Title
Scalable Methods for Big Time-To-Event Data

Author
Kawaguchi, Eric Shinya

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Scalable Methods for Big Time-To-Event Data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Eric Shinya Kawaguchi

2019

ABSTRACT OF THE DISSERTATION

Scalable Methods for Big Time-To-Event Data

by

Eric Shinya Kawaguchi
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2019
Professor Gang Li, Chair

Computational advancements and cost efficiency over the recent years have made big data readily available to researchers. In the biomedical and public health fields analyzing time-to-event data, where the outcome of interest is a time-to-event endpoint, is of particular interest. However, big time-to-event data poses many challenges to currently-available statistical methods due to the large number of covariates and/or observations one can observe. In this dissertation we propose scalable sparse regression methods for both big right-censored and competing risks time-to-event data. We extend the recently-introduced broken adaptive ridge (BAR) regression procedure to both the Cox (1972) proportional hazards for right-censored data and the Fine and Gray (1999) proportional subdistribution hazards model for competing risks data, establish its large-sample properties under diverging dimension, and develop computational software that is scalable to big time-to-event data.

The dissertation of Eric Shinya Kawaguchi is approved.

Xinshu Xiao

Hua Zhou

Marc Adam Suchard

Gang Li, Committee Chair

University of California, Los Angeles

2019

To God, for blessing me with this wonderful opportunity and for placing such incredible mentors, colleagues, and friends into my life. To my parents, Douglas and Mari, and sister, Traci, for encouraging me to explore and pursue my interests and passions. To my wife, Lucia, for her comfort, support, and love throughout this endeavor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation advisor, Dr. Gang Li, for his guidance, advice, support, encouragement, and most importantly, patience, during my academic career at UCLA. I have learned tremendously under his direction. Many thanks are also due to my committee members: Drs. Marc A. Suchard, Hua Zhou, and Xinshu Xiao for their valuable comments, insights, and suggestions. I greatly enjoyed getting to know them during this process. I would like to extend my thanks to Drs. Randall Burd and Sushil Mittal for access to the National Trauma Data Bank dataset used in Chapter 3 and to Dr. Jenny I. Shen for access to the United States Renal Data Systems dataset used in Chapters 4 and 5. Lastly, I am also very thankful for my fellow cohort members and colleagues, I could not imagine starting and finishing this journey without you all.

VITA

2009–2013	B.S. (Mathematics), California State Polytechnic University, Pomona, California.
2013–2015	M.S. (Biostatistics), University of California, Los Angeles, California.
2014–present	Graduate Student Researcher, Department of Biostatistics, UCLA.
2015–present	Teaching Assistant, Department of Biostatistics, UCLA.

CHAPTER 1

Introduction

Advancing informatics tools make big time-to-event data routinely accessible to biomedical researchers. This data deluge offers unprecedented opportunities for new and innovative approaches to improve research and learning (Schuemie et al., 2017) but also presents new computational challenges and barriers for quantitative researchers as many current statistical methodologies and computational tools may grind to a halt as the sample size (n) grows large. Such challenges are particularly common in time-to-event data analyses where the log-likelihood function for commonly-used semi-parametric regression models (such as the Cox (1972) proportional hazards model for right-censored data or Fine and Gray (1999) proportional subdistribution hazards model for competing risks data) and its derivatives typically require $O(n^2)$ number of operations, which will quickly explode as n grows large. The computational burden can be further aggravated as the number of covariates (p_n) increases since 1) the computational cost is multiplied by a factor of p_n for the gradient and p_n^2 for the Hessian matrix, and 2) in addition to estimation, variable selection would add another layer of computational complexity. Statistical methods coupled with high-performance algorithms are critically needed for big time-to-event data analysis.

Generally, not all of the covariates we obtain are expected to be relevant to the outcome of interest. Oftentimes researchers are interested in identifying covariates which have an effect on the outcome. Penalization methods (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Zhang et al., 2010) offer a popular way to perform simultaneous variable selection and parameter estimation through minimizing a penalized objective function. Several methods have been proposed for the Cox proportional hazards model (Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Zhang, 2010; Simon et al., 2011; Johnson et al., 2012; Su et al., 2016) and,

more recently, the Fine-Gray proportional subdistribution hazards model (Ha et al., 2014; Fu et al., 2017; Ahn et al., 2018; Hou et al., 2018).

It is well known that ℓ_0-penalized regression is natural for variable selection and parameter estimation with some optimal properties (Akaike, 1974; Schwarz, 1978; Volinsky and Raftery, 2000; Shen et al., 2012), but also known to have some limitations such as being unstable (Breiman, 1996) and unscalable to high-dimensional settings. The broken adaptive ridge (BAR) estimator, defined as the limit of an ℓ_0-based iteratively reweighted ℓ_2-penalization algorithm, has been recently introduced for simultaneous variable selection and parameter estimation and shown to possess some desirable selection and estimation properties under several model settings (see, e.g. Zhao et al. (2018), Dai et al. (2018), Zhao et al. (2019), and Zhao et al. (2019)). The idea of iteratively reweighted penalizations dates back at least to the well-known Lawson's algorithm (Lawson, 1961) in classical approximation theory, which has been applied to various applications including ℓ_d $(0 < d < 1)$ minimization (Osborne, 1985), sparse signal reconstruction (Gorodnitsky and Rao, 1997), compressive sensing (Candes et al., 2008; Chartrand and Yin, 2008; Gasso et al., 2009; Daubechies et al., 2010; Wipf and Nagarajan, 2010), and variable selection for linear models and generalized linear models (Liu and Li, 2016; Frommlet and Nuel, 2016). The BAR method aims to yield a local solution of ℓ_0-penalized regression that preserves some desirable properties of ℓ_0-penalized regression while avoiding its limitations. First, the BAR estimator is stable and easily scalable to high-dimensional covariates. Second, the BAR estimator has a grouping property for highly-correlated covariates. Lastly, the BAR estimator enjoys the best of ℓ_0-penalized regression and the oracle ridge estimator. Specifically, the reweighted ridge regression at each iteration step shrinks the small values of the initial ridge estimator towards zero and drives its large values towards an oracle ridge estimator. Thus the resulting BAR estimator is selection consistent and its nonzero component behaves like the oracle ridge estimator in that it is asymptotically consistent and Gaussian.

Developing efficient algorithms is crucial in handling large-scale (massive sample size) time-to-event data. We give two examples of such datasets and potential obstacles one may encounter.

- **National Trauma Data Bank**: Sparse high-dimensional massive sample size (sHDMSS) data is a particular type of big data with the following characteristics: 1) high-dimensional with a large number of covariates (p_n in thousands or tens of thousands), 2) massive in sample-size (n in thousands to hundreds of millions), and 3) sparse in covariates with only a very small portion of covariates being nonzero for each subject. For sHDMSS time-to-event data, we also have the issue of rare events (i.e. high right censoring). A typical example of sHDMSS time-to-event data is the pediatric trauma mortality data (Mittal et al., 2014) from the National Trauma Data Bank (NTDB) maintained by the American College of Surgeons (Mittal et al., 2014). This data set includes 210,555 patient records of injured children under 15 collected over 5 years from 2006 -2010. Each patient record includes 125,952 binary covariates that indicate the presence, or absence, of an attribute (ICD9 Codes, AIS codes, etc.) as well as their two-way interactions. The data matrix is extremely sparse with less than 1% of the covariates being non zero. The event rate is also very low at 2%. The massive sample size presents a critical barrier to the application of existing sparse time-to-event regression methods in a high-dimensional setting.

  While many sparse time-to-event regression methods (Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Zhang, 2010; Simon et al., 2011; Johnson et al., 2012; Su et al., 2016) are available, current methods and standard software become inoperable for large datasets due to high computational costs and large memory requirements. Mittal et al. (2014) presented tools for fitting ℓ_2- (ridge) and ℓ_1- (LASSO) penalized Cox's regressions on sHDMSS data. However, it is well known that ridge regression is not sparse and that although LASSO produces a sparse solution, it tends to select too many noise variables and is biased for estimation. Lastly, the commonly used "divide and conquer" strategy for massive-size data is deemed inappropriate for sHDMSS time-to-event data since each of the divided data would typically be too sparse for a meaningful analysis.

- **United States Renal Data Systems**: The United States Renal Data System (USRDS) is a national data system funded by the National Institute of Diabetes and

Digestive and Kidney Diseases (NIDDK) that collects information about end-stage renal disease in the United States. Patients with end-stage renal disease are known to have a shorter life expectancy compared to their disease-free peers (USRDS Annual Report 2017) and kidney transplantation provides better health outcomes for patients with end-stage renal disease (Wolfe et al., 1999; Purnell et al., 2016). However patients may observe competing events such as death or renal function recovery or may wish to discontinue dialysis for quality of life purposes before transplant.

While the number of demographic and clinical covariates is relatively small, the number of subjects can easily exceed hundreds of thousands. Furthermore, the competing risks nature of this dataset makes scalable computing particularly challenging. Current methods calculate key components for parameter estimation in $O(n^2)$ calculations, which prohibits its use for data with massive sample sizes. For example analyzing a subset of 125,000 subjects, a fraction of the data available from the USRDS, with 63 covariates takes over one day to finish.

In addressing the above challenges, the key contributions of this dissertation is four-fold:

1. **Methodology**: We extend the BAR methodology to both the Cox proportional hazards model (Chapter 3) and Fine-Gray subdistribution hazards model (Chapter 4) and rigorously study its asymptotic properties. Specifically, we show that, for each model, the BAR estimator is selection consistent and possesses an oracle property in the sense that with probability tending to 1, it estimates the zero coefficients as zeros and estimates the non-zero coefficients as if the true sub-model is known in advance. Further, we prove that the BAR estimator retains the ℓ_2-property of grouping highly-correlated covariates. The theoretical guarantees are derived in the diverging dimension scenario for both models. Unlike most penalized regression methods that produce a sparse solution in a single step, the BAR method is not sparse, per se, at each iteration and only achieves sparsity at its limit. Consequently, our theoretical derivations for the BAR estimator are quite different from those for a single-step oracle estimator in the literature. Derivations are further complicated due to the log likelihood no longer being a

sum of i.i.d. random variables and the standard martingale central limit theorem does not apply when the number of parameters diverges. We also assess its finite-sample operating characteristics along with other popular ℓ_1-penalization methods.

2. **Extending BAR to sparse high dimensional massive sample size (sHDMSS) right-censored data**: Except for the linear model, current BAR algorithms are not readily applicable to handle sHDMSS data. In Chapter 3, we implement an efficient algorithm to apply BAR to sHDMSS data. The iterative reweighted ℓ_2 nature of our estimator allows us to adapt existing efficient massive ℓ_2-penalized Cox regression techniques. To this end, we implement BAR regression by imbedding an adaptive version of Mittal et al. (2014)'s massive Cox's ridge regression within each iteration of the iteratively reweighted Cox's ridge regression, allowing us to extend the reach of our algorithm to the sHDMSS domain.

3. **A novel cyclic coordinate-wise BAR algorithm**: We propose a novel cyclic coordinate-wise update algorithm, referred to as cycBAR, by deriving a coordinate-wise update for a fixed point problem whose unique solution is the BAR estimator. The cycBAR algorithm computes the BAR estimator without actually carrying out iteratively reweighted ℓ_2-penalizations, resulting in substantial gains in computational efficiency. Obviously, the cycBAR method is of interest on its own since its application can be immediately applied to accelerate the BAR method for a variety of models and data settings such as the linear model, generalized linear models, various time-to-event models, as well as in other applications such as sparse signal reconstruction (Gorodnitsky and Rao, 1997) and compressive sensing (Candes et al., 2008; Chartrand and Yin, 2008; Gasso et al., 2009; Daubechies et al., 2010; Wipf and Nagarajan, 2010) where the ℓ_0-based iteratively reweighted ℓ_2-penalization algorithm are popularly used. We introduce and incorporate this algorithm in Chapter 4 for the Fine-Gray model.

4. **Linearizing parameter estimation for the Fine-Gray model**: As mentioned earlier, calculating the log-pseudo likelihood function and its derivatives typically require $O(n^2)$ number of operations. Commonly-used computational implementations quickly

become inoperable or grind to a halt for massive n. For right-censored time-to-event data Mittal et al. (2014), among others, have made significant progress in reducing the computational complexity for the Cox proportional hazards model from $O(n^2)$ to $O(n)$ by taking advantage of the cumulative structure of the risk set. However, the counterfactual construction of the risk set for the Fine-Gray model does not retain the same structure and presents a barrier to reducing the complexity of the risk set calculation. To the best of our knowledge, no further advancements in reducing the computational complexity required for calculating the subject-specific risk sets exists. By taking advantage of the ordering of the data and the special structure of both the risk set and the subject specific weight functions associated with the Fine-Gray log-pseudo likelihood and its derivatives, we derive a novel *forward-backward scan* algorithm to reduce their computational costs from $O(n^2)$ to $O(n)$, allowing for scalable analyses of competing risks data. We incorporate this algorithm to BAR estimation (Chapter 4) and expands its application to unpenalized and penalized Fine-Gray and cumulative incidence function estimation (Chapter 5).

The rest of the dissertation is organized as follows. We present a brief literature review in Chapter 2. In Chapter 3 we define the BAR estimator for the Cox proportional hazards model, establish its large-sample properties for diverging dimension, and introduce an efficient algorithm to tackle sHDMSS time-to-event data. Then in Chapter 4, we extend the methodology and theory to the Fine-Gray proportional subdistribution hazards model for competing risks data and develop both a novel cyclic coordinate-wise update algorithm (cycBAR) for the BAR estimator and a forward-backward scan algorithm for linearizing parameter estimation. Finally, Chapter 5 extends the forward-backward scan introduced in Chapter 4 for parameter and cumulative incidence function estimation of unpenalized and penalized Fine-Gray regression.

CHAPTER 2

Preliminaries and literature review

The purpose of this chapter is to familiarize readers with the underlying methods that will be presented in this dissertation. We briefly review the literature on the following topics: 1) the Cox proportional hazards model for right-censored time-to-event data; 2) penalized variable selection procedures for the Cox proportional hazards model; and 3) the Fine-Gray proportional subdistribution hazards model for competing risks time-to-event data.

2.1 Modeling the hazard function for right-censored time-to-event data

The hazard function is a quantity of interest when studying right-censored time-to-event data. Letting T be the time to event, we define the hazard function at time t as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{2.1}$$

The Cox (1972) proportional hazards model is the most widely-used model to draw inference about the covariate effect on the hazard function. For a cohort of n independent individuals, let T_i be the event time of interest, C_i be the censoring time, and $\mathbf{z}_i(\cdot) = (z_{i1}(\cdot), \ldots, z_{ip_n}(\cdot))'$ be a p_n-dimensional, possible time dependent, covariate vector. Thus, one observes the following n independent and identically distributed triplets, $\{(X_i, \delta_i, \mathbf{z}_i(\cdot))\}_{i=1}^n$, where $X_i = T_i \wedge C_i$ is the observed event time, and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator where $a \wedge b = \min(a, b)$ and $I(\cdot)$ being an indicator function. It is assumed that for all $i = 1, \ldots, n$, T_i and C_i are independent conditional on $\mathbf{z}_i(\cdot)$.

Cox (1972) proposed to model covariate effects on the conditional hazard function, $h\{t|\mathbf{z}(t)\}$, through the proportional hazards model:

$$h\{t|\mathbf{z}(t)\} = h_0(t)\exp\{\mathbf{z}(t)'\boldsymbol{\beta}\}, \tag{2.2}$$

where $h_0(t)$ is an unspecified baseline hazard and $\boldsymbol{\beta}$ is a p_n-dimensional vector of regression coefficients. Cox (1975) introduced the partial likelihood

$$L_n(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{\frac{\exp\{\mathbf{z}_i(t)'\boldsymbol{\beta}\}}{\sum_{j\in R_i}\exp\{\mathbf{z}_j(t)'\boldsymbol{\beta}\}}\right\}^{\delta_i}, \tag{2.3}$$

where $R_i = \{j : X_j \geq X_i\}$ is the set of those at risk at the ith event time.

Andersen and Gill (1982) define the log-partial likelihood for (2.2) as

$$l_n(\boldsymbol{\beta}) = \log\{L_n(\boldsymbol{\beta})\} = \sum_{i=1}^{n}\int_0^1 \mathbf{z}_i(t)'\boldsymbol{\beta}dN_i(s) - \int_0^1 \ln\left[\sum_{j=1}^{n} Y_j(s)\exp\{\mathbf{z}_j(t)'\boldsymbol{\beta}\}\right]d\bar{N}(s), \tag{2.4}$$

where $N_i(t) = I(X_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$ are the counting and at-risk process for subject i, respectively, and $\bar{N}(t) = \sum_{i=1}^{n} N_i(t)$. Without loss of generality, we work on the time interval $s \in [0, 1]$ as in Andersen and Gill (1982), which can be extended to the time interval $[0, \tau]$ for some $\tau \in (0, \infty)$ without difficulty.

The maximum partial likelihood estimator of $\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}_{mple}$, can be obtained by solving the following score equation

$$U_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^1 \left\{\mathbf{z}_i(t) - \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}\right\}dN_i(t) = 0, \tag{2.5}$$

where $S^{(0)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\exp\{\mathbf{z}_i(t)'\boldsymbol{\beta}\}$ and $S^{(1)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\mathbf{z}_i(t)\exp\{\mathbf{z}_i(t)'\boldsymbol{\beta}\}$. Andersen and Gill (1982) proved that the covariance matrix for $\hat{\boldsymbol{\beta}}_{mple}$ can be consistently estimated by the inverse of the observed information matrix $\hat{\Sigma}^{-1} = -\left\{\partial U_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{mple}}\right\}^{-1}$ and studied the large-sample properties of $\hat{\boldsymbol{\beta}}_{mple}$ under mild regularity conditions.

2.2 Penalized variable selection procedures for the Cox proportional hazards model

The concept of variable selection has been long used in the model building process to achieve a balance between parsimony and goodness of fit. This is especially important today when low costs and computational advancements allow us to collect and store large number of covariates that are potentially related to the outcome of interest. Classical techniques such as stepwise model building or best subset selection are known to be computationally intensive and unstable Breiman (1996) even for moderate dimensions and their theoretical properties remain unknown and underdeveloped. In recent years, penalized regression procedures have been introduced to perform variable selection in a continuous fashion. This is accomplished by minimizing a penalized objective function which consequently shrinks coefficient estimates toward zero or sets them exactly to zero. Tuning parameters typically control the amount of shrinkage imposed on the coefficients. Tibshirani (1996) popularized penalized regression through the development of the least absolute shrinkage and selection operator (LASSO) for ordinary least squares regression. Several well-established methods for linear models have been introduced since the LASSO (see e.g., Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Zhang (2010)) and have been extended to the Cox model. This rest of the section serves to acquaint readers to some popular approaches to penalized variable selection for the Cox model and the list should not be regarded as a comprehensive review.

We define the penalized negative log-partial likelihood for the Cox model (2.3) as

$$\mathrm{pl}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \tag{2.6}$$

where $l_n(\boldsymbol{\beta})$ is defined as in (2.4), and $p_\lambda(\cdot)$ is a penalty function with nonnegative tuning parameter λ_n. When $\lambda_n = 0$, the summation on the right is defined as zero and the usual negative log-partial likelihood is recovered. Estimating the parameters for a penalized Cox regression can be obtained through minimizing (2.6).

Tibshirani (1996) introduced LASSO regression for the linear model and quickly extended

it to the Cox model (Tibshirani, 1997). The LASSO (ℓ_1) penalty is defined as

$$p_\lambda(|\beta_j|) = \lambda_n|\beta_j|, \quad j = 1, \ldots, p_n, \tag{2.7}$$

and the LASSO estimator can be expressed as the minimizer of an ℓ_1-penalized negative log-partial likelihood function. Tibshirani (1996) further showed that the LASSO procedure shrinks all parameter estimates toward 0 and sets some estimates to exactly 0, depending on the choice of the tuning parameter λ_n. While LASSO allows for variable selection, it is also known to perform poorly with highly-correlated covariates and the estimate of $\boldsymbol{\beta}$ may suffer from substantial bias depending on the value of λ_n. The seminal works of Tibshirani (1996) and Tibshirani (1997) have propelled various extensions and improvements to LASSO.

Three such proposals are the smoothly clipped absolute deviation penalty (Fan and Li, 2001, 2002, SCAD), the minimax concave penalty (Zhang, 2010, MCP) and the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007). Both SCAD and MCP aim to address LASSO's significant bias toward 0 for large regression coefficients by initially applying the same rate of penalization as the LASSO but continuously relaxing the amount of penalization toward the unpenalized solution in their own respective manner. The adaptive LASSO is a direct modification of LASSO by allowing each coefficient to be penalized differently based on covariate-specific weights on the tuning parameter. An appealing property of SCAD, MCP, and adaptive LASSO is that they are *oracle* estimators (Fan and Li, 2001); that is, methods that asymptotically estimate the non-zero parameters as accurately and efficiently as if the underlying true model was known *a priori*.

We now focus on ℓ_0- and ℓ_2-penalizations, the key motivation for BAR regression. Best subset selection is a natural choice for variable selection by penalizing model complexity in a straightforward manner. The penalty function associated with best subset selection is the so-called ℓ_0 penalty,

$$p_\lambda(|\beta_j|) = \lambda_n I(\beta_j \neq 0), \quad j = 1, \ldots, p_n. \tag{2.8}$$

Although intuitive, *exact* ℓ_0-penalized regression has several limitations as explained in the introduction of the section. Finding the optimal model using ℓ_0-penalized regression requires the fitting of all possible models and then comparing the fitted models with some information criterion such as AIC (Akaike, 1974) or BIC (Schwarz, 1978; Volinsky and Raftery, 2000). For example, a model with $p_n = 15$ requires $2^{15} = 32768$ model fits. Adding one more covariate to the data will increase the number of candidate models by another 32768. While heuristic surrogates like stepwise selection are available, this combinatorial optimization problem is still infeasible for moderately large p_n and is unstable.

Ridge (ℓ_2-penalized) regression was first introduced to prevent degeneracy due to multicollinearity in ordinary least squares regression (Hoerl and Kennard, 1970) and has been extended to the Cox model (Verweij and Van Houwelingen, 1994). The corresponding penalty function in (2.6) is defined as

$$p_\lambda(|\beta_j|) = \lambda_n \beta_j^2, \quad j = 1, \ldots, p_n. \tag{2.9}$$

Ridge regression is known to have good prediction accuracy and is capable of grouping highly-correlated covariates. The convexity of the penalty also makes it easy to implement in software. On the other hand, ridge regression does not produce a sparse solution (i.e. every variable is preserved in the model) and parameter estimates are known to be downwardly biased.

Zou and Hastie (2005) proposed elastic net regression, which borrows strength from both LASSO (ℓ_1) and ridge (ℓ_2) regression and can be interpreted as a linear combination of the ℓ_1 and ℓ_2 penalties. By taking advantage of both penalties, the authors show that the elastic net penalty allows for sparse regression, a drawback of ℓ_2, while dealing with issues of collinearity, an ℓ_1 limitation. Wu (2012) extended the elastic net penalty to the Cox model and developed a solution path algorithm for it.

Most penalized variable selection methods require the careful selection of a tuning parameter. Data-driven methods to find the "optimal" tuning parameter are generally employed. Typically a grid search is implemented to identify the tuning parameter that minimizes some

criterion. Although cross validation (Craven and Wahba, 1978; Verweij and Van Houwelingen, 1993) has been a popular approach in selecting the tuning parameter, it has been shown to be selection inconsistent, usually resulting in an overfitted model with positive probability (Wang et al., 2007). Recently Ni and Cai (2018) extend the generalized information criterion (Zhang et al., 2010, GIC) to the Cox model. The authors further proved that a family of criteria, which include the BIC and GIC, can identify the true model with probability tending to one as the sample size goes to infinity under mild conditions.

