, *Transplant International*, *27*(1), 19–27.

Li, Y., L. Dicker, and S. D. Zhao (2014), The Dantzig selector for censored linear regression models, *Statistica Sinica*, *24*(1), 251–268.

McKay, J. D., et al. (2017), Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes, *Nature Genetics*, *49*(7), 1126–1132.

Meinshausen, N., and P. Bühlmann (2006), High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics*, *34*(3), 1436–1462.

Mitra, R., and C.-H. Zhang (2016), The benefit of group sparsity in group inference with de-biased scaled group lasso, *Electronic Journal of Statistics*, *10*(2), 1829–1873.

Morris, E., K. He, Y. Li, Y. Li, and J. Kang (2018), Survboost: An R package for high-dimensional variable selection in the stratified proportional hazards model via gradient boosting, *arXiv preprint arXiv:1803.07715*.

Ning, Y., and H. Liu (2017), A general theory of hypothesis tests and confidence regions for sparse high dimensional models, *The Annals of Statistics*, *45*(1), 158–195.

Opelz, G. (1994), Effect of the maintenance immunosuppressive drug regimen on kidney transplant outcome., *Transplantation*, *58*(4), 443–446.

Pintarelli, G., A. Galvan, P. Pozzi, S. Noci, G. Pasetti, F. Sala, U. Pastorino, R. Boffi, and F. Colombo (2017), Pharmacogenetic study of seven polymorphisms in three nicotinic acetylcholine receptor subunits in smoking-cessation therapies, *Scientific Reports*, *7*(1), 16,730.

Purcell, S., et al. (2007), PLINK: a tool set for whole-genome association and population-based linkage analyses, *The American Journal of Human Genetics*, *81*(3), 559–575.

Qiu, L.-X., L. Yao, K. Xue, J. Zhang, C. Mao, B. Chen, P. Zhan, H. Yuan, and X.-C. Hu (2010), BRCA2 N372H polymorphism and breast cancer susceptibility: a meta-analysis involving 44,903 subjects, *Breast Cancer Research and Treatment*, *123*(2), 487–490.

Repapi, E., et al. (2010), Genome-wide association study identifies five loci associated with lung function, *Nature genetics*, *42*(1), 36.

Rodger, R. S. C. (2012), Approach to the management of end-stage renal disease, *Clinical Medicine*, *12*(5), 472–475.

Stevens, V. L., L. J. Bierut, J. T. Talbot, J. C. Wang, J. Sun, A. L. Hinrichs, M. J. Thun, A. Goate, and E. E. Calle (2008), Nicotinic receptor gene variants influence susceptibility to heavy smoking, *Cancer Epidemiology and Prevention Biomarkers*, *17*(12), 3517–3525.

Sur, P., and E. J. Candès (2019), A modern maximum-likelihood theory for high-dimensional logistic regression, *Proceedings of the National Academy of Sciences*, *116*(29), 14,516–14,525.

Tang, D., et al. (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*(1), 104–115.

Taylor, J. G., E.-H. Choi, C. B. Foster, and S. J. Chanock (2001), Using genetic variation to study human disease, *Trends in Molecular Medicine*, *7*(11), 507–512.

Thorgeirsson, T. E., et al. (2008), A variant associated with nicotine dependence, lung cancer and peripheral arterial disease, *Nature*, *452*(7187), 638–642.

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

Tibshirani, R. (1997), The lasso method for variable selection in the Cox model, *Statistics in Medicine*, *16*(4), 385–395.

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

van de Geer, S. A. (2008), High-dimensional generalized linear models and the lasso, *The Annals of Statistics*, *36*(2), 614–645.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W., and J. A. Wellner (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Heidelberg: Springer.

van Houwelingen, H. C., T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. Wessels (2006), Cross-validated Cox regression on microarray gene expression data, *Statistics in Medicine*, *25*(18), 3201–3216.

Veroux, M., G. Grosso, D. Corona, A. Mistretta, A. Giaquinta, G. Giuffrida, N. Sinagra, and P. Veroux (2012), Age is an important predictor of kidney transplantation outcome, *Nephrology Dialysis Transplantation*, *27*(4), 1663–1671.

Vershynin, R. (2010), Introduction to the non-asymptotic analysis of random matrices, *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2012), How close is the sample covariance matrix to the actual covariance matrix?, *Journal of Theoretical Probability*, *25*(3), 655–686.

Wang, L. (2011), GEE analysis of clustered binary data with diverging number of covariates, *The Annals of Statistics*, *39*(1), 389–417.

Wang, S., A. D. van der Vaart, Q. Xu, C. Seneviratne, O. F. Pomerleau, C. S. Pomerleau, T. J. Payne, J. Z. Ma, and M. D. Li (2014), Significant associations of CHRNA2 and CHRNA6 with nicotine dependence in European American and African American populations, *Human Genetics*, *133*(5), 575–586.

Wolfe, R. A., V. B. Ashby, E. L. Milford, A. O. Ojo, R. E. Ettenger, L. Y. Agodoa, P. J. Held, and F. K. Port (1999), Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant, *New England Journal of Medicine*, *341*(23), 1725–1730.

Yu, H., et al. (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*(4), 398–407.

Yu, Y., J. Bradic, and R. J. Samworth (2018), Confidence intervals for high-dimensional Cox models, *arXiv preprint arXiv:1803.01150*.

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

Zhang, C.-H., and S. S. Zhang (2014), Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.

Zhang, H. H., and W. Lu (2007), Adaptive lasso for cox's proportional hazards model, *Biometrika*, *94*(3), 691–703.

Zhang, X., and G. Cheng (2017), Simultaneous inference for high-dimensional linear models, *Journal of the American Statistical Association*, *112*(518), 757–768.

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

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