Finally, the number of parameters, p_n, is generally categorized into three scenarios that reflect its relationship with the sample size n; 1) p_n is considered fixed as $n \to \infty$ (fixed finite dimension), 2) p_n is allowed to increase with n but at a slower rate (diverging dimension), and 3) p_n is assumed to increase exponentially with n (ultrahigh-dimension). This dissertation primarily concerns variable selection in the diverging dimension scenario.

2.3 Modeling the subdistribution hazard function for competing risks data

In biomedical studies with time-to-event data, individuals are oftentimes susceptible to more than one type of event (or cause) and the occurrence of one event oftentimes precludes the others from happening. Such events that are not of primary interest are considered as competing risks. In the USRDS example introduced in Chapter 1, researchers wish to examine how certain covariates affect time until first kidney transplant for kidney dialysis patients with end-stage renal disease. While subjects who are lost to follow up or dropout from the study are generally considered as right censored, they may also observe terminating events such death, renal function recovery, or discontinuation of dialysis. These events are considered to be competing risks as their occurrence will prevent subjects from receiving a transplant.

Before moving forward, we first establish some notation and the formal definition of the data generating process for competing risks. For subject $i = 1, \ldots, n$, let T_i, C_i, and ϵ_i be the event time, possible right-censoring time, and cause (event type), respectively. Without

loss of generality assume there are two event types $\epsilon \in \{1, 2\}$ where $\epsilon = 1$ is the event of interest (or primary event) and $\epsilon = 2$ is the competing risk. With the presence of right-censoring, we generally observe $X_i = T_i \wedge C_i$, $\delta_i = I(T_i \leq C_i)$, where $a \wedge b = \min(a, b)$ and $I(\cdot)$ is the indicator function. Letting $\mathbf{z}_i$ be a p-dimensional vector of time-independent subject-specific covariates, competing risks data consist of the following independent and identically distributed quadruplets $\{(X_i, \delta_i, \delta_i\epsilon_i, \mathbf{z}_i)\}_{i=1}^n$. Assume that there also exists a τ such that 1) for some arbitrary time t, $t \in [0, \tau]$; 2) $\Pr(T_i > \tau) > 0$ and $\Pr(C_i > \tau) > 0$ for all $i = 1, \ldots, n$, and that for simplicity, no ties are observed.

For competing risks data, the cumulative incidence function (CIF) is often of primary interest. The CIF for the primary event conditional on the covariates $\mathbf{z} = (z_1, \ldots, z_p)$ is $F_1(t; \mathbf{z}) = \Pr(T \leq t, \epsilon = 1|\mathbf{z})$. To model the covariate effects on $F_1(t; \mathbf{z})$, Fine and Gray (1999) introduced the now well-appreciated proportional subdistribution hazards (PSH) model:

$$h_1(t|\mathbf{z}) = h_{10}(t) \exp(\mathbf{z}'\boldsymbol{\beta}), \tag{2.10}$$

where

$$\begin{aligned} h_1(t|\mathbf{z}) &= \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T \leq t + \Delta t, \epsilon = 1 | T \geq t \cup (T \leq t \cap \epsilon \neq 1), \mathbf{z}\}}{\Delta t} \\ &= -\frac{d}{dt} \log\{1 - F_1(t; \mathbf{z})\} \end{aligned}$$

is a subdistribution hazard (Gray, 1988), $h_{10}(t)$ is a completely unspecified baseline subdistribution hazard, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. As Fine and Gray (1999) mentioned, the risk set associated with $h_1(t; \mathbf{z})$ is somewhat counterfactual as it includes subjects who are still at risk ($T \geq t$) and those who have already observed the competing risk prior to time t ($T \leq t \cap \epsilon \neq 1$). However, this construction is useful for direct modeling of the CIF.

Parameter estimation and large-sample inference of the PSH model follows from the

log-pseudo likelihood:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \left[\mathbf{z}_i'\boldsymbol{\beta} - \ln \left\{ \sum_k \hat{w}_k(u) Y_k(u) \exp\left(\mathbf{z}_k'\boldsymbol{\beta}\right) \right\} \right] \hat{w}_i(u) dN_i(u), \tag{2.11}$$

where $N_i(t) = I(X_i \leq t, \epsilon_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, and $\hat{w}_i(t)$ is a time-dependent weight based on the inverse probability of censoring weighting (IPCW) technique (Robins and Rotnitzky, 1992). To parallel Fine and Gray (1999), we define the IPCW for subject i at time t as $\hat{w}_i(t) = I(C_i \geq T_i \wedge t)\hat{G}(t)/\hat{G}(X_i \wedge t)$, where $G(t) = \Pr(C \geq t)$ is the survival function of the censoring variable C and $\hat{G}(t)$ is the Kaplan-Meier (Kaplan and Meier, 1958) estimate for $G(t)$. However, we can generalize the IPCW to allow for dependence between C and $\mathbf{z}$.

Let $\hat{\boldsymbol{\beta}}_{mple} = \arg\min_{\boldsymbol{\beta}}\{-l(\boldsymbol{\beta})\}$ be the maximum pseudo likelihood estimator of $\boldsymbol{\beta}$. Fine and Gray (1999) prove that, under certain regularity conditions, $\hat{\boldsymbol{\beta}}_{mple}$ is a consistent estimator for $\boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$, and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{mple} - \boldsymbol{\beta}_0) \to N(0, \boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{-1}), \tag{2.12}$$

where $\boldsymbol{\Omega}$ is the limit of the negative of the partial derivative matrix of the score function evaluated at $\boldsymbol{\beta}_0$, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the limiting distribution of the score function. We refer readers to Section 4 and Appendix A of Fine and Gray (1999) for a more comprehensive derivation of the large-sample properties of $\hat{\boldsymbol{\beta}}_{mple}$ which we have omitted for brevity.

While parameter estimation for the Fine-Gray model is relatively straightforward, the interpretation of the regression coefficients is not without difficulty. For example, the magnitude of the relative effect of the covariate on the subdistribution hazard function (i.e. the subdistribution hazard ratio) is different from the magnitude of the effect of the covariate on the CIF (Austin and Fine, 2017). We can, however, describe the direction of association (e.g. If the subdistribution hazard ratio is greater than 1, then incidence of the event will also increase). In testing statistical significance of the subdistribution hazard ratio, we are also performing a test for the covariate effect on the CIF.

An alternative use of the regression coefficients is to predict the CIF given a set of covariates. Using a Breslow-type estimator (Breslow, 1974), we can obtain a consistent estimate for $H_{10}(t) = \int_0^t h_{10}(s)ds$ through

$$\hat{H}_{10}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{1}{\hat{S}^{(0)}(\hat{\boldsymbol{\beta}}, u)}\hat{w}_i(u)dN_i(u),$$

where $\hat{S}^{(0)}(\hat{\boldsymbol{\beta}}, u) = n^{-1}\sum_{i=1}^{n}\hat{w}_i(u)Y_i(u)\exp(\mathbf{z}_i'\hat{\boldsymbol{\beta}})$. The predicted CIF, conditional on $\mathbf{z} = \mathbf{z}_0$, is then

$$\hat{F}_1(t; \mathbf{z}_0) = 1 - \exp\left\{\int_0^t \exp(\mathbf{z}_0'\hat{\boldsymbol{\beta}})d\hat{H}_{10}(u)\right\}.$$

The quantities needed to estimate $\int_0^t d\hat{H}_{10}(u)$ are already precomputed when estimating $\hat{\boldsymbol{\beta}}$. Fine and Gray (1999) proposed a resampling approach to calculate confidence intervals and confidence bands for $\hat{F}_1(t; \mathbf{z}_0)$.

Variable selection follows directly from Section 2.2 for the Cox proportional hazards model where the log-partial likelihood is replaced with the log-pseudo likelihood (2.11) in (2.6). Recently, Fu et al. (2017) extended LASSO, SCAD, MCP, and adaptive LASSO to the Fine-Gray model and established their large-sample properties for the fixed dimension scenario.

CHAPTER 3

Broken adaptive ridge for the Cox proportional hazards model with applications to sparse high-dimensional massive sample size (sHDMSS) data

This chapter develops the broken adaptive ridge estimator for the Cox proportional hazards model for right-censored time-to-event data with applications to sHDMSS data. In Section 3.1, we formally define the BAR estimator, state its theoretical properties for variable selection and parameter estimation and describe an efficient implementation of BAR for sHDMSS time-to-event data. Simulation studies are presented in Section 3.2 to demonstrate the performance of the BAR estimator with both moderate and massive sample size in various low and high-dimensional settings. A real data example including an application of BAR on the pediatric trauma mortality data (Mittal et al., 2014) is given in Section 3.3. Closing remarks and discussion are given in Section 3.4. Proofs of the theoretical results, regularity conditions needed for the derivations, and supplementary material are collected in the appendix. An R (R Core Development Team, 2019) package for BAR is available at `https:github.com/OHDSI/BrokenAdaptiveRidge`.

3.1 Methodology

3.1.1 Cox's broken adaptive ridge regression and its large sample properties

3.1.1.1 The data structure, model, and estimator

Suppose that one observes a random sample of right-censored time-to-event data consisting of n independent and identically distributed triplets, $\{(X_i, \delta_i, \mathbf{z}_i(\cdot))\}_{i=1}^n$, where for subject i, $X_i = \min(T_i, C_i)$ is the observed event time, $\delta_i = I(T_i \leq C_i)$ is the censoring indicator, T_i is the event time of interest, and C_i is a censoring time that is conditionally independent of T_i given a p_n-dimensional, possibly time-dependent, covariate vector $\mathbf{z}_i(\cdot) = (z_{i1}(\cdot), \ldots, z_{ip_n}(\cdot))'$.

Assume the Cox (1972) proportional hazard model

$$h\{t|\mathbf{z}(t)\} = h_0(t)\exp\{\mathbf{z}(t)'\boldsymbol{\beta}\}, \tag{3.1}$$

where $h\{t|\mathbf{z}(t)\}$ is the conditional hazard function of T_i given $\{\mathbf{z}(u),\ 0 \leq u \leq t,\}$, $h_0(t)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})$ is a vector of regression coefficients. Denote by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ the first q_n and remaining $p_n - q_n$ components of $\boldsymbol{\beta}$, respectively, and define $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ as the true values of $\boldsymbol{\beta}$ where, without loss of generality, $\boldsymbol{\beta}_{01} = (\beta_{01} \ldots, \beta_{0q_n})$ is a vector of q_n non-zero values and $\boldsymbol{\beta}_{02} = \mathbf{0}$ is a $p_n - q_n$ dimensional vector of zeros. Further technical assumptions for $\boldsymbol{\beta}_0$ and p_n are given later in condition (C6) of Appendix A3.1. We work on the time interval $s \in [0, 1]$ as in Andersen and Gill (1982), which can be extended to the time interval $[0, \tau]$ for $0 < \tau < \infty$ without difficulty. Andersen and Gill (1982) defined the log-partial likelihood for the Cox model

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^1 \boldsymbol{\beta}'\mathbf{z}_i(s)dN_i(s) - \int_0^1 \ln\left[\sum_{j=1}^n Y_j(s)\exp\{\boldsymbol{\beta}'\mathbf{z}_j(s)\}\right] d\bar{N}(s), \tag{3.2}$$

where for subject i, $Y_i(s) = I(X_i \geq s)$ is the at-risk process and $N_i(s) = I(X_i \leq s, \delta_i = 1)$ is the counting process of the uncensored event with intensity process $h_i(t|\boldsymbol{\beta}) = h_0(t)Y_i(t)\exp\{\mathbf{z}_i(t)'\boldsymbol{\beta}\}$ and $\bar{N} = \sum_{i=1}^n N_i$. Let $H_i(t) = \int_0^1 h_i(u, \boldsymbol{\beta}_0)du$, then $M_i(t) = N_i(t) - H_i(t)$ is a local square integrable martingale with respect to filtration

$\mathcal{F}_{t,i} = \sigma\{N_i(u), \mathbf{z}_i(u^+), Y_i(u^+), 0 \leq u \leq t\}$, and $\bar{M}(t) = \sum_{i=1}^n M_i(t)$ is a martingale with respect to $\mathcal{F}_t = \cup_{i=1}^n \mathcal{F}_{t,i}$, the smallest σ-algebra containing all $\mathcal{F}_{t,i}$'s.

Our Cox's broken adaptive ridge (BAR) estimation of $\boldsymbol{\beta}$ starts with an initial Cox ridge regression estimator (Verweij and Van Houwelingen, 1994)

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p_n} \beta_j^2 \right\}, \tag{3.3}$$

which is updated iteratively by a reweighed ℓ_2-penalized Cox regression estimator

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\left(\hat{\beta}_j^{(k-1)}\right)^2} \right\}, \quad k \geq 1. \tag{3.4}$$

where ξ_n and λ_n are non-negative penalization tuning parameters. The BAR estimator is defined as

$$\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}. \tag{3.5}$$

Since ℓ_2-penalization yields a non-sparse solution, defining the BAR estimator as the limit is necessary to produce sparsity. Although λ_n is fixed at each iteration, it is weighted inversely by the square of the ridge regression estimates from the previous iteration. Consequently, coefficients whose true values are zero will have larger penalties in the next iteration, whereas penalties for truly non-zero coefficients will converge to a constant. We will show later in Theorem 3.1 that under certain regularity conditions, the estimates of the truly zero coefficients shrink towards zero while the estimates of the truly non-zero coefficients converge to their oracle estimates.

Remark 3.1 *(Computational aspects of BAR) For moderate size data, one may calculate $\hat{\boldsymbol{\beta}}^{(k)}$ in (3.4) using the Newton-Raphson method as in Frommlet and Nuel (2016) who outlined an iterative reweighted ridge regression for generalized linear models. It appears at the first sight that (3.4) will encounter numerical overflow as some of the coefficients $\hat{\beta}_j^{(k-1)}$ will go to zero as k increases. However, it can be shown that after some simple algebraic manipulation, the Newton-Raphson updating formula will only involve multiplications, instead of*

divisions, by $\hat{\beta}_j^{(k-1)}$s and numerical overflow can be avoided. This further implies that once a $\hat{\beta}_j^{(k-1)}$ becomes zero, it will remain as zero in subsequent iterations. Thus one only needs to update $\hat{\boldsymbol{\beta}}^{(k)}$ within the reduced nonzero parameter space, which is an appealing computational advantage for high-dimensional settings. For massive size data with large n and p_n, the Newton-Raphson procedure, which at each iteration calls for calculating both the gradient and Hessian, can become practically infeasible due to high computational costs, memory requirements, and numerical instability. In Section 3.1.2 we will discuss how to adapt an efficient algorithm for massive ℓ_2-penalized Cox regression via cyclic coordinate descent and exploit the sparsity in the covariate structure to make BAR scalable to sHDMSS data.

3.1.1.2 Oracle property

We establish the oracle properties for the BAR estimator for simultaneous variable selection and parameter estimation where we allow both q_n and p_n to diverge to infinity.

Theorem 3.1 (Oracle property) *Assume the regularity conditions (C1) - (C6) in Appendix A3.1 hold. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the first q_n and the remaining $p_n - q_n$ components of the BAR estimator $\hat{\boldsymbol{\beta}}$, respectively. Then, as $n \to \infty$,*

(a) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to one;

(b) $\sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0,1)$, for any q_n-dimensional vector $\mathbf{b}_n$ such that $||\mathbf{b}_n||_2 \leq 1$ and where $\Sigma(\boldsymbol{\beta}_0)_{11}$ is the first $q_n \times q_n$ submatrix of $\Sigma(\boldsymbol{\beta}_0)$, where $\Sigma(\boldsymbol{\beta}_0)$ is defined in Condition (C4).

Theorem 3.1(a) establishes selection consistency of the BAR estimator. Part (b) of the theorem essentially states that the nonzero component of the BAR estimator is asymptotically normal and equivalent to the weighted ridge estimator of the oracle model as shown in the proof provided in Appendix A3.2.

3.1.1.3 The grouping property

When the true model has a group structure, it is desirable for a variable selection method to either retain or drop all variables that are clustered within the same group. Ridge regression has a grouping property, and it is intuitive to conjecture that the BAR method would as well since the estimator is based on an iterative ridge regression. The following theorem states the grouping property of the BAR estimator for highly-correlated covariates.

Theorem 3.2 *Let λ_n, $\{(X_i, \delta_i, \mathbf{z}_i)\}_{i=1}^n$ be given and assume that $Z = (\mathbf{z}_i', \ldots \mathbf{z}_n')$ is standardized. That is, for all $j = 1, \ldots, p_n$, $\sum_{i=1}^n z_{ij} = 0$, $\mathbf{z}_{[,j]}'\mathbf{z}_{[,j]} = n-1$, where $\mathbf{z}_{[,j]}$ is the jth column of Z. Suppose the regularity conditions (C1) - (C6) in Appendix A3.1 hold and let $\hat{\boldsymbol{\beta}}$ be the BAR estimator. Then for any $\hat{\beta}_i \neq 0$ and $\hat{\beta}_j \neq 0$,*

$$|\hat{\beta}_i^{-1} - \hat{\beta}_j^{-1}| \leq \frac{1}{\lambda_n}\sqrt{2\{(n-1)(1-r_{ij})\}}\sqrt{n(1+d_n)^2}, \tag{3.6}$$

with probability tending to one, where $d_n = \sum_{i=1}^n \delta_i$, and $r_{ij} = \frac{1}{n-1}\mathbf{z}_{[,i]}'\mathbf{z}_{[,j]}$ is the sample correlation of $\mathbf{z}_{[,i]}$ and $\mathbf{z}_{[,j]}$.

We can see that as $r_{ij} \to 1$, the absolute difference between $\hat{\beta}_i$ and $\hat{\beta}_j$ approaches 0 implying that the estimated coefficients of two highly correlated variables will be similar in magnitude. The proof is provided in Appendix A3.3.

3.1.1.4 Selection of tuning parameters

Model complexity depends critically on the choice of the tuning parameters. The BAR estimator depends on two tuning parameters: ξ_n for the initial ridge estimator in (3.3) and λ_n for the iterative ridge step in (3.4). Our simulations in Section 3.2.1 illustrate that while fixing λ_n, the BAR estimator is insensitive to the choice of ξ_n over a wide interval (Figure 3.1).

We optimize with respect to λ_n in a similar manner to currently-used penalization methods. A popular strategy for tuning parameter selection is to perform optimization with re-

spect to a data-driven selection criterion such as cross-validation (Craven and Wahba, 1978; Verweij and Van Houwelingen, 1993), Akaike information criterion (AIC) (Akaike, 1974), and Bayesian information criterion (BIC) (Schwarz, 1978; Volinsky and Raftery, 2000; Ni and Cai, 2018). Although cross validation has been used extensively in the literature, it has been known to asymptotically overfit models with a positive probability (Wang et al., 2007; Zhang et al., 2010). Recent theoretical work has shown that for penalized Cox models that possess the oracle property, BIC-based tuning parameter selection identifies the true model with probability tending to one (Ni and Cai, 2018).

3.1.2 Efficient implementation BAR for sparse high-dimensional massive sample size (sHDMSS) data

As mentioned in Remark 3.1, the Newton-Raphson algorithm used for each iteration of the BAR algorithm will become infeasible in large-scale settings with large n and p_n due to high computational costs, high memory requirements, and numerical instability. Because BAR only involves fitting a reweighted Cox's ridge regression at each iteration step, it allows us to adapt an efficient algorithm developed by Mittal et al. (2014) for massive Cox ridge regression.

Mittal et al. (2014) developed an efficient implementation of the massive Cox's ridge regression for sHDMSS data. For parameter estimation, the authors adopted the column relaxation with logistic loss (CLG) algorithm of Zhang and Oles (2001), which is a type of cyclic coordinate descent algorithm that estimates the coefficients using one-dimensional updates. The CLG easily scales to high-dimensional data (Wu and Lange, 2008; Simon et al., 2011; Gorst-Rasmussen and Scheike, 2012) and has been recently implemented for fitting ℓ_2- and ℓ_1-penalized generalized linear models (Suchard et al., 2013), parametric time-to-event models (Mittal et al., 2013), and Cox's model (Mittal et al., 2014). Readers are encouraged to refer Section A3.4 of the Appendix for a detailed explanation of the algorithm.

The design matrix Z for sHDMSS data has few non-zero entries for each subject. Storing such a sparse matrix as a dense matrix is inefficient and may increase computation time

and/or cause standard software to crash due to insufficient memory allocation. To the best of our knowledge, popular penalization packages such as **glmnet** (Friedman et al., 2010) and **ncvreg** (Breheny and Huang, 2011) do not support a sparse data format as an input for right-censored time-to-event models, although the former supports the input for other generalized linear models. For sHDMSS data, we propose to use specialized, column-data structures as in Suchard et al. (2013) and Mittal et al. (2014). The advantage of this structure is two-fold: it significantly reduces the memory requirement needed to store the covariate information, and performance is enhanced when employing cyclic coordinate descent. For example when updating β_j, efficiency is gained when computing and storing the inner product $r_i = \mathbf{z}_i'\boldsymbol{\beta}$ using a low-rank update $r_i^{(new)} = r_i + z_{ij} + \Delta\beta_j$ for all i (Zhang and Oles, 2001; Genkin et al., 2007; Wu and Lange, 2008; Suchard et al., 2013; Mittal et al., 2014).

Furthermore, to calculate the gradient and Hessian diagonal, one requires a series of cumulative sums introduced through the risk set $R_i = \{j : X_j > X_i\}$ for each subject i. These cumulative sums would need to be calculated when updating each parameter estimate in the optimization routine. This can prove to be computationally costly, especially when both n and p_n are large. By taking advantage of the sparsity of the design matrix, one can reduce the computational time needed to calculate these cumulative sums by entering into this operation only if at least one observation in the risk set has a non-zero covariate value along dimension j and embarking on the scan at the first non-zero entry rather than from the beginning. Suchard et al. (2013) and Mittal et al. (2014) have implemented these efficiency techniques for conditional Poisson regression and Cox's regression, respectively. Our BAR implementation naturally exploits the sparsity in the design matrix and the partial likelihood by imbedding an adaptive version of Mittal et al. (2014)'s massive Cox's ridge regression within each iteration of the iteratively reweighted Cox's ridge regression.

Remark 3.2 (Ultrahigh-dimensional time-to-event data) *The asymptotic properties of the BAR estimator in the Section 3.1.1.2 are derived for $p_n < n$. In an ultrahigh dimensional setting where the number of covariates far exceeds the number of observations ($p_n >> n$), one may couple a sure screening (Fan et al., 2010) method with the BAR estimator to obtain a two-step estimator with desirable selection and estimation properties.*

There are a number of screening methods for right-censored time-to-event data, which include marginal screening methods (Fan et al., 2010; Zhao and Li, 2012; Gorst-Rasmussen and Scheike, 2013; Song et al., 2014) and joint screening methods (Yang et al., 2016). For example, the sure joint screening (SJS) method of Yang et al. (2016) is based on the joint partial likelihood of potentially important covariates using a sparsity-restricted maximum partial likelihood estimate. As an illustration, we consider a two-step estimator, referred to as SJS-BAR, obtained by first performing SJS to reduce the covariate space to a subset $\hat{s}$ of m_n covariates and then fit BAR to the screened model $\hat{s}$. Additional regularity conditions, the conditional oracle property, and a proof are provided in Appendix A3.3.1.

3.2 Simulations

This section presents three simulation studies. First, we demonstrate in Section 3.2.1 that for fixed λ_n, the BAR estimator is insensitive to the tuning parameter ξ_n of its initial ridge estimator and does well in terms of performing variable selection and correcting possible bias of the initial ridge estimator. Then in Section 3.2.2, we evaluate and compare the operating characteristics of BAR with some popular penalized Cox regression methods, where we only consider settings with moderate sample sizes because most of the competing methods are inoperable for massive sample size data. Finally in Section 3.2.3, we use a sHDMSS setting to illustrate the performance of BAR over its closest competitor.

Sections 3.2.1 and 3.2.2 employ the same simulation structure. Event times are drawn from an exponential proportional hazards model with baseline hazard $h_0(t) = 1$ and $\boldsymbol{\beta}_0 = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80, \mathbf{0}_{p_n-10})$, representing small to moderate effect sizes, the design matrix $Z = (\mathbf{z}_1', \ldots, \mathbf{z}_n')$ is generated from a p_n-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$ with an autoregressive structure such that $\sigma_{ij} = 0.5^{|i-j|}$ and independent censoring times are generated from uniform distribution $U(0, u_{\max})$, where $u_{\max}$ is chosen to achieve different percentages of censoring. We describe how we simulate sHDMSS time-to-event data in Section 3.2.3.

3.2.1 BAR estimator for varying values of ξ_n

We illustrate below how the BAR estimator behaves by fixing λ_n and varying the tuning parameter ξ_n of the initial Cox ridge regression. Figure 3.1 (panels (b), (c) and (d)) depicts the solution path plots average over 100 Monte Carlo simulations of the BAR estimator with respect to ξ_n over a wide interval $[10^{-2}, 10^2]$ for $n = 300$, $p_n = 100$, $\approx 25\%$ censoring, and $\lambda_n = \log(p_n), 0.5\log(p_n), 0.75\log(p_n)$, respectively. The resulting BAR estimator is essentially unchanged, regardless of the choice of λ_n, over a large interval of ξ_n, suggesting that the BAR estimator is relatively insensitive to original ridge estimator.

As a reference, we also display the solution path plots of the corresponding initial ridge estimator in panel (a). The initial ridge estimator starts to introduce over shrinkage and, consequently, estimation bias when ξ_n exceeds 10^1. However, its bias has been effectively corrected by BAR. Therefore, by iteratively refitting reweighted Cox ridge regression, the BAR estimator not only performs variable selection by shrinking estimates of the true zero parameters to zero, but also effectively corrects the estimation bias from the initial Cox ridge estimator. Similar results are obtained for several different simulation scenarios and can be found in Appendix A3.5.

3.2.2 Model selection and parameter estimation

In this simulation, we evaluate and compare the variable selection and parameter estimation performance of BAR with four popular penalized Cox regression methods: LASSO (Tibshirani, 1997), SCAD (Fan and Li, 2002) , adaptive LASSO (ALASSO) (Zhang and Lu, 2007), and MCP (Zhang, 2010). We fix $\xi_n = 1$ for the BAR methods since Section 3.2.1 yields evidence that the BAR estimator is insensitive to the selection of ξ_n. BIC-score minimization is used to select the optimal tuning parameter for all five penalization methods.

Estimation bias is summarized through the mean squared bias (MSB), $E(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2)$. Variable selection performance is measured by a number of indices: the mean number of false positives (FP), the mean number of false negatives (FN); and average similarity measure (SM) for support recovery where $SM = ||\hat{\mathcal{S}} \cap \mathcal{S}_0||_0 / \sqrt{||\hat{\mathcal{S}}||_0 \cdot ||\mathcal{S}_0||_0}$ and $\mathcal{S}_0$ and $\hat{\mathcal{S}}$ are the set

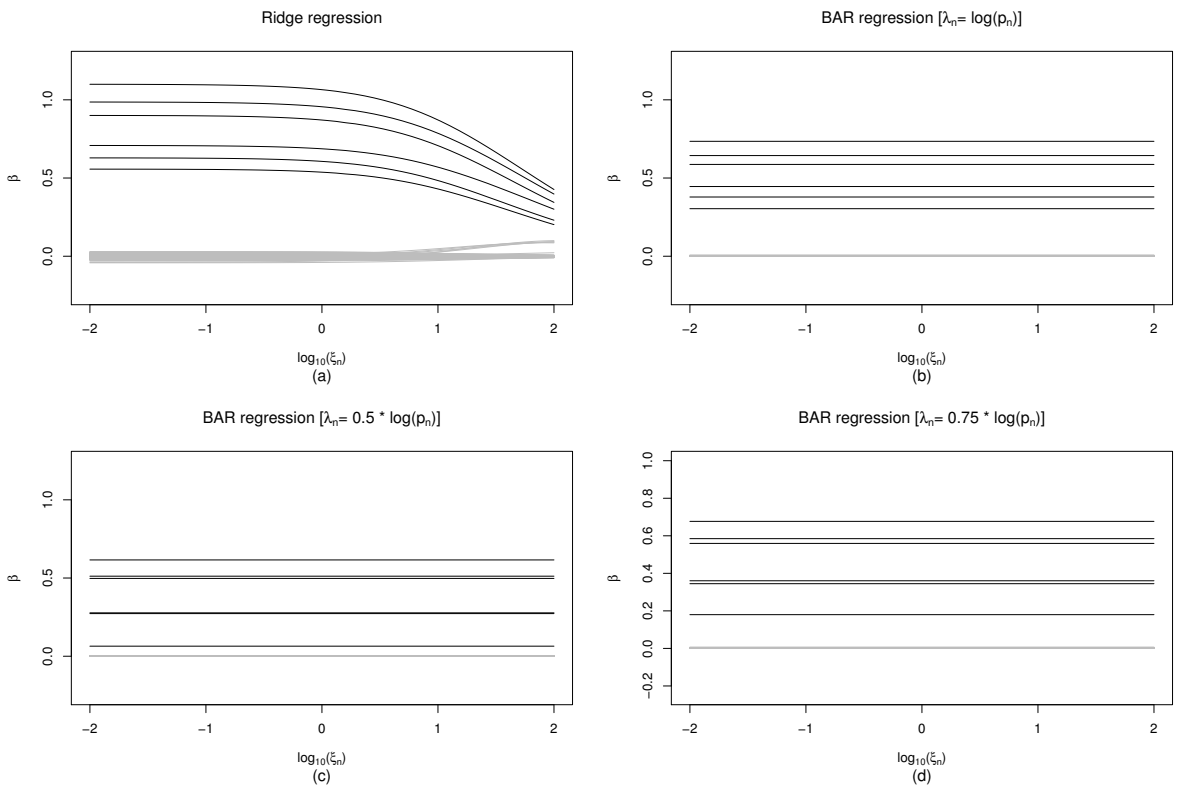

Figure 3.1: Path plot for BAR regression with varying ξ_n and: (b) $\lambda_n = \log(p_n)$, (c) $\lambda_n = 0.5\log(p_n)$, and (d) $\lambda_n = 0.75\log(p_n)$ with estimates averaged over 100 Monte Carlo simulations of size $n = 300$, $p_n = 100$, and censoring rate $\approx 25\%$. Path plot for ridge regression (d) with varying ξ_n is also included as a comparison.

of indices for the non-zero components of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, respectively (Zhang and Cheng, 2017). The similarity measure can be viewed as a continuous measure for true model recovery: it is close to 1 when the estimated model is similar to the true model and close to 0 when the estimated model is highly dissimilar to the true model. We use the R package **ncvreg** to perform LASSO, adaptive LASSO (ALASSO), SCAD, and MCP penalizations in our simulations. For ALASSO, we let the initial weight be the maximum partial likelihood estimator since $p_n < n$. Partial simulation results are summarized in Table 3.1 where we fix $n = 300, 1000$, $p_n = 100$, a censoring rate of $\approx 25\%$, and average results over 100 replications.

It is observed from Table 3.1 that when the tuning parameter λ is selected by minimizing the BIC score as the other methods, the performance of BAR is generally comparable to other methods with respect to all measures across all scenarios. We have conducted more extensive

Table 3.1: (Moderate dimension and sample size) Simulated estimation and variable selection performance of BAR, LASSO, SCAD, ALASSO, and MCP where the BIC criterion was used to select the tuning parameters via a grid search. (MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = similarity measure; BIC = average BIC score; Each entry is based on 100 Monte Carlo samples of size $n = 300$ (top), and 1000 (bottom), $p_n = 100$, censoring rate $\approx 25\%$.)

	MSB	FN	FP	SM	BIC
BAR	0.11	0.01	1.79	0.89	1919.26
LASSO	0.27	0.01	3.32	0.82	1958.40
SCAD	0.12	0.01	2.23	0.87	1933.43
ALASSO	0.11	0.04	1.48	0.90	1935.60
MCP	0.09	0.02	1.21	0.92	1929.33
BAR	0.02	0.00	0.73	0.95	8196.51
LASSO	0.10	0.00	2.77	0.84	8236.76
SCAD	0.01	0.00	0.23	0.98	8203.00
ALASSO	0.02	0.00	0.26	0.98	8204.58
MCP	0.01	0.00	0.08	0.99	8202.04

simulations with different combinations of model dimension, censoring rates, sample sizes, and model sparsity, which yielded consistent findings and are reported in Appendix A3.6

3.2.3 Sparse high-dimensional massive sample size data

In this simulation, we simulate a sHDMSS time-to-event dataset with $n = 200,000$ and $p_n = 20,000$. Event times are generated from an exponential hazards model with baseline hazard $h_0(t) = 1$, regression coefficients $\boldsymbol{\beta}_0 = (\mathbf{0.7}_{10}, \mathbf{0.5}_{10}, \mathbf{1}_{10}, -\mathbf{0.7}_{10}, -\mathbf{0.5}_{10}, -\mathbf{1}_{10}, \mathbf{0}_{p_n-60})$, and a censoring rate of 95%. The covariates for each subject are simulated such, on average, 2% are assigned a non-zero value. The amount of memory used to store this dense design matrix would require over `16GB`, which exceeds the functional capacity of most statistical software packages on standard hardware. To overcome this difficulty, we efficiently store the information in a coordinate list fashion and compare our BAR method with the massive sparse Cox's regression for LASSO (mCox-LASSO) using the **Cyclops** package (Suchard et al., 2013; Mittal et al., 2014) which, to the best of our knowledge, is the fastest software available today that exploits the sparsity of sHDMSS time-to-event data for efficient computing and offers > 10-fold speedup (Mittal et al., 2014) over its competitors such as

Table 3.2: (Sparse high dimensional and massive sample size) Estimation and variable selection results for BAR and massive Cox regression with LASSO penalty (mCox-LASSO, Mittal et al. (2014)) for a simulated sHDMSS dataset with $n = 200,000$, $p_n = 20,000$, and $q_n = 60$. (Bias = $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2$; FP= number of false positives; FN = number of false negatives.)

Method	Bias	FP	FN	BIC score
BAR (BIC)	0.82	5	0	226200.5
mCox-LASSO (BIC)	2.49	5	0	227059.5
mCox-LASSO (CV)	2.02	120	0	227955.3

CoxNet (Simon et al., 2011) and **FastCox** (Yang and Zou, 2012). For LASSO, cross validation (mCox-LASSO (CV)), combined with a nonconvex optimization technique which is more efficient than the classical grid search approach, and BIC score minimization (mCox-LASSO (BIC)), implemented with the classical grid search approach, were used to find the optimal value for the tuning parameter. For the BAR method, we also implement BIC score minimization using a classical grid search. We report the bias ($||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2$), number of false positives (FP), false negatives (FN), and BIC score ($-2l_n(\hat{\boldsymbol{\beta}}) + \log(n) \sum_j I(\hat{\boldsymbol{\beta}}_j \neq 0)$) in Table 3.2.

All three methods retain the 60 true nonzero coefficients; however, mCox-LASSO using cross validation selects a large number of noise variables (120) compared to BAR and mCox-LASSO using BIC minimization (5). In addition, of the five noise variables selected by both BAR (BIC) and mCox-LASSO (BIC), four of them are overlapping. In terms of parameter estimation, BAR is less biased (0.82) than mCox-LASSO (2.49 for BIC and 2.02 for CV). For model fit, BAR has a much lower BIC score when compared to the mCox-LASSO methods. In summary, this simulation illustrates that BAR produces a sparse model with less bias and better model fit compared to mCox-LASSO.

We further examined the solution paths of mCox-LASSO and BAR in Figure 3.2. The vertical solid and dashed lines in the mCox-LASSO solution path plot (Figure 3.2(a)) represent the estimates at the optimal tuning parameter obtained via cross validation and BIC minimization, respectively. We can see that the mCox-LASSO solution path changes rapidly as its tuning parameter varies. In contrast, the BAR solution path plot (Figure 3.2(b)) with respect to λ_n changes very slowly over a relatively where the vertical line represents the es-

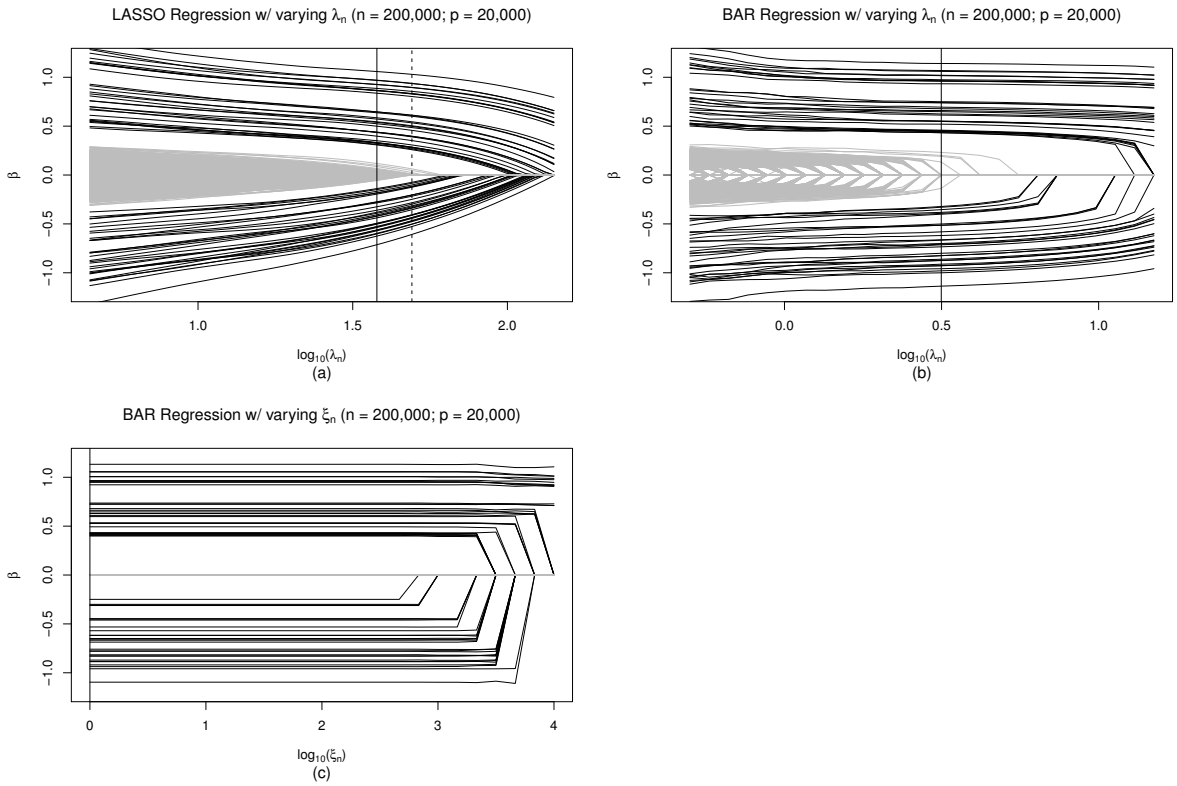

Figure 3.2: Path plots for mCox-LASSO and BAR regression: (a) Path plot for mCox-LASSO regression, where the black dashed line represents the estimates when using cross validation to find the optimal value of the tuning parameter; (b) Path plot for BAR regression with $\xi_n = 1$ and varying λ_n, where the black solid and dashed line represent estimates for $\lambda_n = \ln(n)$ and $\lambda_n = \ln(d_n)$, respectively; (c) Path plot for BAR regression with $\lambda_n = \ln(n)$ and varying ξ_n, where the black solid line represent the estimates for BAR when $\xi_n = 1$.

timates at the optimal tuning parameter selected by BIC minimization and selects a model that estimates with are less biased than LASSO (see Table 3.2). For the BAR method, we also made a BAR solution path plot with respect to ξ_n, while fixing $\lambda_n = \ln(n)$ in Figure 3.2(c). It shows that the BAR estimates are very stable over a large range of ξ_n, affirming our observation in Section 3.2.1 with small-scale data that BAR is generally insensitive ξ_n.

3.3 Pediatraic trauma mortality

For an application of BAR regression in the sHDMSS setting, we consider a subset of the National Trauma Data Bank that involves children and adolescents. This dataset was previ-

Table 3.3: (Pediatric NTDB data) Comparison of mCox-LASSO and BAR regression for the pediatric NTDB data. (mCox-LASSO (CV) and mCox-LASSO (BIC) correspond to mCox-LASSO using cross validation and BIC selection criterion, respectively. BAR (BIC) denotes BAR using the BIC selection criterion while fixing $\xi_n = \log(p_n)$. The training set has a sample size of $168,000$ while the test set used for the c-index has a sample size of $45,555$.)

Method	# Selected	BIC score	c-index
BAR (BIC)	83	51269.43	0.93
mCox-LASSO (BIC)	100	52544.90	0.91
mCox-LASSO (CV)	253	53165.44	0.92

ously analyzed by Mittal et al. (2014) as an example for efficient massive Cox regression with mCox-LASSO and ridge regression to sHDMSS data. The dataset includes 210,555 patient records of injured children under 15 that were collected over 5 years (2006 - 2010). Each patient record includes 125,952 binary covariates which indicate the presence, or absence, of an attribute (ICD9 Codes, AIS codes, etc.) as well as the two-way interactions. The outcome of interest is mortality after time of injury. The data is extremely sparse, with less than 1% of the covariates being non-zero and has a censoring rate of 98%. We randomly split the data into training and test sets of 168,000 and 42,555, respectively. The mortality rate of both sets were approximately equal to the combined rate. Similar to Section 3.2.3, we were unable to load the training set ($n = 168,000$, $p_n = 125,000$) into other popular oracle procedures due to the memory requirements needed to support a dense design matrix of that size and compare BAR to mCox-LASSO. BIC-score minimization over a coarse penalization path of 10 tuning parameters was used to select the final model for both BAR (fixing $\xi_n = \log(p_n)$) and mCox-LASSO. In addition, we perform mCox-LASSO using cross validation. The BIC score based on the training data is used to compare selection performance between models and discriminatory performance is measured using Harrell's c-statistic (Harrell et al., 1982, 1996) based on the test data.

Table 3.3 summarizes the findings for our example. BAR, using BIC minimization, selects fewer covariates than both mCox-LASSO methods. Both model selection and discriminatory performance are similar to slightly superior for BAR over mCox-LASSO.

3.4 Discussion

Although many penalized Cox regression methods for simultaneous variable selection and parameter estimation are available, most current algorithms and software will grind to a halt and become inoperable for sHDMSS data. We develop a new sparse Cox regression method by iteratively performing reweighted ℓ_2-penalized Cox regression where the penalty is adaptively reweighted to approximate the ℓ_0 penalty. The resulting estimator can be viewed as a special local ℓ_0-penalized Cox regression method and is shown to enjoy properties of both ℓ_0- and ℓ_2-penalized Cox regression: it is selection consistent, oracle for parameter estimation, computationally stable, and has a grouping property for highly-correlated covariates. We illustrate through empirical studies that the BAR estimator has comparable or better performance for variable selection and parameter estimation as compared to current penalized Cox regression methods and, most importantly, can directly fit sHDMSS time-to-event data. Its scalability to sHDMSS data is primarily due to the fact that the BAR algorithm allows us to easily adapt existing algorithms and software for massive ℓ_2-penalized Cox regression (Mittal et al., 2014).

It is also worth noting that our ℓ_0-based BAR method and theory can be easily extended to an ℓ_d-based BAR method for any $d \in [0, 1]$, by replacing $(\hat{\beta}_j^{(k-1)})^2$ with $|\hat{\beta}_j^{(k-1)}|^{2-d}$ in (3.4). We have observed empirically that as d increases towards 1, the resulting estimator becomes less sparse, and the average number of false positives as well as estimation bias tend to increase, especially for larger p_n, while the average number of false negatives tends to decrease. In practice, d can be used as a resolution tuning parameter. Finally, the proposed BAR method can extended to obtain scalable sparse regression methods for more complex sampling schemes such as cohort sampling, which is currently under investigation.

Appendix to Chapter 3

A3.1 Regularity conditions for Theorem 3.1

Define

$$
\begin{aligned}
S^{(k)}(\boldsymbol{\beta}, s) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(s)\mathbf{z}_i(s)^{\otimes k}\exp\{\boldsymbol{\beta}'\mathbf{z}_i(s)\}, \quad k = 0, 1, 2, \\
\mathbf{E}(\boldsymbol{\beta}, s) &= S^{(1)}(\boldsymbol{\beta}, s)/S^{(0)}(\boldsymbol{\beta}, s), \\
V(\boldsymbol{\beta}, s) &= S^{(2)}(\boldsymbol{\beta}, s)/S^{(0)}(\boldsymbol{\beta}, s) - \mathbf{E}(\boldsymbol{\beta}, s)^{\otimes 2},
\end{aligned}
$$

where $\mathbf{z}^{\otimes k} = 1, \mathbf{z}, \mathbf{z}\mathbf{z}'$ for $k = 0, 1, 2$, respectively. Let $||\cdot||_p$ be the ℓ_p-norm for vectors and the norm induced by the vector p-norm for matrices. The following technical conditions will be needed in our derivations to establish the statistical properties of the BAR estimator.

(C1) $\int_0^1 h_0(t)dt < \infty$;

(C2) There exists some compact neighborhood, $\mathcal{B}_0$, of the true value $\boldsymbol{\beta}_0$ such that for $k = 0, 1, 2$, there exists a scalar, vector, and matrix function $s^{(k)}(\boldsymbol{\beta}, t)$ defined on $\mathcal{B}_0 \times [0, 1]$ such that

$$
\sup_{t\in[0,1],\boldsymbol{\beta}\in\mathcal{B}_0} \left\|S^{(k)}(\boldsymbol{\beta}, t) - s^{(k)}(\boldsymbol{\beta}, t)\right\|_2 = o_p(1), \quad \text{as } n \to \infty;
$$

(C3) Let $s^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial\boldsymbol{\beta}}s^{(0)}(\boldsymbol{\beta}, t)$ and $s^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial\boldsymbol{\beta}}s^{(1)}(\boldsymbol{\beta}, t)$. For $k = 0, 1, 2$, the functions $s^{(k)}(\boldsymbol{\beta}, t)$ are continuous with respect to $\boldsymbol{\beta} \in \mathcal{B}_0$, uniformly in $t \in [0, 1]$, and $s^{(k)}(\boldsymbol{\beta}, t)$ are bounded; furthermore, $s^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from zero on $\mathcal{B}_0 \times [0, 1]$;

(C4) Let $e(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$, $v(\boldsymbol{\beta}, t) = s^{(2)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t) - e(\boldsymbol{\beta}, t)^{\otimes 2}$, and $\Sigma(\boldsymbol{\beta}) = \int_0^1 v(\boldsymbol{\beta}, t)s^{(0)}(\boldsymbol{\beta}, t)h_0(t)dt$.

There exists some constant $C_1 > 0$ such that

$$0 < C_1^{-1} < \text{eigen}_{\min}\{\Sigma(\boldsymbol{\beta})\} \leq \text{eigen}_{\max}\{\Sigma(\boldsymbol{\beta})\} < C_1 < \infty,$$

uniformly in $\boldsymbol{\beta} \in \mathcal{B}_0$, where for any matrix A, $\text{eigen}_{\min}(A)$ and $\text{eigen}_{\max}(A)$ represent its smallest and largest eigenvalues, respectively;

(C5) Let $\mathbf{U}_i = \int_0^1 \{\mathbf{z}_i(t) - e(\boldsymbol{\beta}_0, t)\}\, dM_i(t)$. There exists a constant C_2 such that $\sup_{1 \leq i \leq n} E(U_{ij}^2 U_{il}^2) < C_2 < \infty$ for all $1 \leq j, l \leq p_n$, where U_{ij} is the j-th element of $\mathbf{U}_i$;

(C6) As $n \to \infty$, $p_n^4/n \to 0$, $\lambda_n \to \infty$, $\xi_n \to \infty$, $\xi_n b_n/\sqrt{n} \to 0$, $p_n/(na_n^2) \to 0$, $\lambda_n b_n^3 \sqrt{q_n}/\sqrt{n} \to 0$ and $\lambda_n \sqrt{q_n}/(a_n^3 \sqrt{n}) \to 0$, where $a_n = \min_{j=1,\ldots,q_n}(|\beta_{0j}|)$ and $b_n = \max_{j=1,\ldots q_n}(|\beta_{0j}|)$.

Condition (C1) ensures a finite baseline cumulative hazard over the interval $[0, 1]$. Condition (C2) ensures the asymptotic stability of $S^{(k)}(\boldsymbol{\beta}, t)$, as required for Cox regression under fixed dimension. Under diverging dimension, it follows from Theorem 3.1 of Kosorok and Ma (2007) that under certain regularity conditions, $\sup_{t \in [0,1], \boldsymbol{\beta} \in \mathcal{B}_0} \left\| S^{(k)}(\boldsymbol{\beta}, t) - s^{(k)}(\boldsymbol{\beta}, t) \right\|_2 \leq \sqrt{p_n \ln p_n / n}$, which implies that (C2) holds if $p_n \ln p_n / n \to 0$. Condition (C3) is an asymptotic regularity condition similar to that for the fixed dimension Cox model. Condition (C4) guarantees that the covariance matrix of the score function is positive definite and has uniformly bounded eigenvalues for all n and $\boldsymbol{\beta} \in \mathcal{B}_0$. Other authors in the variable selection literature have also required a slightly weaker condition (Fan and Peng, 2004; Cai et al., 2005; Cho and Qu, 2013; Ni et al., 2016). Condition (C5) is needed to prove the Lindeberg condition under diverging dimension in our proof. Condition (C6) specifies the divergence or convergence rates for the model size, the penalty tuning parameters, and the lower and upper bound of the true signal. These technical assumptions are only sufficient conditions for our theoretical derivations and it is possible that our theoretical results hold under weaker conditions. For instance, we have observed in empirical studies that the BAR method has good performance even when p_n is at the same order as n. Further efforts to relax these technical conditions are warranted in future research.

A3.2 Proof of Theorem 3.1

To prove Theorem 3.1, we first establish five lemmas.

Lemma 3.1 (Asymptotic Variance of $\mathbf{U}_i$) *Let $\mathbf{U}_i = \int_0^1 \{\mathbf{z}_i(t) - e(\boldsymbol{\beta}_0, t)\}\, dM_i(t)$ be defined as in Condition (C5) and $\Sigma = \Sigma(\boldsymbol{\beta}_0) = \int_0^1 v(\boldsymbol{\beta}_0, t) s^{(0)}(\boldsymbol{\beta}_0, t) h_0(t) dt$, $\mathbf{e}(\boldsymbol{\beta}_0, t)$, and $v(\boldsymbol{\beta}_0, t)$ be defined as in Condition (C4). Then under Conditions (C1) - (C4),*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Var(\mathbf{U}_i) - \Sigma \right\|_2 = o_p(1), \tag{3.7}$$

as $n \to \infty$.

Proof. Denote by U_{ij} the jth element of $\mathbf{U}_i$ and $e_j(\boldsymbol{\beta}_0, s)$ as the jth element of $\mathbf{e}(\boldsymbol{\beta}_0, s)$. Then,

$$\begin{aligned} Cov(U_{ij}, U_{ik}) &= \left\langle \int_0^1 \{z_{ij}(s) - e_j(\boldsymbol{\beta}_0, s)\} dM_i(s), \int_0^1 \{z_{ik}(s) - e_k(\boldsymbol{\beta}_0, s)\} dM_i(s) \right\rangle \\ &= \int_0^1 \{z_{ij}(s) - e_j(\boldsymbol{\beta}_0, s)\}\{z_{ik}(s) - e_k(\boldsymbol{\beta}_0, s)\} Y_i(s) h_0(s) \exp\{\boldsymbol{\beta}' \mathbf{z}_i(s)\} ds. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} Var(\mathbf{U}_i) &= \int_0^1 \frac{1}{n} \sum_{i=1}^{n} h_0(s) Y_i(s) \mathbf{z}_i(s)^{\otimes 2} \exp\{\boldsymbol{\beta}_0' \mathbf{z}_i(s)\} ds \\ &- \int_0^1 \frac{1}{n} \sum_{i=1}^{n} h_0(s) Y_i(s) \mathbf{z}_i(s) \mathbf{e}(\boldsymbol{\beta}_0, s)' \exp(\boldsymbol{\beta}_0' \mathbf{z}_i(s)\} ds \\ &- \int_0^1 \frac{1}{n} \sum_{i=1}^{n} h_0(s) Y_i(s) \mathbf{e}(\boldsymbol{\beta}_0, s) \mathbf{z}_i'(s) \exp\{\boldsymbol{\beta}_0' \mathbf{z}_i(s)\} ds \\ &+ \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s)^{\otimes 2} \frac{1}{n} \sum_{i=1}^{n} h_0(s) Y_i(s) \exp(\boldsymbol{\beta}_0' \mathbf{z}_i(s)\} ds \\ &= \int_0^1 S^{(2)}(\boldsymbol{\beta}_0, s) h_0(s) ds - \int_0^1 S^{(1)}(\boldsymbol{\beta}_0, s) \mathbf{e}(\boldsymbol{\beta}_0, s)' h_0(s) ds \\ &- \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s) S^{(1)}(\boldsymbol{\beta}_0, s)' h_0(s) ds + \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s)^{\otimes 2} S^{(0)}(\boldsymbol{\beta}_0, s) h_0(s) ds. \end{aligned}$$

Also note that

$$
\begin{aligned}
\Sigma(\boldsymbol{\beta}_0) &= \int_0^1 v(\boldsymbol{\beta}_0, s) s^{(0)}(\boldsymbol{\beta}_0, s) h_0(s) ds \\
&= \int_0^1 \left\{ \frac{s^{(2)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} - \mathbf{e}(\boldsymbol{\beta}_0, s)^{\otimes 2} \right\} s^{(0)}(\boldsymbol{\beta}_0, s) h_0(s) ds \\
&= \int_0^1 s^{(2)}(\boldsymbol{\beta}_0, s) h_0(s) ds - \int_0^1 s^{(1)}(\boldsymbol{\beta}_0, s) \mathbf{e}(\boldsymbol{\beta}_0, s)' h_0(s) ds \\
&- \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s) s^{(1)}(\boldsymbol{\beta}_0, s)' h_0(s) ds + \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s)^{\otimes 2} s^{(0)}(\boldsymbol{\beta}_0, s) h_0(s) ds,
\end{aligned}
$$

since $e(\boldsymbol{\beta}_0, t) = s^{(1)}(\boldsymbol{\beta}_0, t)/s^{(0)}(\boldsymbol{\beta}_0, t)$. Therefore,

$$
\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n Var(\mathbf{U}_i) - \Sigma(\boldsymbol{\beta}_0) \right\|_2 &\leq \left\| \int_0^1 \left\{ S^{(2)}(\boldsymbol{\beta}_0, s) - s^{(2)}(\boldsymbol{\beta}_0, s) \right\} h_0(s) ds \right\|_2 \\
&+ \left\| \int_0^1 \left\{ S^{(1)}(\boldsymbol{\beta}_0, s) - s^{(1)}(\boldsymbol{\beta}_0, s) \right\} \mathbf{e}(\boldsymbol{\beta}_0, s)' h_0(s) ds \right\|_2 \\
&+ \left\| \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s) \left\{ S^{(1)}(\boldsymbol{\beta}_0, s) - s^{(1)}(\boldsymbol{\beta}_0, s) \right\}' h_0(s) ds \right\|_2 \\
&+ \left\| \int_0^1 \mathbf{e}(\boldsymbol{\beta}_0, s)^{\otimes 2} \left\{ S^{(0)}(\boldsymbol{\beta}_0, s) - s^{(0)}(\boldsymbol{\beta}_0, s) \right\} h_0(s) ds \right\|_2 \\
&= o(1),
\end{aligned}
$$

where the last step follows from Conditions (C1), (C2), and (C3). □

Lemma 3.2 (Asymptotic Normality of the Score Function) *Let $l_n(\boldsymbol{\beta})$ be the log-partial likelihood as defined in (3.2). For any p_n-dimensional vector $\mathbf{d}_n$ such that $||\mathbf{d}_n||_2 = 1$, under Conditions (C1) - (C6), we have*

$$
n^{-1/2} \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} \dot{l}_n(\boldsymbol{\beta}_0) \xrightarrow{D} N(0, 1), \tag{3.8}
$$

where $\dot{l}_n(\boldsymbol{\beta}_0)$ is the first derivative of $l_n(\boldsymbol{\beta}_0)$ and $\Sigma(\boldsymbol{\beta}_0)$ is defined in Condition (C4).

Proof: First, observe that

$$\begin{aligned}
\dot{l}_n(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n} \int_0^1 \{\mathbf{z}_i(t) - \mathbf{E}(\boldsymbol{\beta}_0, s)\}\, dM_i(s) \\
&= \sum_{i=1}^{n} \int_0^1 \{\mathbf{z}_i(t) - \mathbf{e}(\boldsymbol{\beta}_0, s)\}\, dM_i(s) - \sum_{i=1}^{n} \int_0^1 \{\mathbf{E}(\boldsymbol{\beta}_0, s) - \mathbf{e}(\boldsymbol{\beta}_0, s)\}\, dM_i(s) \\
&= \sum_{i=1}^{n} \mathbf{U}_i + o_p(\sqrt{n}), \qquad (3.9)
\end{aligned}$$

where $\mathbf{U}_i$ is defined as in condition (C4), and the right-hand side of the last equality is due to $||\mathbf{E}(\boldsymbol{\beta}_0, s) - \mathbf{e}(\boldsymbol{\beta}_0, s)||_2 \to o_p(1)$ from conditions (C2) and (C3), and $n^{-1/2} \sum_{i=1}^{n} \int_0^1 dM_i(s) = O_p(1)$. Therefore

$$n^{-1/2} \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} \dot{l}_n(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} Y_{ni} + o_p(1),$$

where $Y_{ni} = n^{-1/2} \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} \mathbf{U}_i$. Note that Y_{ni} has mean zero and

$$\begin{aligned}
s_n^2 = \sum_{i=1}^{n} Var(Y_{ni}) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} Var(\mathbf{U}_i) \Sigma(\boldsymbol{\beta}_0)^{-1/2} \mathbf{d}_n \\
&= \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} Var(\mathbf{U}_i) \right\} \Sigma(\boldsymbol{\beta}_0)^{-1/2} \mathbf{d}_n \to 1,
\end{aligned}$$

where the last step follows from Lemma 3.1. Hence by the Lindeberg-Feller central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_{ni}}{s_n} \xrightarrow{D} N(0, 1), \qquad (3.10)$$

if the following Lindeberg condition for Y_{ni} holds: for all $\epsilon > 0$,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} E\{Y_{ni}^2 I(|Y_{ni}| \geq \epsilon s_n)\} \to 0, \qquad (3.11)$$

as $n \to \infty$. To verify (3.11) we note that

$$
\begin{aligned}
\sum_{i=1}^{n} E(Y_{ni}^4) &= n^{-2} \sum_{i=1}^{n} E\left[\left\{\mathbf{d}_n' \Sigma^{-1/2} \mathbf{U}_i\right\}^4\right] \\
&\leq n^{-2} \sum_{i=1}^{n} E\left[||\mathbf{d}_n||_2^4 \cdot ||\Sigma(\boldsymbol{\beta}_0)^{-1/2}||_2^4 \cdot ||\mathbf{U}_i||_2^4\right] \\
&= n^{-2} \text{eigen}_{\max}^2\{\Sigma(\boldsymbol{\beta}_0)^{-1}\} \sum_{i=1}^{n} E(||\mathbf{U}_i||_2^4) \\
&= n^{-2} \text{eigen}_{\max}^2\{\Sigma(\boldsymbol{\beta}_0)^{-1}\} \sum_{i=1}^{n} \sum_{j=1}^{p_n} \sum_{k=1}^{p_n} E(U_{ij}^2 U_{ik}^2) \\
&= O(p_n^2/n), \qquad (3.12)
\end{aligned}
$$

where the first inequality is due to Cauchy-Schwarz, the second equality is due to $||\mathbf{d}_n||_2 = 1$, Condition (C4) and the definition of the spectral norm, and the last step follows from Condition (C5). Therefore for any $\epsilon > 0$,

$$
\begin{aligned}
\frac{1}{s_n^2} \sum_{i=1}^{n} E\left\{Y_{ni}^2 I(|Y_{ni}| > \epsilon s_n)\right\} &\leq \frac{1}{s_n^2} \sum_{i=1}^{n} \left\{E(Y_{ni}^4)\right\}^{1/2} \left[E\left\{I(|Y_{ni}| > \epsilon s_n)\right\}^2\right]^{1/2} \\
&\leq \frac{1}{s_n^2} \left\{\sum_{i=1}^{n} E(Y_{ni}^4)\right\}^{1/2} \cdot \left\{\sum_{i=1}^{n} \Pr(|Y_{ni}| > \epsilon s_n)\right\}^{1/2} \\
&\leq \frac{1}{s_n^2} O(p_n/\sqrt{n}) \cdot \left\{\sum_{i=1}^{n} \frac{Var(Y_{ni})}{\epsilon^2 s_n^2}\right\}^{1/2} \\
&= \frac{1}{s_n^2 \epsilon} O(p_n/\sqrt{n}) \to 0,
\end{aligned}
$$

where the third inequality follows (3.12) and Chebyshev inequality, and last step is a consequence of $s_n^2 \to 1$ and the assumption $p_n^4/n \to 0$. Thus, (3.11) is satisfied and consequently

$$
n^{-1/2} \mathbf{d}_n' \Sigma(\boldsymbol{\beta}_0)^{-1/2} \dot{l}_n(\boldsymbol{\beta}_0) = s_n \frac{1}{s_n} \sum_{i=1}^{n} Y_{ni} + o_p(1) \xrightarrow{D} N(0,1),
$$

by the Lindeberg-Feller central limit theorem and Slutsky's theorem. This completes the proof. $\square$

Lemma 3.3 (Consistency of Ridge Estimator) *Let*

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} \xi_n \beta_j^2 \right\},$$

be the Cox ridge estimator defined in Equation (3.3). Assume that Conditions (C1) - (C5), and (C6)(i) and (C6)(iii) hold. Then

$$||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p\left[\sqrt{p_n}\{n^{-1/2}(1 + \xi_n b_n/\sqrt{n})\}\right] = O_p(\sqrt{p_n/n}), \tag{3.13}$$

where b_n is an upper bound of the true nonzero $|\beta_{0j}|$'s defined in Condition (C6).

Proof. Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + \xi_n b_n/n)$ and $\ell_n(\boldsymbol{\beta}) = -2l_n(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p_n} \beta_j^2$. To prove Lemma 3.3, it is sufficient to show that for any $\epsilon > 0$, there exists a large enough constant K_0 such that

$$\Pr\left\{ \inf_{||\mathbf{u}||_2 = K_0} L_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) > L_n(\boldsymbol{\beta}_0) \right\} \geq 1 - \epsilon, \tag{3.14}$$

since (3.14) implies that there exists a local minimum, $\hat{\boldsymbol{\beta}}_{ridge}$, inside the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : ||\mathbf{u}||_2 \leq K_0\}$ such that $||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p(\alpha_n)$, with probability tending to one. To prove (3.14), we first note

$$\begin{aligned}
\frac{1}{n}L_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \frac{1}{n}L_n(\boldsymbol{\beta}_0) &= -\frac{1}{n}\{2l_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \frac{1}{n}2l_n(\boldsymbol{\beta}_0)\} + \frac{\xi_n}{n}\sum_{j=1}^{p_n}\left\{(\beta_{0j} + \alpha_n u_j)^2 - \beta_{0j}^2\right\} \\
&= -\frac{1}{n}\{2l_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - 2l_n(\boldsymbol{\beta}_0)\} + \frac{\xi_n}{n}\sum_{j=1}^{p_n}\left(2\beta_{0j}\alpha_n u_j + \alpha_n^2 u_j^2\right) \\
&\geq -\frac{1}{n}\{2l_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - 2l_n(\boldsymbol{\beta}_0)\} + \frac{2\xi_n\alpha_n}{n}\sum_{j=1}^{p_n}\beta_{0j}u_j \\
&= -\frac{1}{n}\{2l_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - 2l_n(\boldsymbol{\beta}_0)\} + \frac{2\xi_n\alpha_n}{n}\sum_{j=1}^{q_n}\beta_{0j}u_j \\
&\equiv W_1 + W_2.
\end{aligned}$$

By Taylor expansion, we have

$$\begin{aligned} W_1 &= -\frac{2}{n}\alpha_n \mathbf{u}'\dot{l}_n(\boldsymbol{\beta}_0) - \frac{1}{n}\alpha_n^2 \mathbf{u}'\ddot{l}_n(\boldsymbol{\beta}^*)\mathbf{u} \\ &= W_{11} + W_{12}, \end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}$, and $\dot{l}_n(\boldsymbol{\beta})$ and $\ddot{l}_n(\boldsymbol{\beta})$ denote the first and second derivatives of $l_n(\boldsymbol{\beta})$, respectively. By the Cauchy-Schwartz inequality,

$$W_{11} = -\frac{2}{n}\alpha_n \mathbf{u}'\dot{l}_n(\boldsymbol{\beta}_0) \leq \frac{2}{n}\alpha_n ||\dot{l}_n(\boldsymbol{\beta}_0)||_2 \cdot ||\mathbf{u}||_2 = \frac{2}{n}\alpha_n O_p(\sqrt{np_n})||\mathbf{u}||_2 \leq O_p(\alpha_n^2)||\mathbf{u}||_2,$$

where the second equality holds because $||\dot{l}_n(\boldsymbol{\beta}_0)||_2 = O_p(\sqrt{np_n})$ from Lemma 3.2 under Conditions (C1) - (C5), and the last inequality is due to $\sqrt{p_n/n} \leq \alpha_n$. By equation (A.4) of Cai et al. (2005), under conditions (C1) - (C5) and $p_n^4/n \to 0$, we have

$$\left\| n^{-1}\ddot{l}_n(\boldsymbol{\beta}) + \Sigma(\boldsymbol{\beta}) \right\|_2 = o_p(p_n^{-1}), \tag{3.15}$$

in probability, uniformly in $\boldsymbol{\beta} \in \mathcal{B}_0$. Hence

$$W_{12} = -\frac{1}{n}\alpha_n^2 \mathbf{u}'\ddot{l}_n(\boldsymbol{\beta}^*)\mathbf{u} = \alpha_n^2 \mathbf{u}'\Sigma(\boldsymbol{\beta}_0)\mathbf{u}\{1 + o_p(1)\}.$$

Since $\text{eigen}_{min}\{\Sigma(\boldsymbol{\beta}_0)\} \geq C_1^{-1} > 0$ by Condition (C4), W_{12} dominates W_{11} uniformly in $||\mathbf{u}||_2 = K_0$ for a sufficiently large K_0. Furthermore

$$W_2 \leq \frac{2\xi_n \alpha_n}{n}|\boldsymbol{\beta}_{01}'\mathbf{u}| \leq \frac{2\sqrt{q_n}\xi_n \alpha_n b_n}{n}||\mathbf{u}||_2 = O_p(\alpha_n^2)||\mathbf{u}||_2,$$

where the last step follows from the fact that $\sqrt{q_n}\xi_n b_n/n < \sqrt{p_n}(n^{-1/2} + \xi_n b_n/n) = \alpha_n$. Therefore for a sufficiently large K_0, we have that W_{12} dominates W_{11} and W_2 uniformly in $||\mathbf{u}||_2 = K_0$. Since W_{12} is positive, (3.14) holds and therefore $||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p(\alpha_n) = O_p\left[\sqrt{p_n}\{n^{-1/2}(1 + \xi_n b_n/\sqrt{n})\}\right] = O_p(\sqrt{p_n/n})$, where the last step follows from condition (C6)(iii). □

Remark 3.3 *Let $\hat{\boldsymbol{\beta}}_{ridge,1}$ and $\hat{\boldsymbol{\beta}}_{ridge,2}$ denote the first q_n and the remaining $p_n - q_n$ components of $\hat{\boldsymbol{\beta}}_{ridge}$, respectively. Then, Lemma 3.3 and condition (C6) imply that for $j = 1, \ldots, q_n$ and sufficiently large n, $a_n/2 \leq |\hat{\beta}_{ridge,1j}| \leq 2b_n$, where $\hat{\beta}_{ridge,1j}$ is the jth component of $\hat{\boldsymbol{\beta}}_{ridge,1}$ and $||\hat{\boldsymbol{\beta}}_{ridge,2}||_2 = O(\sqrt{p_n/n})$.*

Lemma 3.4 *Let $M_n = \max\{2/a_n, 2b_n\}$. Define $\mathcal{H}_n \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' : |\boldsymbol{\beta}_1| = (|\beta_1|, \ldots, |\beta_{q_n}|)' \in [1/M_n, M_n]^{q_n}, 0 < \|\boldsymbol{\beta}_2\|_2 \leq \delta_n\sqrt{p_n/n}, \}$, where δ_n is a sequence of positive real numbers satisfying $\delta_n \to \infty$ and $p_n\delta_n^2/\lambda_n \to 0$. For any given $\boldsymbol{\beta} \in \mathcal{H}_n$, define*

$$Q_n(\boldsymbol{\theta}|\boldsymbol{\beta}) = -2l_n(\boldsymbol{\theta}) + \lambda_n\boldsymbol{\theta}'D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{3.16}$$

where $l_n(\boldsymbol{\theta})$ is the p_n-dimensional log-partial likelihood and $D(\boldsymbol{\beta}) = diag(\beta_1^{-2}, \ldots, \beta_{p_n}^{-2})$. Let $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})', g_2(\boldsymbol{\beta})')'$ be a solution to $\dot{Q}_n(\boldsymbol{\theta}|\boldsymbol{\beta}) = \mathbf{0}$, where

$$\dot{Q}_n(\boldsymbol{\theta}|\boldsymbol{\beta}) = -2\dot{l}_n(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{3.17}$$

is the derivative of $Q(\boldsymbol{\theta}|\boldsymbol{\beta})$ with respective to $\boldsymbol{\theta}$. Assume that conditions (C1) - (C6) hold. Then, as $n \to \infty$, with probability tending to 1,

(a) $\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \frac{\|g_2(\boldsymbol{\beta})\|_2}{\|\boldsymbol{\beta}_2\|_2} \leq \frac{1}{K_1}$, *for some constant $K_1 > 1$;*

(b) $|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}$.

Proof. By the first-order Taylor expansion and the definition of $g(\boldsymbol{\beta})$, we have

$$\dot{Q}_n(\boldsymbol{\beta}_0|\boldsymbol{\beta}) = \dot{Q}_n\{g(\boldsymbol{\beta})|\boldsymbol{\beta}\} + \ddot{Q}_n(\boldsymbol{\beta}^*|\boldsymbol{\beta})\{\boldsymbol{\beta}_0 - g(\boldsymbol{\beta})\} = \ddot{Q}_n(\boldsymbol{\beta}^*|\boldsymbol{\beta})\{\boldsymbol{\beta}_0 - g(\boldsymbol{\beta})\}, \tag{3.18}$$

where $\boldsymbol{\beta}_0$ is the true parameter vector, and $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $g(\boldsymbol{\beta})$. Rearranging terms, we have

$$\ddot{Q}_n(\boldsymbol{\beta}^*|\boldsymbol{\beta})g(\boldsymbol{\beta}) = -\dot{Q}_n(\boldsymbol{\beta}_0|\boldsymbol{\beta}) + \ddot{Q}_n(\boldsymbol{\beta}^*|\boldsymbol{\beta})\boldsymbol{\beta}_0, \tag{3.19}$$

which can be rewritten as

$$\left\{-2\ddot{l}_n(\boldsymbol{\beta}^*) + 2\lambda_n D(\boldsymbol{\beta})\right\} g(\boldsymbol{\beta}) = -\left\{-2\dot{l}_n(\boldsymbol{\beta}_0) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\beta}_0\right\} + \left\{-2\ddot{l}_n(\boldsymbol{\beta}^*) + 2\lambda_n D(\boldsymbol{\beta})\right\}\boldsymbol{\beta}_0$$
$$= 2\dot{l}_n(\boldsymbol{\beta}_0) - 2\ddot{l}_n(\boldsymbol{\beta}^*)\boldsymbol{\beta}_0.$$

Write $H_n(\boldsymbol{\beta}) \equiv -n^{-1}\ddot{l}_n(\boldsymbol{\beta})$, we have

$$\left\{H_n(\boldsymbol{\beta}^*) + \frac{\lambda_n}{n} D(\boldsymbol{\beta})\right\} g(\boldsymbol{\beta}) = H_n(\boldsymbol{\beta}^*)\boldsymbol{\beta}_0 + \frac{1}{n}\dot{l}_n(\boldsymbol{\beta}_0), \tag{3.20}$$

which can be further written as

$$\{g(\boldsymbol{\beta}) - \boldsymbol{\beta}_0\} + \frac{\lambda_n}{n} H_n(\boldsymbol{\beta}^*)^{-1} D(\boldsymbol{\beta}) g(\boldsymbol{\beta}) = \frac{1}{n} H_n(\boldsymbol{\beta}^*)^{-1}\dot{l}_n(\boldsymbol{\beta}_0). \tag{3.21}$$

Now we partition $H_n(\boldsymbol{\beta}^*)^{-1}$ into

$$H_n(\boldsymbol{\beta}^*)^{-1} = \begin{bmatrix} A & B \\ B' & G \end{bmatrix}$$

and partition $D(\boldsymbol{\beta})$ into

$$D(\boldsymbol{\beta}) = \begin{bmatrix} D_1(\boldsymbol{\beta}_1) & \mathbf{0} \\ \mathbf{0}' & D_2(\boldsymbol{\beta}_2) \end{bmatrix}$$

where $D_1(\boldsymbol{\beta}_1) = \text{diag}(\beta_1^{-2}, ..., \beta_{q_n}^{-2})$ and $D_2(\boldsymbol{\beta}_2) = \text{diag}(\beta_{q_n+1}^{-2}, ..., \beta_{p_n}^{-2})$. Then (3.21) can be re-written as

$$\begin{pmatrix} g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01} \\ g_2(\boldsymbol{\beta}) \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \\ B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \end{pmatrix} = \frac{1}{n} H_n(\boldsymbol{\beta}^*)^{-1}\dot{l}_n(\boldsymbol{\beta}_0). \tag{3.22}$$

Moreover, it follows from (3.15), condition (C4) and Lemma 3.2 that

$$\left\| n^{-1} H_n(\boldsymbol{\beta}^*)^{-1}\dot{l}_n(\boldsymbol{\beta}_0) \right\|_2 = O_p(\sqrt{p_n/n}). \tag{3.23}$$

Therefore,

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|g_2(\boldsymbol{\beta})+\frac{\lambda_n}{n}B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})+\frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2=O_p(\sqrt{p_n/n}). \tag{3.24}$$

Furthermore,

$$\begin{aligned}\|g(\boldsymbol{\beta})-\boldsymbol{\beta}_0\|_2 &= \left\|-\left\{H_n(\boldsymbol{\beta}^*)+\frac{\lambda_n}{n}D(\boldsymbol{\beta})\right\}^{-1}\left\{\frac{\lambda_n}{n}D(\boldsymbol{\beta})\boldsymbol{\beta}_0-\frac{1}{n}\dot{l}_n(\boldsymbol{\beta}_0)\right\}\right\|_2\\ &\le \left\|\{H_n(\boldsymbol{\beta}^*)\}^{-1}\left\{\frac{\lambda_n}{n}D(\boldsymbol{\beta})\boldsymbol{\beta}_0-\frac{1}{n}\dot{l}_n(\boldsymbol{\beta}_0)\right\}\right\|_2\\ &\le \left\|\{H_n(\boldsymbol{\beta}^*)\}^{-1}\right\|_2\cdot\left\{\left\|\frac{\lambda_n}{n}D_1(\boldsymbol{\beta}_1)\boldsymbol{\beta}_{01}\right\|_2+\left\|\frac{1}{n}\dot{l}_n(\boldsymbol{\beta}_0)\right\|_2\right\}\\ &= O_p(1)\left\{O(n^{-1}\lambda_n M_n^3\sqrt{q_n})+O_p(\sqrt{p_n/n})\right\}\\ &= O_p(\sqrt{p_n/n}),\end{aligned}$$

where the first equality follows from (3.20) and the fourth step follows from (3.15), condition (C4), $\|n^{-1}\lambda_n D_1(\boldsymbol{\beta}_1)\beta_{01}\|_2 = O(n^{-1}\lambda_n M_n^3\sqrt{q_n})$, and $\left\|n^{-1}\dot{l}_n(\boldsymbol{\beta}_0)\right\|_2 = O_p(\sqrt{p_n/n})$, and the last step holds since $n^{-1}\lambda_n M_n^3\sqrt{q_n}=o(1/\sqrt{n})$ under condition (C6). Hence,

$$\|g(\boldsymbol{\beta})\|_2\le\|\boldsymbol{\beta}_0\|_2+\|g(\boldsymbol{\beta})-\boldsymbol{\beta}_0\|_2=O_p(M_n\sqrt{q_n}). \tag{3.25}$$

Also note that $\|B\|_2 = O_p(1)$ since $\|BB'\|_2 \le \|A^2+BB'\|_2+\|A^2\|_2 \le 2\|A^2+BB'\|_2 \le 2\|H_n(\boldsymbol{\beta}^*)^{-2}\|_2 = O_p(1)$. This, combined with (3.25), implies that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|\frac{\lambda_n}{n}B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})\right\|_2\le\frac{\lambda_n}{n}\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\|B\|_2\,\|D_1(\boldsymbol{\beta}_1)\|_2\,\|g_1(\boldsymbol{\beta})\|_2=O_p\left(\frac{\lambda_n M_n^3\sqrt{q_n}}{n}\right)=o(1/\sqrt{n}). \tag{3.26}$$

It then follows from (3.24) and (3.26) that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|g_2(\boldsymbol{\beta})+\frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2\le O_p(\sqrt{p_n/n})+o(1/\sqrt{n})=O_p(\sqrt{p_n/n}).$$

Since G is positive definite and symmetric with probability tending to one, by the spectral decomposition theorem, $G=\sum_{i=1}^{p_n-q_n}r_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}'$, where r_{2i} and $\mathbf{u}_{2i}$ are the eigenvalues and

eigenvectors of G, respectively. Now with probability tending to one,

$$\begin{aligned}\frac{\lambda_n}{n}\left\|GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 &= \frac{\lambda_n}{n}\left\|\left(\sum_{i=1}^{p_n-q_n} r_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}'\right)D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \\ &\geq \frac{\lambda_n}{n}\left\|\frac{1}{C_1}\left(\sum_{i=1}^{p_n-q_n} \mathbf{u}_{2i}\mathbf{u}_{2i}'\right)D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \\ &\geq \frac{1}{C_1}\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2, \end{aligned} \tag{3.27}$$

where the first inequality is due to (3.15) and condition (C4) since we can assume that for all $i = 1, \ldots, p_n - q_n$, $r_{2i} \in (1/C_1, C_1)$ for some $C_1 > 1$ with probability tending to one. Therefore with probability tending to one,

$$\frac{1}{C_1}\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 - \left\|g_2(\boldsymbol{\beta})\right\|_2 \leq \left\|g_2(\boldsymbol{\beta}) + \frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \leq \delta_n\sqrt{p_n/n}, \tag{3.28}$$

where δ_n diverges to ∞. Let $\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2} = (g_2(\beta_{q_n+1})/\beta_{q_n+1}, \ldots, g_2(\beta_{p_n})/\beta_{p_n})'$. Because $||\boldsymbol{\beta}_2||_2 \leq \delta_n\sqrt{p_n/n}$, we have

$$\frac{1}{C_1}\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 = \frac{1}{C_1}\frac{\lambda_n}{n}\left\|D_2(\boldsymbol{\beta}_2)^{1/2}\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \geq \frac{1}{C_1}\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2, \tag{3.29}$$

and

$$\left\|g_2(\boldsymbol{\beta})\right\|_2 = \left\|D_2(\boldsymbol{\beta}_2)^{-1/2}\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \left\|D_2(\boldsymbol{\beta}_2)^{-1/2}\right\|_2 \cdot \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2. \tag{3.30}$$

Hence it follows from (3.28), (3.29), and (3.30) that with probability tending to one,

$$\frac{1}{C_1}\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 - \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \delta_n\sqrt{p_n/n}.$$

This implies that with probability tending to one,

$$\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \frac{1}{\lambda_n/(C_1p_n\delta_n^2) - 1} < \frac{1}{K_1}, \tag{3.31}$$

for some constant $K_1 > 1$ provided that $\lambda_n/(p_n\delta_n^2) \to \infty$ as $n \to \infty$. Now from (3.31), we have

$$\|g_2(\boldsymbol{\beta})\|_2 \leq \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \max_{q_n+1\leq j\leq p_n} |\beta_j| \leq \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \|\boldsymbol{\beta}_2\|_2 \leq \frac{1}{K_1}\|\boldsymbol{\beta}_2\|_2, \tag{3.32}$$

with probability tending to one. Thus

$$\Pr\left(\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \frac{\|g_2(\boldsymbol{\beta})\|_2}{\|\boldsymbol{\beta}_2\|_2} < \frac{1}{K_1}\right) \to 1 \quad \text{as } n \to \infty$$

and (a) is proved.

To prove part (b), we first note from (3.32) that as $n \to \infty$, $\Pr(\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \delta_n\sqrt{p_n/n}) \to 1$. Therefore it is sufficient to show that for any $\boldsymbol{\beta} \in \mathcal{H}_n$, $|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}$ with probability tending to 1. By (3.22) and (3.23), we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|(g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}) + \frac{\lambda_n}{n}AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + \frac{\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 = O_p(\sqrt{p_n/n}). \tag{3.33}$$

Similar to (3.26), it can be shown that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|\frac{\lambda_n}{n}AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})\right\|_2 = O_p\left(\frac{\lambda_n M_n^3\sqrt{q_n}}{n}\right) = o_p(1/\sqrt{n}), \tag{3.34}$$

where the last equality holds trivially under condition (C6). Furthermore, with probability tending to one,

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|\frac{\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \leq \frac{\lambda_n}{n}\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\|B\|_2 \cdot \|D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\|_2 \leq \frac{\lambda_n}{n}\sqrt{2K_3}\left(\delta_n\sqrt{\frac{p_n}{n}}\right)^2, \tag{3.35}$$

for some $K_3 > 0$, since $||g_2(\boldsymbol{\beta})|| \leq \delta_n\sqrt{p_n/n}$, $||B||_2 = O_p(1)$ and $\|D_2(\boldsymbol{\beta}_2)\|_2 \leq \delta_n\sqrt{p_n/n}$. Therefore, combining (3.33), (3.34) and (3.35) gives

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}\|_2 \leq \frac{\lambda_n}{n}\sqrt{2K_3}\left(\delta_n\sqrt{\frac{p_n}{n}}\right)^2 + \frac{\delta_n\sqrt{p_n}}{\sqrt{n}},$$

with probability tending to one. Because $\lambda_n/n \to 0$ and $\delta_n\sqrt{p_n/n} = \sqrt{p_n\delta_n^2/\lambda_n}\sqrt{\lambda_n/n} \to 0$ as $n \to \infty$, we have $\Pr(|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}) \to 1$. This completes the proof of part (b). $\square$

Lemma 3.5 *Let $\boldsymbol{\beta}_1$ be the first q_n components of $\boldsymbol{\beta}$. Define $f(\boldsymbol{\beta}_1) = \arg\min_{\boldsymbol{\theta}_1}\{Q_{n1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1)\}$, where $Q_{n1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = -2l_{n1}(\boldsymbol{\theta}_1) + \lambda_n\boldsymbol{\theta}_1'D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1$, is a weighted ℓ_2-penalized -2log-partial likelihood for the oracle model of model size q_n, and $D_1(\boldsymbol{\beta}_1) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_{q_n}^{-2})$. Assume that conditions (C1) - (C6) hold. Then with probability tending to one,*

(a) $f(\boldsymbol{\beta}_1)$ is a contraction mapping from $[1/M_n, M_n]^{q_n}$ to itself;

(b) $\sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0,1)$, for any q_n-dimensional vector $\mathbf{b}_n$ such that $\mathbf{b}_n'\mathbf{b}_n = 1$ and where $\hat{\boldsymbol{\beta}}_1^\circ$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$ and $\Sigma(\boldsymbol{\beta}_0)_{11}$ is the first $q_n \times q_n$ submatrix of $\Sigma(\boldsymbol{\beta}_0)$.

Proof: (a) First we show that $f(\cdot)$ is a mapping from $[1/M_n, M_n]^{q_n}$ to itself with probability tending to one. Again through a first order Taylor expansion, we have

$$\{f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01}\} + \frac{\lambda_n}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1) = \frac{1}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_{n1}(\boldsymbol{\beta}_{01}), \tag{3.36}$$

where $H_{n1}(\boldsymbol{\beta}_1^*) = -n^{-1}\ddot{l}_{n1}(\boldsymbol{\beta}_1^*)$ exists and is invertible for $\boldsymbol{\beta}_1^*$ between $\boldsymbol{\beta}_{01}$ and $f(\boldsymbol{\beta}_1)$. We have

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_n,M_n]^{q_n}} \left\| f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01} + \frac{\lambda_n}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1)\right\|_2 = O_p(\sqrt{q_n/n}),$$

where the right-hand side follows in the same fashion as (3.24). Similar to (3.26) we have

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_0,M_0]^{q_n}} \left\| \frac{\lambda_n}{n}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1)\right\|_2 = O_p\left(\frac{\lambda_n M_n^3}{\sqrt{n}}\sqrt{\frac{q_n}{n}}\right) = o_p\left(1/\sqrt{n}\right).$$

Therefore, with probability tending to one

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_n,M_n]^{q_n}} \|f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01}\|_2 \le \delta_n\sqrt{q_n/n}, \tag{3.37}$$

where δ_n is a sequence of real numbers diverging to ∞ and satisfies $\delta_n\sqrt{p_n/n} \to 0$. As a result, we have

$$\Pr(f(\boldsymbol{\beta}_1) \in [1/M_n, M_n]^{q_n}) \to 1$$

as $n \to \infty$. Hence $f(\cdot)$ is a mapping from the region $[1/M_n, M_n]^{q_n}$ to itself. To prove that $f(\cdot)$ is a contraction mapping, we need to further show that

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2 = o_p(1). \tag{3.38}$$

Since $f(\boldsymbol{\beta}_1)$ is a solution to $\dot{Q}_{n1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = 0$, we have

$$-\frac{1}{n}\dot{l}_{n1}(f(\boldsymbol{\beta}_1)) = -\frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1). \tag{3.39}$$

Taking the derivative of (3.39) with respect to $\boldsymbol{\beta}_1'$ and rearranging terms, we obtain

$$\left\{ H_{n1}(f(\boldsymbol{\beta}_1)) + \frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) \right\} \dot{f}(\boldsymbol{\beta}_1) = \frac{2\lambda_n}{n} diag\{f_1(\boldsymbol{\beta}_1)/\beta_1^3, \ldots, f_{q_n}(\boldsymbol{\beta}_1)/\beta_{q_n}^3\}. \tag{3.40}$$

With probability tending to one, we have

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \frac{2\lambda_n}{n} \left\| diag\{f_1(\boldsymbol{\beta}_1)/\beta_1^3, \ldots, f_{q_n}(\boldsymbol{\beta}_1)/\beta_{q_n}^3\} \right\|_2 = O_p\left(\frac{\lambda_n M_n^4}{n}\right) = o_p(1),$$

where the last step follows from condition (C6). This, combined with (3.40) implies that

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \left\| \left\{ H_{n1}(f(\boldsymbol{\beta}_1)) + \frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) \right\} \dot{f}(\boldsymbol{\beta}_1) \right\|_2 = o_p(1). \tag{3.41}$$

Now, it can be shown that probability tending to one,

$$\left\| H_{n1}(f(\boldsymbol{\beta}_1)) \dot{f}(\boldsymbol{\beta}_1) \right\|_2 \geq \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2 \cdot \left\| H_{n1}(f(\boldsymbol{\beta}_1))^{-1} \right\|_2^{-1} \geq \frac{1}{K_2} \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2,$$

for some $K_2 > 0$, and that

$$\frac{\lambda_n}{n}\left\|D_1(\boldsymbol{\beta}_1)\dot{f}(\boldsymbol{\beta}_1)\right\|_2 \geq \frac{\lambda_n}{n}\left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2 \left\|D_1(\boldsymbol{\beta}_1)^{-1}\right\|_2^{-1} \geq \frac{\lambda_n}{n}\frac{1}{M_n^2}\left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2.$$

Therefore, combining the above two inequalities with (3.40) and (3.41) gives

$$\left(\frac{1}{K_2} - \frac{\lambda_n}{nM_n^2}\right) \sup_{|\boldsymbol{\beta}_1|\in[1/M_n, M_n]^{q_n}} \left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2 = o_p(1).$$

This, together with the fact that $\frac{\lambda_n}{n}\frac{1}{M_n^2} = o(1)$, implies that (3.38) holds. Therefore, with probability tending to one, $f(\cdot)$ is a contraction mapping and consequently has a unique fixed point, say $\hat{\boldsymbol{\beta}}_1^\circ$, such that $\hat{\boldsymbol{\beta}}_1^\circ = f(\hat{\boldsymbol{\beta}}_1^\circ)$.

We next prove part (b). By (3.36) we have

$$f(\boldsymbol{\beta}_1) = \left\{H_{n1}(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\boldsymbol{\beta}_1)\right\}^{-1}\left\{H_{n1}(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01} + \frac{1}{n}\dot{l}_{n1}(\boldsymbol{\beta}_{01})\right\}.$$

Now,

$$\begin{aligned}
\sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) &= \sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\left[\left\{H_{n1}(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1} H_{n1}(\boldsymbol{\beta}_1^*) - I_{q_n}\right]\boldsymbol{\beta}_{01} \\
&\quad + \sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\left[\left\{H_{n1}(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_{n1}(\boldsymbol{\beta}_{01})\right] \\
&= I_1 + I_2. \qquad (3.42)
\end{aligned}$$

Note that for any two conformable invertible matrices Φ and Ψ, we have

$$(\Phi + \Psi)^{-1} = \Phi^{-1} - \Phi^{-1}\Psi(\Phi + \Psi)^{-1},$$

Thus we can rewrite I_1 as

$$
\begin{aligned}
I_1 &= \sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\left[\left\{H_{n1}(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}H_{n1}(\boldsymbol{\beta}_1^*)-I_{q_n}\right]\boldsymbol{\beta}_{01}\\
&= -\frac{\lambda_n}{\sqrt{n}}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\left\{H_{n1}(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}H_{n1}(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01}.
\end{aligned}
$$

$$
\begin{aligned}
\|I_1\|_2 &\le \frac{\lambda_n}{\sqrt{n}}\left\|\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\right\|_2\left\|H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\right\|_2\left\|D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\|_2\left\|\left\{H_{n1}(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\right\|_2\|H_{n1}(\boldsymbol{\beta}_1^*)\|_2\,\|\boldsymbol{\beta}_{01}\|_2\\
&= \frac{\lambda_n}{\sqrt{n}}\cdot O(1)\cdot O_p(1)\cdot M_n^2\cdot O_p(1)\cdot O_p(1)\cdot M_n\sqrt{q_n}\\
&= O_p(\lambda_n M_n^3\sqrt{q_n}/\sqrt{n}) = o_p(1), \qquad (3.43)
\end{aligned}
$$

where the first equality follows from (3.15) and condition (C4), and the last equality is a consequence of condition (C6). Similarly, we can rewrite I_2 as

$$
\begin{aligned}
I_2 &= \sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\left[\left\{H_{n1}(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_{n1}(\boldsymbol{\beta}_{01})\right]\\
&= \mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\frac{1}{\sqrt{n}}\dot{l}_{n1}(\boldsymbol{\beta}_{01})\\
&\quad -\frac{\lambda_n}{\sqrt{n}}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\left\{H_{n1}(\boldsymbol{\beta}_1^*)^{-1}+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_{n1}(\boldsymbol{\beta}_{01})\\
&= \mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\frac{1}{\sqrt{n}}\dot{l}_{n1}(\boldsymbol{\beta}_{01})+o_p(1). \qquad (3.44)
\end{aligned}
$$

We now establish the asymptotic normality of $n^{-1/2}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_{n1}(\boldsymbol{\beta}_{01})$, which will be derived in a similar fashion to Lemma 3.2. By (3.15), (3.37), and the continuity of $\Sigma(\boldsymbol{\beta}_0)$, we can deduce that $H_{n1}(\boldsymbol{\beta}^*) = \Sigma(\boldsymbol{\beta}_0)_{11} + o_p(1)$, where $\Sigma(\boldsymbol{\beta}_0)_{11} = \Sigma(\boldsymbol{\beta}_0)_{11}$ is the first $q_n \times q_n$

submatrix of $\Sigma(\boldsymbol{\beta}_0)$. This, together with (3.9) and (3.44), implies that

$$\begin{aligned} I_2 &= n^{-1/2}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\mathbf{U}_{i1} + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{U}_{i1} + \left\{n^{-1/2}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\mathbf{U}_{i1}\right\}o_p(1) + o_p(1) \\ &= I_{21} + I_{22}\cdot o_p(1) + o_p(1), \end{aligned} \tag{3.45}$$

where $\mathbf{U}_{i1}$ consists of the first q_n components of $\mathbf{U}_i$. Letting $Y_{ni} = n^{-1/2}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{U}_{i1}$, then

$$\begin{aligned} s_n^2 = \sum_{i=1}^{n} Var(Y_{ni}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}Var(\mathbf{U}_{i1})\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{b}_n \\ &= \mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}Var(\mathbf{U}_{i1})\right\}\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{b}_n \to 1. \end{aligned}$$

To prove the asymptotic normality of I_{21}, we need to verify the Lindeberg condition: for all $\epsilon > 0$,

$$\frac{1}{s_n^2}\sum_{i=1}^{n}E\{Y_{ni}^2 I(|Y_{ni}| \geq \epsilon s_n)\} \to 0, \tag{3.46}$$

as $n \to \infty$. Note that

$$\begin{aligned} \sum_{i=1}^{n}E(Y_{ni}^4) &= n^{-2}\sum_{i=1}^{n}E\left[\left\{\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{U}_{i1}\right\}^4\right] \\ &\leq n^{-2}\sum_{i=1}^{n}E\left[||\mathbf{b}_n||_2^4\cdot||\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}||_2^4\cdot||\mathbf{U}_{i1}||_2^4\right] \\ &= n^{-2}\text{eigen}_{\max}^2\{\Sigma(\boldsymbol{\beta}_0)^{-1}\}\sum_{i=1}^{n}E(||\mathbf{U}_{i1}||_2^4) \\ &= n^{-2}\text{eigen}_{\max}^2\{\Sigma(\boldsymbol{\beta}_0)^{-1}\}\sum_{i=1}^{n}\sum_{j=1}^{p_n}\sum_{k=1}^{p_n}E(U_{ij}^2U_{ik}^2) \\ &= O(p_n^2/n), \end{aligned} \tag{3.47}$$

where the first inequality is due to Cauchy-Schwarz, the second equality is due to $||\mathbf{b}_n||_2 = 1$ and the last step follows from conditions (C4) and (C5). Therefore for any $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{s_n^2}\sum_{i=1}^{n} E\left\{Y_{ni}^2 I(|Y_{ni}| > \epsilon s_n)\right\} &\leq \frac{1}{s_n^2}\sum_{i=1}^{n}\left\{E(Y_{ni}^4)\right\}^{1/2}\left[E\left\{I(|Y_{ni}| > \epsilon s_n)\right\}^2\right]^{1/2} \\
&\leq \frac{1}{s_n^2}\left\{\sum_{i=1}^{n} E(Y_{ni}^4)\right\}^{1/2} \cdot \left\{\sum_{i=1}^{n} \Pr(|Y_{ni}| > \epsilon s_n)\right\}^{1/2} \\
&\leq \frac{1}{s_n^2}\left\{\sum_{i=1}^{n} E(Y_{ni}^4)\right\}^{1/2} \cdot \left\{\sum_{i=1}^{n} \frac{Var(Y_{ni})}{\epsilon^2 s_n^2}\right\}^{1/2} \\
&= \frac{1}{s_n^2}\left\{O(p_n^2/n)\right\}^{1/2}\frac{1}{\epsilon} \to 0.
\end{aligned}$$

Thus, (3.46) is satisfied and by the Lindeberg-Feller central limit theorem and Slutsky's theorem

$$I_{21} = s_n\left(\frac{1}{s_n}\sum_{i=1}^{n} Y_{ni}\right) \xrightarrow{D} N(0,1). \tag{3.48}$$

Similarly, it can be shown that as $n \to \infty$,

$$\frac{I_{22}}{\sqrt{\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^2\mathbf{b}_n}} \xrightarrow{D} N\left(0,1\right). \tag{3.49}$$

since $\left\|\left\{\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^2\mathbf{b}_n + o(1)\right\}^{-1}\right\|_2 = O(1)$. Therefore $I_{22} = O_p(1)$ and by Slutsky's theorem,

$$\begin{aligned}
n^{-1/2}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}H_{n1}(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_{n1}(\boldsymbol{\beta}_{01}) &= n^{-1/2}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\mathbf{U}_{i1} \\
&\quad + \left\{n^{-1/2}\sum_{i=1}^{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}\mathbf{U}_{i1}\right\}o_p(1) + o_p(1) \\
&= I_{21} + I_{22}\cdot o_p(1) + o_p(1) \\
&\xrightarrow{D} N(0,1).
\end{aligned}$$

Hence, combining (3.42), (3.43), (3.45), (3.48) and (3.49) gives

$$\sqrt{n}\mathbf{b}_n'\Sigma(\boldsymbol{\beta}_0)_{11}^{1/2}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0,1),$$

which proves part (b). □

Proof of Theorem 3.1. Part (a) of the theorem follows immediately from part (a) of Lemma 3.4. Part (b) of the theorem will follow from part (b) Lemma 3.5 and the following

$$\Pr\left(\lim_{k\to\infty}\left\|g_1(\boldsymbol{\beta}^{(k)}) - \hat{\boldsymbol{\beta}}_1^\circ\right\|_2 = 0\right) \to 1, \tag{3.50}$$

where $\hat{\boldsymbol{\beta}}_1^\circ$ is the fixed point of $f(\boldsymbol{\beta}_1)$ defined in Lemma 3.5. Note that $g(\boldsymbol{\beta})$ is a solution to

$$-\frac{1}{n}D(\boldsymbol{\beta})^{-1}\dot{l}_n(\boldsymbol{\theta}) + \frac{1}{n}\lambda_n\boldsymbol{\theta} = \mathbf{0}, \tag{3.51}$$

where $D(\boldsymbol{\beta})^{-1} = diag\{\beta_1^2, \ldots, \beta_{q_n}^2, \beta_{q_n+1}^2, \ldots, \beta_{p_n}^2\}$. It is easy to see from (3.51) that

$$\lim_{\boldsymbol{\beta}_2\to 0} g_2(\boldsymbol{\beta}) = \mathbf{0}_{p_n-q_n}.$$

This, combined with (3.51), implies that for any $\boldsymbol{\beta}_1$

$$\lim_{\boldsymbol{\beta}_2\to 0} g_1(\boldsymbol{\beta}) = f(\boldsymbol{\beta}_1).$$

Hence, $g(\cdot)$ is continuous and thus uniform continuous on the compact set $\boldsymbol{\beta} \in \mathcal{H}_n$. Hence as $k \to \infty$,

$$\omega_k \equiv \sup_{|g_1(\boldsymbol{\beta})|\in[1/M_n, M_n]^{q_n}} \left\|g_1(\boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_2^{(k)}) - f(\boldsymbol{\beta}_1)\right\|_2 \to 0, \tag{3.52}$$

with probability tending to one. Furthermore,

$$\left\|\hat{\boldsymbol{\beta}}_1^{(k+1)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2 \leq \left\|g_1(\hat{\boldsymbol{\beta}}^{(k)}) - f(\hat{\boldsymbol{\beta}}_1^{(k)})\right\|_2 + \left\|f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2 \leq \omega_k + \frac{1}{K_4}\left\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2, \tag{3.53}$$

for some $K_4 > 1$ where the last inequality follows from (3.38) and the definition of ω_k. Denote by $a_k = \left\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2$, we can rewrite (3.53) as

$$a_{k+1} \leq \frac{1}{K_4} a_k + \omega_k.$$

By (3.52), for any $\epsilon > 0$, there exists an $N > 0$ such that $\omega_k < \epsilon$ for all $k > N$. Therefore for $k > N$,

$$\begin{aligned}
a_{k+1} &\leq \frac{1}{K_4} a_k + \omega_k \\
&\leq \frac{a_{k-1}}{K_4^2} + \frac{\omega_{k-1}}{K_4} + \omega_k \\
&\leq \frac{a_1}{K_4^k} + \frac{\omega_1}{K_4^{k-1}} + \cdots + \frac{\omega_N}{K_2^{k-N}} + \left(\frac{\omega_{N+1}}{K_4^{k-N-1}} + \cdots + \frac{\omega_{k-1}}{K_4} + \omega_k\right) \\
&\leq (a_1 + \omega_1 + ... + \omega_N)\frac{1}{K_4^{k-N}} + \frac{1-(1/K_4)^{k-N}}{1-1/K_4}\epsilon \to 0, \quad \text{as } k \to \infty,
\end{aligned}$$

with probability tending to one. Therefore,

$$\Pr\left(\lim_{k\to\infty}\left\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2 = \mathbf{0}\right) = 1$$

with probability tending to one, or equivalently

$$\Pr(\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^{\circ}) = 1 \tag{3.54}$$

with probability tending to one. This proves (3.50) and thus complete the proof of the theorem. □

A3.3 Proof of Theorem 3.2.

Under Conditions (C1) - (C6), by Theorem 3.1 we have that $\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}$, where

$$\hat{\boldsymbol{\beta}}^{(k+1)} = g(\hat{\boldsymbol{\beta}}^{(k)}) = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p_n} \frac{I(\beta_j \neq 0)\beta_j^2}{\left(\hat{\beta}_j^{(k)}\right)^2} \right\}.$$

Looking at (3.17) we have,

$$D(\hat{\boldsymbol{\beta}}^{(k)})^{-1}\dot{l}_n(\hat{\boldsymbol{\beta}}^{(k+1)}) = \lambda_n \hat{\boldsymbol{\beta}}^{(k+1)}$$

Therefore for any $l = i, j$ where $\hat{\beta}_i \neq 0$, $\hat{\beta}_j \neq 0$,

$$\hat{\beta}_l^{(k+1)} = \frac{(\hat{\beta}_l^{(k)})^2}{\lambda_n}\dot{l}_{nl}(\hat{\boldsymbol{\beta}}^{(k+1)}).$$

From Theorem 3.1, we also have that as $k \to \infty$, $\hat{\boldsymbol{\beta}}^{(k)} \to \hat{\boldsymbol{\beta}}$ and hence as $k \to \infty$, (4.60) can be rewritten as

$$\hat{\beta}_l^{-1} = \frac{1}{\lambda_n}\dot{l}_{nl}(\hat{\boldsymbol{\beta}}).$$

Let $\boldsymbol{\eta} = Z\boldsymbol{\beta}$ and

$$\zeta(\eta_i) = \frac{\partial}{\partial \eta_i} l_n(\boldsymbol{\beta}) = N_i(1) - \int_0^1 \frac{Y_i(s)\exp(\eta_i)}{\sum_{j=1}^n Y_j(s)\exp(\eta_j)} d\bar{N}(s) \quad i = 1, \ldots, n.$$

Then

$$|\zeta(\hat{\eta}_i)| \leq |N_i(1)| + \left| \int_0^1 \frac{Y_i(s)\exp(\hat{\eta}_i)}{\sum_{j=1}^n Y_j(s)\exp(\hat{\eta}_j)} d\bar{N}(s) \right| \leq 1 + d_n \quad i = 1, \ldots, n,$$

since the integrand is at most one and where $d_n = \sum_{i=1}^n \delta_i$. Hence

$$\|\zeta(\hat{\boldsymbol{\eta}})\|_2 \leq \|\mathbf{1} + d_n\mathbf{1}\|_2 = \sqrt{n(1+d_n)^2}.$$

Let $\mathbf{z}_{[,i]}$ denote the ith column of Z. Since Z is assumed to be standardized, $\mathbf{z}'_{[,i]}\mathbf{z}_{[,i]} = n-1$ and $\mathbf{z}'_{[,i]}\mathbf{z}_{[,j]} = (n-1)r_{ij}$, for all $i \neq j$ and where r_{ij} is the sample correlation between $\mathbf{z}_{[,i]}$ and $\mathbf{z}_{[,j]}$. Since

$$\hat{\beta}_i^{-1} = \frac{1}{\lambda_n}\mathbf{z}'_{[,i]}\zeta(\hat{\boldsymbol{\eta}}) \quad \text{and} \quad \hat{\beta}_j^{-1} = \frac{1}{\lambda_n}\mathbf{z}'_{[,j]}\zeta(\hat{\boldsymbol{\eta}}),$$

we have

$$\begin{aligned}
\left|\hat{\beta}_i^{-1} - \hat{\beta}_j^{-1}\right| &= \left|\frac{1}{\lambda_n}\mathbf{z}'_{[,i]}\zeta(\hat{\boldsymbol{\eta}}) - \frac{1}{\lambda_n}\mathbf{z}'_{[,j]}\zeta(\hat{\boldsymbol{\eta}})\right| \\
&= \left|\frac{1}{\lambda_n}(\mathbf{z}_{[,i]} - \mathbf{z}_{[,j]})'\zeta(\hat{\boldsymbol{\eta}})\right| \\
&\leq \frac{1}{\lambda_n}\left\|(\mathbf{z}_{[,i]} - \mathbf{z}_{[,j]})\right\| \left\|\zeta(\hat{\boldsymbol{\eta}})\right\| \\
&\leq \frac{1}{\lambda_n}\sqrt{2\{(n-1) - (n-1)r_{ij}\}}\sqrt{n(1+d_n)^2}
\end{aligned}$$

for any $\hat{\beta}_i \neq 0$ and $\hat{\beta}_j \neq 0$. $\square$

A3.3.1 High-dimensional data: additional regularity conditions, conditional oracle property of SJS-CoxBAR and proof

Let x_j be the observed event time for subject j. Assume no ties are present, $\mathbf{z}$ is time independent, and that $q_n = q$ is fixed. Define

$$M_j = \sum_{i=1}^{n}\int_0^{x_j}\left\{\mathbf{z}_i - \frac{\sum_{j=1}^{n} Y_j(s)\mathbf{z}_j \exp(\mathbf{z}'_j\boldsymbol{\beta})}{\sum_{j=1}^{n} Y_j(s)\exp(\mathbf{z}'_j\boldsymbol{\beta})}\right\} dM_i$$

Note that $E(M_j|\mathcal{F}_{j-1}) = M_{j-1}$ or equivalently $E(M_j - M_{j-1}|\mathcal{F}_{j-1}) = 0$. If $b_j = M_j - M_{j-1}$, then $\{b_j\}_{j=1,2,\ldots}$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$, so b_j is bounded almost surely and $E(b_j|\mathcal{F}_{j-1}) = 0$ a.s. for $j = 1, 2, \ldots$

Assume that the screening procedure retains m out of p_n covariates such that $q < m < p_n$. Let s denote an arbitrary subset of $\{1, \ldots, p_n\}$ which defines the submodel with covariates $\mathbf{z}_s = \{z_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. Let s_0 denote the true model. For instance, in our construction of $\boldsymbol{\beta}_0$, $s_0 = \{1, 2, \ldots, q\}$ with $|||s_0||_0 = q$, where $||a||_0$ is the

cardinality of a.

(C8) There exist $w_1, w_2 > 0$ and some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$ with $\min_{1 \le j \le q} |\beta_{0j}| \ge w_1 n^{-\tau_1}$ and $q < m \le w_2 n^{\tau_2}$;

(C9) $\log p_n = O(n^\kappa)$ for some $0 \le \kappa < 1 - 2(\tau_1 + \tau_2)$;

(C10) There exists constants $c_1 > 0, \delta_1 > 0$ such that for sufficiently large n, $\text{eigen}_{\min}[H_n(\boldsymbol{\beta}_0)] \ge c_1$ for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : ||\boldsymbol{\beta}_s - \boldsymbol{\beta}_{0s}||_2 \le \delta_1\}$ and $s \in S_+^{2m}$, where $S_+^{2m} = \{s : s_0 \subset s; ||s||_0 \le 2m\}$;

(C11) There exists constants $C_3, C_4 > 0$ such that $\max_{ij} |z_{ij}| < C_3$ and $\max_i |\mathbf{z}_i \boldsymbol{\beta}_0| < C_4$.

(C12) There exists nonnegative constants γ_j such that for every real number t,

$$E\{\exp(tb_j)|\mathcal{F}_{j-1}\} \le \exp(\gamma_j^2 t^2/2),$$

almost surely for $j = 1, 2, \ldots, n$. Further, for each j, define $\eta(b_j) = \min_j(\gamma_j)$. Now $|b_j| \le K_j$ almost surely for $j = 1, \ldots, n$ and $E\{b_{j_1}, b_{j_2}, \ldots b_{j_k}\} = 0$ for $b_{j_1} < b_{j_2} < \ldots < b_{j_k}$, $k = 1, 2, \ldots$.

Arguably some of the conditions in Yang et al. (2016) are much stronger than some conditions (C1) - (C7). For example, Condition (C10) requires the covariates and $\boldsymbol{\beta}_0$ to be bounded by a constant. This is much stronger than our Condition (C7) where we allow $\boldsymbol{\beta}_0$ to diverge to infinity. The authors also included the Lindeberg condition as a regularity condition in their paper, which we note is unnecessary due to the bounded covariate assumption in Condition (C10) [c.f. Section 4 in Andersen and Gill (1982)].

Theorem 3.3 *Let $\ln(p_n) = O(n^\kappa)$, $0 \le \kappa < 1$. Suppose that Conditions (C1) - (C12). Let $s_0 = \{j : \beta_{0j} \ne 0\}$ be the set of indices where $\beta_{0j} \ne 0$, the true signal in the model and $\hat{s}$ be the set of indices obtained after the sure joint screening procedure of Yang et al. (2016). Define $l_{\hat{s}}(\boldsymbol{\beta})$ as the log-partial likelihood of the model corresponding to $\hat{s}$. Then*

(a) (Sure screening property) $\Pr(s_0 \subset \hat{s}) \to 1$;

(b) *(Oracle Property) Conditional on $s_0 \subset \hat{s}$, with probability tending to one, the CoxBAR estimator $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2'\right)'$ exists and is unique with $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_1$ being the unique fixed point of $f(\boldsymbol{\beta}_1)$, where $f(\boldsymbol{\beta}_1)$ is a solution to $\dot{Q}_{\hat{s}1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = \mathbf{0}$ for*

$$Q_{\hat{s}1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = -2l_{\hat{s}1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) + \lambda_n \boldsymbol{\theta}_1' D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1,$$

with $D_1(\boldsymbol{\beta}_1) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_{q_n}^{-2})$ and $l_{\hat{s}1}(\boldsymbol{\theta}_1)$ being the first q_n components of $l_{\hat{s}}(\boldsymbol{\theta})$, and furthermore

$$\sqrt{n}\mathbf{b}_n' \Sigma_{11}^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0,1),$$

where $\mathbf{b}_n$ is a q_n-dimensional vector such that $||\mathbf{b}_n||_2 \leq 1$, and Σ is defined in Condition (C4).

Proof: Part (a) is a direct consequence of Theorem 2 of Yang et al. (2016) and part (b) is a consequence of part (a) and Theorem 3.1. □

A3.4 The CLG algorithm for Cox ridge regression as explained in Section 3.1.2

The CLG algorithm involves finding $\beta_j^{(new)}$, the value of the jth entry of $\boldsymbol{\beta}$, that minimizes the negative penalized log-partial likelihood, $-l_p(\boldsymbol{\beta})$, assuming that the other values of β_j's are held constant at their current values. For a Cox ridge regression with tuning parameters ϕ_j for $j = 1, \ldots, p_n$, finding $\beta_j^{(new)}$ is equivalent to finding the v that minimizes,

$$g(v) = -v\sum_{i=1}^{n}\delta_i z_{ij} + \sum_{i=1}^{n}\delta_i \ln\left\{\sum_{y\in R_i}\exp\left(\sum_{k=1,k\neq j}^{p_n}\beta_k z_{yk} + v z_{yj}\right)\right\} + \frac{v^2\phi_j}{2},$$

where $R_i = \{y : X_y > X_i\}$ is the risk set for observation i. Here we allow each parameter to be penalized differently. For the BAR algorithm, $\phi_j = \xi_n$ and $\phi_j = \lambda_n/(\hat{\beta}_j^{(k-1)})^2$ in Equations (3.3) and (3.4), respectively. An optimization procedure needs to be used since there is no closed form solution. Using a Taylor series approximation at the current β_j, one

can approximate $g(\cdot)$ through

$$g(v) \approx g(\beta_j) + g'(\beta_j)(v - \beta_j) + \frac{1}{2} g''(\beta_j)(v - \beta_j)^2, \tag{3.55}$$

where

$$g'(\beta_j) = \frac{d}{dv} g(v)\bigg|_{v=\beta_j} = -\sum_{i=1}^{n} z_{ij}\delta_i + \sum_{i=1}^{n} \delta_i \frac{\sum_{y \in R_i} z_{yj} \exp(\mathbf{z}_y' \boldsymbol{\beta})}{\sum_{y \in R_i} \exp(\mathbf{z}_y' \boldsymbol{\beta})} + \beta_j \phi_j, \tag{3.56}$$

and

$$\begin{aligned} g''(\beta_j) = \frac{d^2}{dv^2} g(z)\bigg|_{v=\beta_j} &= \sum_{i=1}^{n} \delta_i \frac{\sum_{y \in R_i} z_{yj}^2 \exp(\mathbf{z}_y' \boldsymbol{\beta})}{\sum_{y \in R_i} \exp(\mathbf{z}_y' \boldsymbol{\beta})} \\ &\quad - \left(\sum_{i=1}^{n} \delta_i \frac{\sum_{y \in R_i} z_{yj} \exp(\mathbf{z}_y' \boldsymbol{\beta})}{\sum_{y \in R_i} \exp(\mathbf{z}_y' \boldsymbol{\beta})} \right)^2 + \phi_j. \end{aligned} \tag{3.57}$$

Consequently, the Taylor series approximation in Equation (3.55) has its minimum at

$$\beta_j^{(new)} = \beta_j + \Delta\beta_j = \beta_j - \frac{g'(\beta_j)}{g''(\beta_j)}.$$

It is worth noting that as $\phi_j \to \infty$, $g'(\beta_j)/g''(\beta_j) \to \beta_j$ and thus $\beta_j^{(new)} \to 0$, which is an important feature of our BAR algorithm as discussed in Remark 3.1. Furthermore, the above algorithm of Mittal et al. (2014) adopts multiple aspects of the work by Zhang and Oles (2001) and Genkin et al. (2007). For CLG, a trust region approach is implemented so that $|\Delta\beta_j|$ is not allowed to be too large on a single iteration. This prevents large updates in regions where a quadratic is a poor approximation to the objective. Second, rather than iteratively updating $\beta_j^{(new)} = \beta_j + \Delta\beta_j$ until convergence, CLG does this only once before going on to the next variable. Since the optimal value of $\beta_j^{(new)}$ depends on the current value of the other β_j's, there is little reason to tune each $\beta_j^{(new)}$ with high precision. Instead, we simply want to decrease $-l_p(\boldsymbol{\beta})$ before going on to the next β_j.

A3.5 Additional simulation results for Section 3.2.1

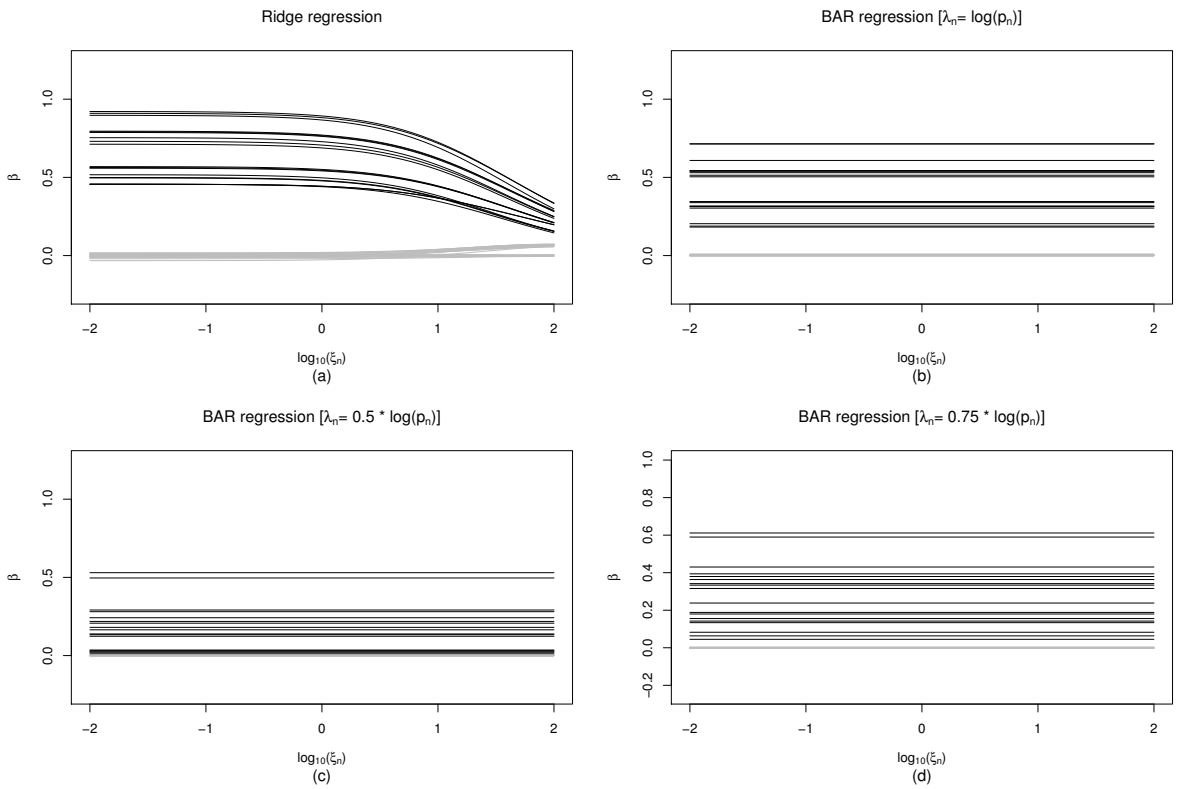

Figure A3.1: Path plot for BAR regression with varying ξ_n and: (b) $\lambda_n = \log(p_n)$, (c) $\lambda_n = 0.5\log(p_n)$, and (d) $\lambda_n = 0.75\log(p_n)$ with estimates averaged over 100 Monte Carlo simulations of size $n = 300$, $p_n = 40$, censoring rate $\approx 25\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}_{p_n-30})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80)$. Path plot for ridge regression (d) with varying ξ_n is also included as a comparison.

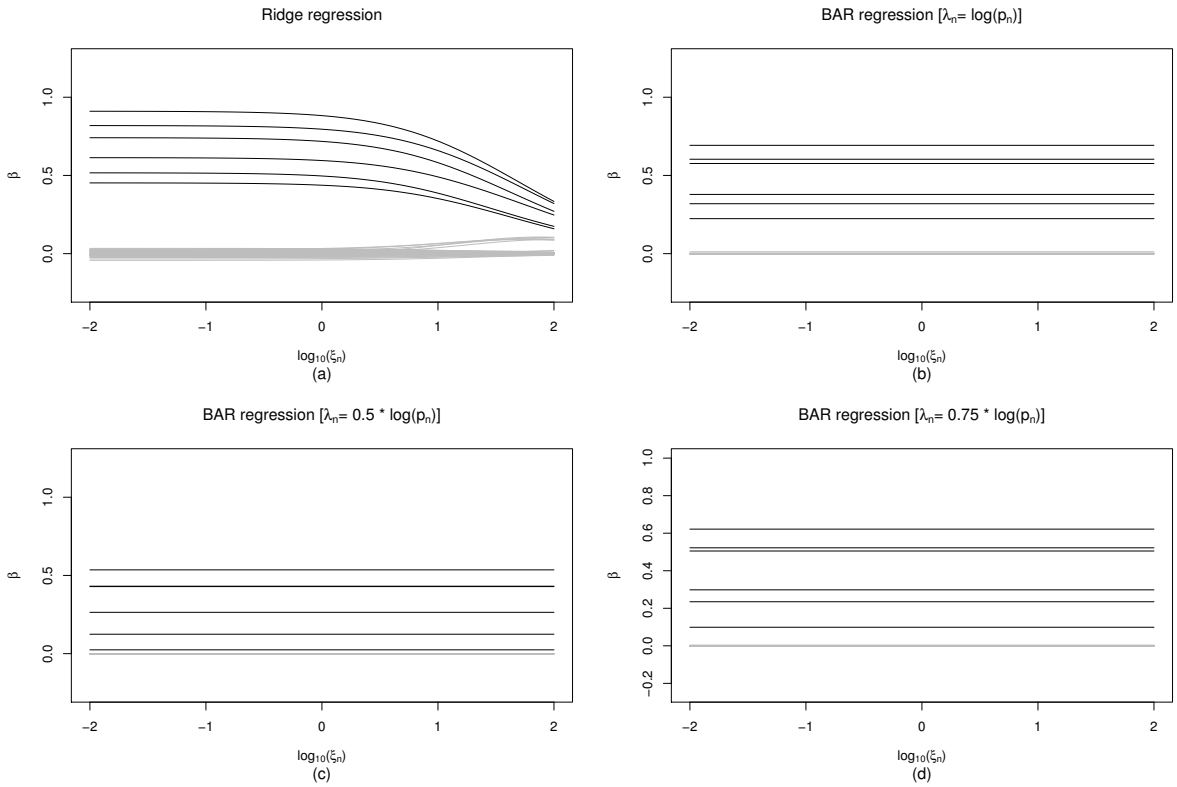

Figure A3.2: Path plot for BAR regression with varying ξ_n and: (b) $\lambda_n = \log(p_n)$, (c) $\lambda_n = 0.5\log(p_n)$, and (d) $\lambda_n = 0.75\log(p_n)$ with estimates averaged over 100 Monte Carlo simulations of size $n = 300$, $p_n = 40$, censoring rate $\approx 60\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \mathbf{0}_{p_n-10})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80)$. Path plot for ridge regression (d) with varying ξ_n is also included as a comparison.

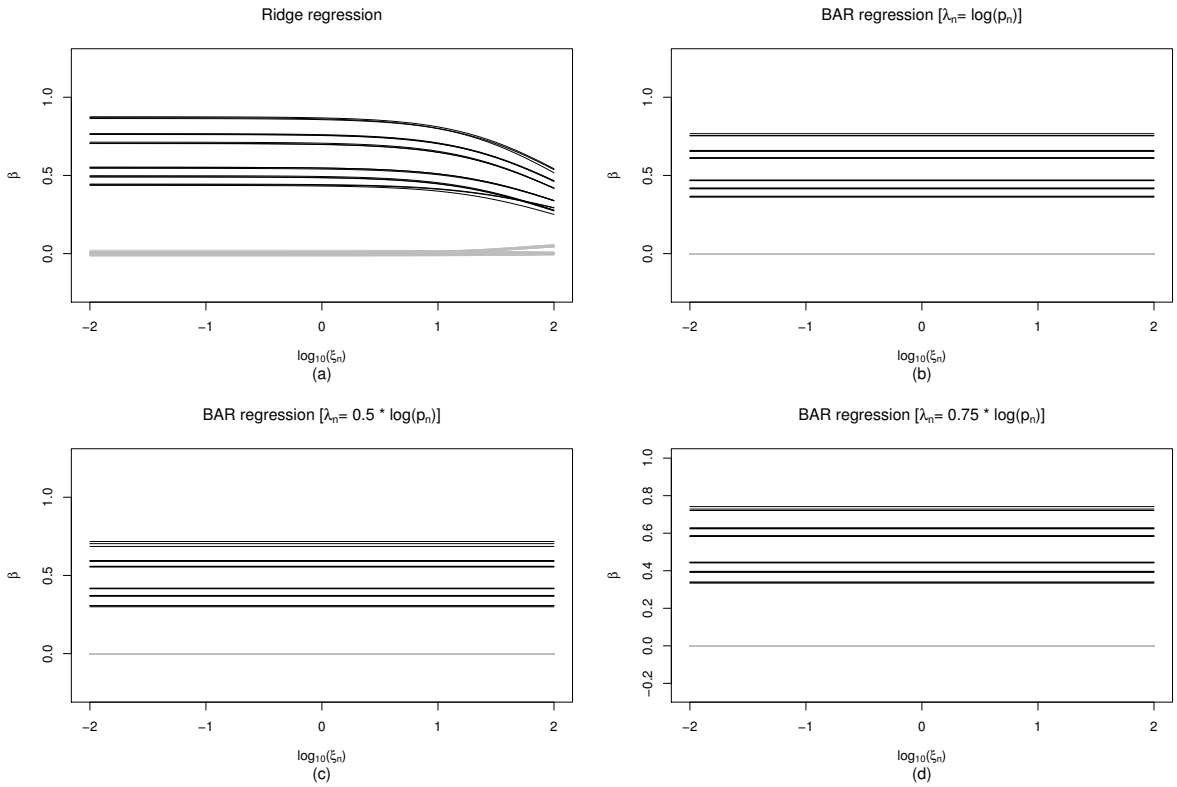

Figure A3.3: Path plot for BAR regression with varying ξ_n and: (b) $\lambda_n = \log(p_n)$, (c) $\lambda_n = 0.5\log(p_n)$, and (d) $\lambda_n = 0.75\log(p_n)$ with estimates averaged over 100 Monte Carlo simulations of size $n = 1000$, $p_n = 100$, censoring rate $\approx 25\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}_{p_n-30})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80)$. Path plot for ridge regression (d) with varying ξ_n is also included as a comparison.

A3.6 Additional simulation results for Section 3.2.2

Table A3.1: Simulated estimation and variable selection performance of BAR, LASSO, SCAD, ALASSO, and MCP where the BIC criterion was used to select the tuning parameters via a grid search. (MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = similarity measure; BIC = average BIC score; Each entry is based on 100 Monte Carlo samples of size $n = 300$, $p_n = 40$, censoring rate $\approx 25\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}_{p_n-30})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80))$

	MSB	FN	FP	SM	BIC
BAR	0.27	0.11	0.74	0.98	1600.72
LASSO	0.46	0.00	9.13	0.82	1608.98
SCAD	0.34	0.07	1.94	0.95	1601.59
ALASSO	0.30	0.06	2.36	0.94	1603.78
MCP	0.32	0.10	1.04	0.97	1600.12

Table A3.2: Simulated estimation and variable selection performance of BAR, LASSO, SCAD, ALASSO, and MCP where the BIC criterion was used to select the tuning parameters via a grid search. (MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = similarity measure; BIC = average BIC score; Each entry is based on 100 Monte Carlo samples of size $n = 300$, $p_n = 40$, censoring rate $\approx 60\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \mathbf{0}_{p_n-10})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80))$

	MSB	FN	FP	SM	BIC
BAR	0.15	0.14	0.67	0.94	1047.77
LASSO	0.31	0.02	3.21	0.82	1063.76
SCAD	0.27	0.11	2.22	0.86	1052.36
ALASSO	0.20	0.16	1.24	0.91	1053.77
MCP	0.23	0.23	1.14	0.91	1050.78

Table A3.3: Simulated estimation and variable selection performance of BAR, LASSO, SCAD, ALASSO, and MCP where the BIC criterion was used to select the tuning parameters via a grid search. (MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = similarity measure; BIC = average BIC score; Each entry is based on 100 Monte Carlo samples of size $n = 1000$, $p_n = 100$, censoring rate $\approx 25\%$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}_{p_n-30})$ where $\boldsymbol{\beta}^* = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80))$

	MSB	FN	FP	SM	BIC
BAR	0.06	0.00	0.71	0.98	6733.09
LASSO	0.43	0.00	11.98	0.79	6801.82
SCAD	0.06	0.00	0.50	0.99	6739.34
ALASSO	0.09	0.00	1.34	0.97	6744.73
MCP	0.05	0.00	0.26	0.99	6737.06

CHAPTER 4

Broken adaptive ridge for the Fine-Gray proportional subdistribution hazards model with applications to large-scale competing risks data

This chapter develops the broken adaptive ridge estimator for the Fine-Gray proportional subdistribution hazards (PSH) model for competing risks data with applications to large-scale competing risks data. The rest of this chapter is organized as follows. In Section 4.1.1, we review the mathematical formulation of competing risks data and the Fine and Gray (1999) proportional subdistribution hazards model. Section 4.1.2, introduces the BAR estimator for the PSH model and state its large-sample properties. Section 4.1.3 derives the cyclic coordinate-wise BAR algorithm. The forward-backward scan method for the PSH model is described in Section 4.1.4. Section 4.2 presents some simulation studies to demonstrate the selection and estimation performance and the computational efficiency gains of our proposed method. A proof-of-concept real data example for fitting large competing risks data is provided in Section 4.3 using a subset of the United States Renal Data System (USRDS). Concluding remarks are given in Section 4.4. An R package for BAR is available at `https:github.com/erickawaguchi/pshBAR`.

4.1 Methodology

4.1.1 Preliminaries: Competing risks data, model, and parameter estimation for fixed model dimension

Competing risks time-to-event data arises frequently in clinical trials, reliability testing, social science, and many other fields (Prentice et al., 1978; Pintilie, 2006; Putter et al., 2007). Competing risks occur when individuals are susceptible to more than one types of possibly correlated events or causes and the occurrence of one event precludes the others from happening. For example, one may wish to study time until first kidney transplant for kidney dialysis patients with end-stage renal disease. Then terminating events such as death, renal function recovery, or discontinuation of dialysis are competing risks as their occurrence will prevent subjects from receiving a transplant. For $i = 1, \ldots, n$, let T_i, C_i, ϵ_i, and $\mathbf{z}_i$ be the event time, possible right-censoring time, cause (event type), and a p_n-dimensional vector of time-independent covariates, respectively, for subject i. Without loss of generality assume there are two event types $\epsilon \in \{1, 2\}$ where $\epsilon = 1$ is the event of interest and $\epsilon = 2$ is the competing risk. With the presence of right-censoring, we generally observe $X_i = T_i \wedge C_i$, $\delta_i = I(T_i \leq C_i)$, where $a \wedge b = \min(a, b)$ and $I(\cdot)$ is the indicator function. Competing risks data consists of the following independent and identically distributed quadruplets $\{(X_i, \delta_i, \delta_i\epsilon_i, \mathbf{z}_i\}_{i=1}^n$. Assume that there exists a τ such that (1) $t \in [0, \tau]$; (2) $\Pr(T_i > \tau) > 0$ and $\Pr(C_i > \tau) > 0$ for all $i = 1, \ldots, n$.

An important quantity for competing risks data is the cumulative incidence function (CIF), which describes the probability of failing from a certain cause of interest before the other causes. The CIF for cause 1 events conditional on the covariates is defined as $F_1(t; \mathbf{z}) = \Pr(T \leq t, \epsilon = 1|\mathbf{z})$. To model the covariate effects on $F_1(t; \mathbf{z})$, Fine and Gray (1999) introduced the now well- appreciated proportional subdistribution hazards (PSH) model:

$$h_1(t|\mathbf{z}) = h_{10}(t) \exp(\mathbf{z}'\boldsymbol{\beta}), \tag{4.1}$$

where

$$h_1(t|\mathbf{z}) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T \leq t + \Delta t, \epsilon = 1 | T \geq t \cup (T \leq t \cap \epsilon \neq 1), \mathbf{z}\}}{\Delta t} = -\frac{d}{dt} \log\{1 - F_1(t; \mathbf{z})\}$$

is a subdistribution hazard (Gray, 1988), $h_{10}(t)$ is a completely unspecified baseline subdistribution hazard, and $\boldsymbol{\beta}$ is a $p_n \times 1$ vector of regression coefficients. As Fine and Gray (1999) mentioned, the risk set associated with $h_1(t; \mathbf{z})$ is somewhat counterfactual as it includes subjects who are still at risk ($T \geq t$) and those who have already observed the competing risk prior to time t ($T \leq t \cap \epsilon \neq 1$). However, this construction is useful for direct modeling of the CIF.

Assume that p_n is fixed. Fine and Gray (1999) developed a parameter estimation method and large sample inference for the PSH model based on the following log-pseudo likelihood:

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left(\mathbf{z}_i' \boldsymbol{\beta} - \log \left\{ \sum_j \hat{w}_j(s) Y_j(s) \exp(\mathbf{z}_j' \boldsymbol{\beta}) \right\} \right) \times \hat{w}_i(s) dN_i(s), \tag{4.2}$$

where $N_i(t) = I(T_i \leq t, \epsilon_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, $\hat{w}_i(t)$ is a time-dependent weight for subject i at time t defined as $\hat{w}_i(t) = I(C_i \geq T_i \wedge t)\hat{G}(t)/\hat{G}(X_i \wedge t)$, and $\hat{G}(t)$ is the Kaplan-Meier estimate for $G(t) = \Pr(C \geq t)$, the survival function of the censoring variable C. Note that, for any subject i and time t, $\hat{w}_i(t)Y_i(t) = 0$ if an individual is right censored or has experienced the event of interest; and $\hat{w}_i(t)Y_i(t) = 1$ if $t < X_i$, and $\hat{w}_i(t)Y_i(t) = \hat{G}(t)/\hat{G}(X_i)$ for events due to the competing risk. Define

$$\hat{\boldsymbol{\beta}}_{mple} = \arg\min_{\boldsymbol{\beta}} \{-l_n(\boldsymbol{\beta})\}.$$

Fine and Gray (1999) showed that $\hat{\boldsymbol{\beta}}_{mple}$ is consistent and asymptotically normal with a sandwich-type variance estimator.

4.1.2 Broken adaptive ridge estimation for the proportional subdistribution hazards model under diverging model dimension

Penalized regression is useful for simultaneous variable selection and parameter estimation. Below we extend the broken adaptive ridge (BAR) estimator to the proportional subdistribution hazards model and establish its large sample properties when the model dimension p_n is allowed to diverge with n.

Let $l_n(\boldsymbol{\beta})$ be the log-pseudo likelihood defined by (4.2). The BAR estimator of $\boldsymbol{\beta}$ starts with an initial ℓ_2-penalized (or ridge) estimator

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}} \{-2l_n(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p} \beta_j^2\}, \tag{4.3}$$

which is updated iteratively by a reweighted ℓ_2-penalized estimator

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} \frac{\beta_j^2}{|\hat{\beta}_j^{(k-1)}|^2} \right\}, \quad k \geq 1, \tag{4.4}$$

where ξ_n and λ_n are non-negative penalization tuning parameters. The BAR estimator of $\boldsymbol{\beta}$ is defined as the limit of this iterative algorithm:

$$\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}, \tag{4.5}$$

which can be viewed as a surrogate to ℓ_0-penalized regression.

Note that adaptively reweighting the penalty of a coefficient by the inverse of its squared estimate from the previous iteration allows each coefficient to be penalized differently. At each successive iteration, coefficients whose true values are zero will have larger penalties that will shrink the estimate further towards zero. On the other hand, at each iteration of the BAR algorithm the solution will be non-sparse. In Theorem 4.1 below, we will show that with probability tending to 1, the limit of the BAR algorithm defined by (4.5) will be sparse, selection consistent, and has both an oracle and grouping property for estimation.

Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be the first q_n and remaining $p_n - q_n$ components of $\boldsymbol{\beta}$, respectively, and

define $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{02})'$ as the true values of $\boldsymbol{\beta}$ where, without loss of generality, $\boldsymbol{\beta}_{01} = (\beta_{01} \ldots, \beta_{0q_n})$ is a vector of q_n non-zero values and $\boldsymbol{\beta}_{02} = \mathbf{0}$ is a $p_n - q_n$ dimensional vector of zeros. Further technical assumptions for $\boldsymbol{\beta}_0$ and p_n are given later in condition (C6) in Appendix A4.1. Below we state the asymptotic properties of the BAR estimator for the PSH model under certain regularity conditions.

Theorem 4.1 (Oracle property) *Assume the regularity conditions (C1) - (C6) in Appendix A4.1 hold. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the first q_n and the remaining $p_n - q_n$ components of the BAR estimator $\hat{\boldsymbol{\beta}}$, respectively. Then,*

(a) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to one;

(b) $\sqrt{n}\mathbf{d}'_n \Gamma_{11}^{-1/2} \Omega_{11}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \to N(0,1)$, for any q_n-dimensional vector $\mathbf{d}_n$ such that $||\mathbf{d}_n||_2 \leq 1$ and where Γ_{11} and Ω_{11} are the first $q_n \times q_n$ submatrices of Γ and Ω, respectively, defined in Condition (C4).

Theorem 4.1(a) establishes that with large probability, the true zero coefficients will be estimated as zeros by the BAR estimator. Part (b) of the theorem essentially states that the nonzero component of the BAR estimator is asymptotically normal and equivalent to the weighted ridge estimator of the oracle model.

An appealing property of ℓ_2-penalized regression is its tendency to shrink correlated covariates toward each other. As an ℓ_2-based procedure, the BAR method also retains this grouping property for highly-correlated covariates. A proof is provided in Appendix A4.4.

Theorem 4.2 (Grouping property) *Assume that $Z = (\mathbf{z}_i^T, \ldots \mathbf{z}_n^T)$ is standardized. That is, for all $j = 1, \ldots, p_n$, $\sum_{i=1}^n z_{ij} = 0$, $\mathbf{z}_{[,j]}^T \mathbf{z}_{[,j]} = n - 1$, where $\mathbf{z}_{[,j]}$ is the jth column of Z. Suppose the regularity conditions (C1) - (C6) in Appendix A4.1 hold and let $\hat{\boldsymbol{\beta}}$ be the BAR estimator. Then for any $\hat{\beta}_i \neq 0$ and $\hat{\beta}_j \neq 0$,*

$$|\hat{\beta}_i^{-1} - \hat{\beta}_j^{-1}| \leq \frac{1}{\lambda_n} \sqrt{2\{(n-1)(1 - r_{ij})\}} \sqrt{n(1 + e_n)^2}, \tag{4.6}$$

with probability tending to one, where $e_n = \sum_{i=1}^{n} I(\epsilon_i = 1)$*, and* $r_{ij} = \frac{1}{n-1}\mathbf{z}_{[,i]}^T\mathbf{z}_{[,j]}$ *is the sample correlation of* $\mathbf{z}_{[,i]}$ *and* $\mathbf{z}_{[,j]}$*.*

We can see that as $r_{ij} \to 1$, the absolute difference between $\hat{\beta}_i$ and $\hat{\beta}_j$ approaches 0 implying that the estimated coefficients of two highly-correlated variables will be similar in magnitude.

Remark 4.1 *(Penalization parameter selection) The BAR estimator has two penalization parameters:* ξ_n *for the initial ridge estimator and* λ_n *for the subsequent reweighted ridge estimators. We have observed through extensive simulations that the BAR estimator is not sensitive to* ξ_n *and is stable over a huge interval as illustrated in Figures A4.3, A4.4, and A4.5 of Appendix A4.8.1. In Table 4.1 and Tables A4.1, A4.2, and A4.3 of Appendix A4.8.2, the BAR*(λ_n) *estimator with* $\xi_n = log(p_n)$ *and a grid search on* λ_n *showed essentially the same performance as the BAR*(ξ_n, λ_n) *estimator using a two-dimensional grid search on* ξ_n *and* λ_n *based on the BIC criterion.*

Remark 4.2 *(Computational aspects) The BAR estimator can be implemented using the algorithm outlined in Algorithm 2 of Appendix A4.7 in which cyclic coordinate decent (CCD) algorithm is employed to accelerate each reweighted* ℓ_2*-penalized regression. We point out that there are still some key computational aspects that impact the runtime of the algorithm. First, at the highest level (line 2), it runs a sequence* $(k = 0, 1, \ldots)$ *of adaptively reweighted ridge regressions, which seems to add an extra layer of computational complexity as compared to other popular single-step penalization methods such as LASSO and can become a bottleneck when a large number of iterations is needed. Second, because ridge regression is not sparse and thus the limit is never achieved at any given step of the BAR algorithm, an arbitrarily small cutoff value* ϵ^* *has to be used to induce sparsity in Algorithm 2 (line 18), which is an unpleasant feature. These are obviously general issues for the BAR approach and are not limited to the PSH model. Lastly, for the PSH model, direct calculation of the log-pseudo likelihood and its derivatives involves* $O(n^2)$ *operations as explained later in Section 4.1.4, which is highly impactful on the algorithm since it occurs at the lowest level (the innermost loop). This is a common issue also shared by other (penalized or unpenalized) estimation*

methods for the PSH model. Below we derive new algorithms to address the above issues in Sections 4.1.3 and 4.1.4.

4.1.3 A cyclic coordinate-wise BAR algorithm

In this section we derive a fast cyclic coordinate-wise BAR algorithm that will result in the elimination of performing multiple ridge regressions and avoid using a cutoff ϵ^* to introduce sparsity as required by the original BAR algorithm (Algorithm 2). Let $\dot{l}(\boldsymbol{\beta})$ and $\ddot{l}(\boldsymbol{\beta})$ denote the first and second derivative of $l_n(\boldsymbol{\beta})$, respectively. For a consistent estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, consider the Cholesky decomposition $-\ddot{l}(\tilde{\boldsymbol{\beta}}) = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ and define $\tilde{\mathbf{y}} = (\tilde{\mathbf{X}}')^{-1}\{-\ddot{l}(\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}}+\dot{l}(\tilde{\boldsymbol{\beta}})\}$ as the pseudo-response vector. Approximating the negative log-pseudo likelihood by $-l_n(\boldsymbol{\beta}) \approx (1/2)(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})$ using a second-order Taylor expansion in (4.4) leads to the following solution

$$\hat{\boldsymbol{\beta}}^{(k)} = g(\hat{\boldsymbol{\beta}}^{(k-1)}),$$

where

$$g(\boldsymbol{\beta}) = \{\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda_n D(\boldsymbol{\beta})\}^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}.$$

and $D(\boldsymbol{\beta}) = \mathrm{diag}(\beta_1^{-2}, \ldots, \beta_{p_n}^{-2})$. Hence, as $k \to \infty$, the limit of the sequence $\{\hat{\boldsymbol{\beta}}^{(k)}\}$ is the fixed point of the function $g(\cdot)$ or the solution of $g(\boldsymbol{\beta}) = \boldsymbol{\beta}$.

The next theorem shows that each component of the fixed-point solution of g can be expressed as a function of all other components. The proof is deferred to Appendix A4.5.

Theorem 4.3 *Let $\hat{\boldsymbol{\beta}}$ be the fixed-point solution of $g(\cdot)$. Then, for each $j = 1, \ldots, p_n$, the jth component of $\hat{\boldsymbol{\beta}}$ can be expressed as follows*

$$\hat{\beta}_j = g_j(\hat{\boldsymbol{\beta}}_{-j}) \equiv \begin{cases} 0, & \text{if } |b_j| < 2\sqrt{\lambda_n \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j}, \\ \frac{b_j + sign(b_j)\sqrt{(b_j)^2 - 4\lambda_n \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j}}{2\tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j}, & \text{otherwise,} \end{cases} \tag{4.7}$$

where $b_j = \tilde{\mathbf{x}}_j'(\tilde{\mathbf{y}} - \sum_{i\neq j}\tilde{\mathbf{x}}_i\hat{\beta}_i)$ and $\hat{\boldsymbol{\beta}}_{-j} = (\hat{\beta}_1, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_{p_n})'$.

The above result motivates our cyclic coordinate-wise broken adaptive ridge (CYCBAR)

Algorithm 1: The CYCBAR Algorithm

```
Set β^(0) = β̂_ridge;
for s = 1, 2, ... do
    # Enter cyclic coordinate-wise BAR algorithm
    for j = 1, ... p_n do
        Calculate c_1j = −l̇_j(β^(s−1)), c_2j = −l̈_jj(β^(s−1)) and b_j^(s) = c_2j β_j^(s−1) − c_1j;
        if |b_j^(s)| < 2√(c_2j λ_n) then
            β_j^(s) = 0;
        else
            β_j^(s) = (b_j^(s) + sign(b_j^(s)) √((b_j^(s))^2 − 4 c_2j λ_n)) / (2 c_2j);
        end
    end
    if ‖β^(s) − β^(s−1)‖ < tol then
        β̂_BAR = β^(s) and break;
    end
end
```

algorithm which performs cyclic coordinate-wise updates for the fixed point of $g(\cdot)$ using equation (4.7) as outlined in Algorithm 1 below. In Algorithm 1, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ are initially estimated using the initial ridge estimate $\boldsymbol{\beta}^{(0)}$ and then subsequently updated at step s using the previous estimate $\boldsymbol{\beta}^{(s-1)}$ for $s \geq 1$. Consequently, at step s, we have

$$b_j^{(s)} \equiv \tilde{\mathbf{x}}_j' \left\{ \tilde{\mathbf{y}} - \sum_{i \neq j} \tilde{\mathbf{x}}_i \beta_i^{(s-1)} \right\} = -\ddot{l}_{jj}(\boldsymbol{\beta}^{(s-1)})\beta_j^{(s-1)} + \dot{l}_j(\boldsymbol{\beta}^{(s-1)}), \quad \text{for } j = 1, \ldots, p_n,$$

where $\dot{l}_j(\boldsymbol{\beta})$ is the jth element of $-\dot{l}(\boldsymbol{\beta})$ and $-\ddot{l}_{jj}(\boldsymbol{\beta})$ is the jth diagonal element of the matrix $\ddot{l}(\boldsymbol{\beta})$.

Remark 4.3 *(Convergence of* CYCBAR*) The* CYCBAR *algorithm resembles the well-known cyclic coordinate decent (CCD) algorithm. On the other hand, unlike CCD that decreases an objective function with each coordinate update, the* CYCBAR *algorithm makes coordinate-wise updates for a fixed-point problem without explicitly decreasing an objective function. Hence the numerical convergence of the* CYCBAR *algorithm is guaranteed by a different mechanism. To appreciate how the* CYCBAR *algorithm behaves, in Figures A4.1 and A4.2 of Appendix A4.8.1, we illustrate under a simple scenario with* $p_n = 2$ *that the* CYCBAR *algo-*

rithm converges to the fixed point of $g(\beta_1, \beta_2)$ along the graphs of $\beta_1 = g_1(\beta_2)$ and $\beta_2 = g_2(\beta_1)$, with each coordinate-wise update moving monotonically a step closer to the fixed point. This is further corroborated in our extensive numerical studies. A rigorous proof of the numerical convergence of the CYCBAR *algorithm is however not trivial and needs to be investigated in future research.*

4.1.4 Scalable parameter estimation via forward-backward scan

Within the innermost loop of the CYCBAR algorithm (Algorithm 1), we need to calculate

$$\dot{l}_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) z_{ij} - \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) \frac{\sum_{k \in R_i} z_{kj} \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)},$$
$$\ddot{l}_{jj}(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) \left[\frac{\sum_{k \in R_i} z_{kj}^2 \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)} - \left\{ \frac{\sum_{k \in R_i} z_{kj} \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)} \right\}^2 \right],$$

where

$$\tilde{w}_{ik} = \hat{w}_k(X_i) = \hat{G}(X_i)/\hat{G}(X_i \wedge X_k), \quad k \in R_i,$$

$R_i = \{y : (X_y \geq X_i) \cup (X_y \leq X_i \cap \epsilon_y = 2)\}$ and $\eta_k = \mathbf{z}_k' \boldsymbol{\beta}$. Direct calculations using the above formulas will need $O(n^2)$ operations because of the double summations a dis computationally taxing for large n.

Before going further, we note that for the Cox proportional hazards model with no competing risks, $R_i = \{y : X_y \geq X_i\}$ and $\tilde{w}_{ik} \equiv 1$ for all i and k. Therefore the score function can be written as

$$\dot{l}_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i = 1) z_{ij} - \sum_{i=1}^{n} I(\delta_i = 1) \frac{\sum_{k \in R_i} z_{kj} \exp(\eta_k)}{\sum_{k \in R_i} \exp(\eta_k)}, \tag{4.8}$$

for $j = 1, \ldots, p_n$. Again, if done directly, calculating $\dot{l}_j(\boldsymbol{\beta})$ will require $O(n^2)$ calculations. Suchard et al. (2013) and Mittal et al. (2014), among others, have implemented the following technique to calculate (4.8) in $O(n)$ calculations. Note that if the event times are arranged in decreasing order, both $\sum_{k \in R_i} z_{kj} \exp(\eta_k)$ and $\sum_{k \in R_i} \exp(\eta_k)$ are a series of cumulative sums. For example, given $X_i > X_{i'}$, the set $R_{i'}$ consists of the observations

from R_i and the set of observations $\{y : X_y \in [X_{i'}, X_i)\}$, therefore $\sum_{k\in R_{i'}} z_{kj}\exp(\eta_k) = \sum_{k\in R_i} z_{kj}\exp(\eta_k) + \sum_{k\in\{y:X_y\in[X_{i'},X_i)\}} z_{kj}\exp(\eta_k)$ and calculating both $\sum_{k\in R_i} z_{kj}\exp(\eta_k)$ and $\sum_{k\in R_i}\exp(\eta_k)$, and consequently its ratio, for all $i = 1,\ldots,n$ will only require $O(n)$ calculations in total. Furthermore, the outer summation of subjects who observe the event of interest is also a cumulative sum since, provided that $X_i > X_{i'}$ and both $\delta_i = 1$ and $\delta_{i'} = 1$,

$$\sum_{l=1}^{i} I(\delta_l = 1)\frac{\sum_{k\in R_l} z_{kj}\exp(\eta_k)}{\sum_{k\in R_l}\exp(\eta_k)} = \sum_{l=1}^{i'} I(\delta_l = 1)\frac{\sum_{k\in R_l} z_{kj}\exp(\eta_k)}{\sum_{k\in R_l}\exp(\eta_k)} \tag{4.9}$$

$$+ I(\delta_i = 1)\frac{\sum_{k\in R_i} z_{kj}\exp(\eta_k)}{\sum_{k\in R_i}\exp(\eta_k)}, \tag{4.10}$$

which will also only require $O(n)$ calculations since the ratio can be precomputed in $O(n)$ calculations. The diagonal elements of the Hessian also follow a similar derivation and can be calculated in $O(n)$ calculations.

For the PSH model, however, $\sum_{k\in R_i}\tilde{w}_{ik}\exp(\eta_j)$, $i = 1,\ldots,n$, are not a series of simple cumulative sums because 1) the risk sets R_i are not monotone over time, and 2) for each i, a different set of weights $\tilde{w}_{ik} = \hat{G}(X_i)/\hat{G}(X_i \wedge X_k)$, $k \in R_i$ are required. To overcome this problem, we show in Lemma 4.1 below that $\sum_{k\in R_i}\tilde{w}_{ik}\exp(\eta_j)$ can be decomposed into a forward cumulative sum and a backward cumulative sum over two disjoint monotone sets. A simple proof is provided in Appendix A4.6.

Lemma 4.1 *Assume that no ties are present. Then*

$$\sum_{k\in R_i}\tilde{w}_{ik}\exp(\eta_k) = \sum_{k\in R_i(1)}\exp(\eta_k) + \hat{G}(X_i)\sum_{k\in R_i(2)}\exp(\eta_k)/\hat{G}(X_k) \tag{4.11}$$

where $R_i(1) = \{y : (X_y \geq X_i)\}$ and $R_i(2) = \{y : (X_y < X_i \cap \epsilon_y = 2)\}$ are distinct partitions of R_i. Furthermore, $R_i(1)$ is monotonically decreasing over time and $R_i(2)$ is monotonically increasing over time.

Because $R_i(1)$ grows cumulatively as the event times decrease from the largest to the smallest, whereas $R_i(2)$ grows cumulatively as the observed event times increase from the smallest to the largest since it only involves subjects who observed a competing risk and had

an observed event time smaller than subject i. Thus, similar to the Cox model, the ratio of summations for the score and diagonal Hessian values can be calculated in linear time via a forward-backward scan where one scan goes in one direction to calculate the cumulative sums associated with $R_i(1)$ and the other scan goes in the opposite direction to calculate the cumulative sum associated with $R_i(2)$. Therefore, we can effectively reduce the number of operations from $O(n^2)$ to $O(n)$.

4.2 Simulation study

Two simulation studies are presented in this section. First, in Section 4.2.2, we illustrate the selection and estimation performance of the BAR method for the PSH model along with several popular ℓ_1-based penalization methods: LASSO, adaptive LASSO (ALASSO), smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP). We use the R package **crrp** (Fu et al., 2017) to perform LASSO ALASSO, SCAD, and MCP. Second, in Section 4.2.3 we illustrate the computational efficiency gains for BAR obtained by the CYCBAR algorithm described in Section 4.1.3 and the forward-backward scan algorithm described in Section 4.1.4. All simulations were performed on a system with an Intel Core i5 2.9 GHz processor and 16GB of memory.

4.2.1 Simulation setup

We simulate datasets under various sample sizes and parameter dimensions. The design matrix, $\mathbf{Z}$ was generated from a p_n-dimensional standard normal distribution with mean zero and pairwise correlation $\text{corr}(z_i, z_j) = \rho^{|i-j|}$, where $\rho = 0.5$ simulates moderate correlation. The vector of regression parameters for cause 1, the cause of interest, is $\boldsymbol{\beta}_1 = (0.40, 0.45, 0, 0.50, 0, 0.60, 0.75, 0, 0, 0.80, \mathbf{0}_{p-10})$. The data generation scheme follows a similar design to that of Fine and Gray (1999) and Fu et al. (2017). The CIF for cause 1 is $F_1(t; \mathbf{z}_i) = \Pr(T_i \leq t, \epsilon_i = 1|\mathbf{z}_i) = 1 - [1 - \pi\{1 - \exp(-t)\}]^{\exp(\mathbf{z}_i'\boldsymbol{\beta}_1)}$, which is a unit exponential mixture with mass $1 - \pi$ at ∞ when $\mathbf{z}_i = \mathbf{0}$. Unless otherwise noted, the value of π is set to 0.5, which corresponds to a cause 1 event rate of approximately 41%. The

CIF for cause 2 is obtained by setting $\Pr(\epsilon_i = 2|\mathbf{z}_i) = 1 - \Pr(\epsilon_i = 1|\mathbf{z}_i)$ and then using an exponential distribution with rate $\exp(\mathbf{z}_i'\boldsymbol{\beta}_2)$ for the conditional CIF $\Pr(T_i \leq t|\epsilon_i = 2, \mathbf{z}_i)$ with $\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_1$. Censoring times are independently generated from a uniform distribution $U(0, u_{max})$ where u_{max} controls the censoring percentage. The average censoring percentage for our simulations vary between $30 - 35\%$.

4.2.2 Variable selection and parameter estimation performance

The operating characteristics of BAR with different tuning parameter selection strategies along with LASSO, adaptive LASSO (ALASSO), SCAD and MCP are assessed by the following measures. As a gold standard, we also fit the oracle model (ORACLE) as if the true model was known *a priori*. Estimation bias is summarized through the mean squared bias (MSB), $E\{\sum_{i=1}^{p}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{0i})^2\}$. Variable selection performance is measured by a number of indices: the mean number of false positives (FP), the mean number of false negatives (FN); and average similarity measure (SM) for support recovery where $SM = ||\hat{\mathcal{S}}_1 \cap \mathcal{S}_1||_0 / \sqrt{||\hat{\mathcal{S}}_1||_0 \cdot ||\mathcal{S}_1||_0}$ and $\mathcal{S}_1$ and $\hat{\mathcal{S}}_1$ are the set of indices for the non-zero components of $\boldsymbol{\beta}_1$ and $\hat{\boldsymbol{\beta}}_1$, respectively (Zhang and Cheng, 2017). The similarity measure can be viewed as a continuous measure for true model recovery: it is close to 1 when the estimated model is similar to the true model, and close to 0 when the estimated model is highly dissimilar to the true model.

For BAR, we investigate three tuning parameter selection approaches: 1) ξ_n and λ_n are selected via a two-dimensional grid search to minimize the BIC criterion (BAR(ξ_n, λ_n)); 2) λ_n is selected via grid search to minimize the BIC criterion; and $\xi_n = \log(p_n)$ (BAR(λ_n)); and 3) fixed $\lambda_n = \log(p_n)$ and $\xi_n = \log(p_n)$ (BAR$_{EBIC}$), which corresponds to a local solution to the extended BIC criterion (Chen and Chen, 2008; Gao and Carroll, 2017). The grids for ξ_n and λ_n were chosen from a log-spaced interval of 25 values between $[0.001, 3\log(p_n)]$. Unless otherwise noted, we implement BAR using both the CYCBAR and forward-backward scan outlined in the section above. The tuning parameter for LASSO, ALASSO, SCAD, and MCP is selected by minimizing the BIC-score through a grid search of 25 possible values for λ_n. We only consider the $p_n < n$ scenario and thus use the maximum pseudo likelihood estimator

Table 4.1: Estimation and selection performance of BAR along with LASSO, ALASSO, SCAD, and MCP. (BAR(ξ_n, λ_n): ξ_n and λ_n are found using a two-dimensional grid search; BAR(λ_n): $\xi_n = \log(p_n)$ and λ_n is found through a grid search; BAR$_{EBIC}$: $\xi_n = \log(p_n)$ and $\lambda_n = \log(p_n)$; MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = average similarity measure; Censoring rate $\approx 30\%$; $p_n = 40; q_n = 6$. The BIC criterion is used for tuning parameter selection for all methods except for BAR$_{EBIC}$. Each entry is based on 100 Monte Carlo samples.)

	$n = 300$				$n = 700$			
Method	MSB	FN	FP	SM	MSB	FN	FP	SM
ORACLE	0.09	0.00	0.00	1.00	0.04	0.00	0.00	1.00
BAR(ξ_n, λ_n)	0.20	0.26	0.69	0.93	0.05	0.00	0.39	0.97
BAR(λ_n)	0.20	0.29	0.64	0.93	0.05	0.00	0.41	0.97
BAR$_{EBIC}$	0.40	1.29	0.00	0.88	0.06	0.06	0.00	0.99
LASSO	0.31	0.02	3.09	0.82	0.15	0.00	2.69	0.84
ALASSO	0.23	0.22	1.05	0.91	0.06	0.00	0.61	0.96
SCAD	0.32	0.20	1.70	0.88	0.08	0.01	0.95	0.94
MCP	0.28	0.41	0.96	0.90	0.06	0.05	0.42	0.97

as the initial estimator for ALASSO. Tables 4.1 summarizes some results for $p_n = 40$ and two sample sizes $n = 300$, 700.

First, we observe from Table 4.1 that for BAR, the selection and estimation performances between optimizing over both ξ_n and λ_n (BAR(ξ_n, λ_n)) and over only λ_n (BAR(λ_n)) are similar, suggesting that the BAR estimator is insensitive over the choice of ξ_n. This is corroborated by further simulations in Appendix A4.8.1 (Figures A4.3, A4.4, and A4.5) where the solution path of the BAR estimator with various choices of λ_n are stable over a large interval of ξ_n. Secondly, BAR(ξ_n, λ_n), BAR(λ_n), and MCP are generally the top performers. Third, for the smaller sample size $n = 300$, BAR$_{EBIC}$ can sometimes be overly aggressive in misclassifying small effects as null, thus showing a tradeoff between a slightly higher false zeros (FN) and a lower false non-zeros (FP). However when the sample size gets sufficiently large ($n = 700$), BAR$_{EBIC}$ performs as well as or better than other methods with respect to all considered measures. This suggests that for large sample data, computational savings can be achieved using BAR$_{EBIC}$ by prefixing both ξ_n and λ_n at $\log(p_n)$ and thus avoiding costly data-driven tuning parameter selection. Lastly, we have conducted more extensive simulations with different combinations of model dimension, event rate, signal

value, sample sizes, and model sparsity, which yielded consistent findings. Some further results are reported in Appendix A4.8.2.

4.2.3 Computational efficiencies

In this simulation we illustrate the computational savings obtained from CYCBAR and the forward-backward scan described in Sections 4.1.3 and 4.1.4. We compare three implementations of BAR for the PSH model: BAR(λ_n) with both the CYCBAR algorithm and forward-backward scan, BAR(λ_n) with the CYCBAR algorithm and without the forward-backward scan, and BAR(λ_n) without either. We let n vary from 600 to 2000, $p_n = 100$, and $\rho = 0.5$ and compute the runtime of each method averaged over 100 simulations. We report timing on a system with an Intel Core i5 2.9 GHz processor and 16GB of memory.

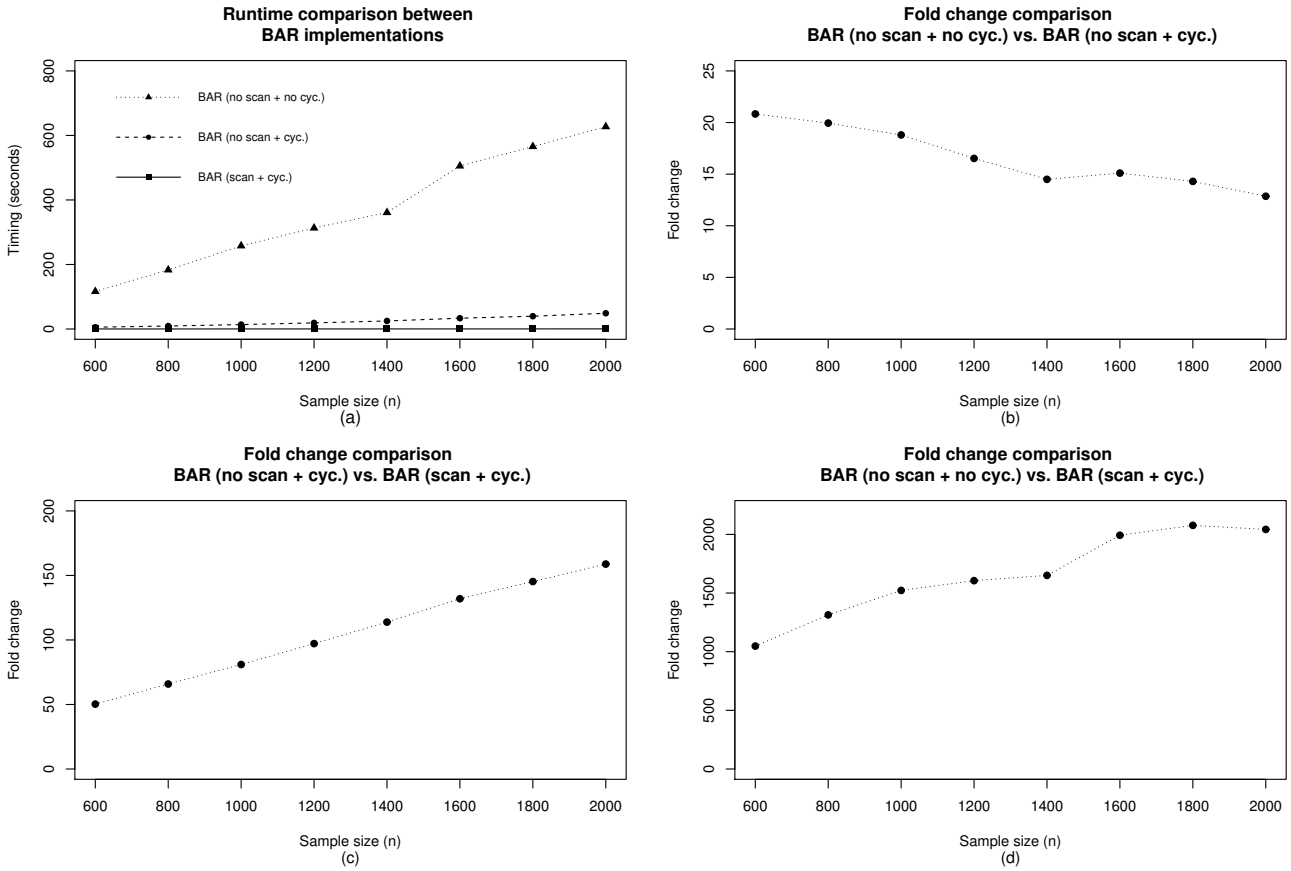

Figure 4.1: Runtime comparison between three BAR(λ_n) implementations (cyc. = CYCBAR described in Section 4.1.3; scan = forward-backward scan described in Section 4.1.4).

Figure 4.1(a) displays the mean runtime (in seconds) for each method as the sample size increases, which shows that the runtime of the original BAR(λ_n) increases quickly while the runtime of BAR(λ_n) with both CYCBAR and forward-backward scan grows at a much lower rate. Panels (b) and (c) further demonstrate the separate contributions of CYCBAR and the forward-backward scan method, respectively, using fold change. Panel (b) shows a 15-20 fold decrease in runtime between CYCBAR and the original BAR. Panel (c) shows the benefit of linearized estimation, with a 50-150 fold decrease in runtime between CYCBAR with and without the forward-backward scan. Panel (d) illustrates that using both CYCBAR and the forward-backward scan results in a multiplicative gain, yielding a 1,000-2,000 fold speedup in runtime. Finally, the runtime reduction is expected to increase as n and/or p_n grow larger as illustrated by the real data example in the following section.

4.3 End-stage renal disease

The United States Renal Data System (USRDS) is a national data system that collects information about end-stage renal disease in the United States. Patients with end-stage renal disease are known to have a shorter life expectancy compared to their disease-free peers (USRDS Annual Report 2017) and kidney transplantation has been shown to provide better health outcomes for patients with end-stage renal disease (Wolfe et al., 1999; Purnell et al., 2016). As an illustration of the scalability of various methods for large data, we run penalized regressions for a PSH model with 63 demographic and clinical variables using a subset of $n = 225,000$ patients from the USRDS that spans a 10-year study time between January 2005 to June 2015. The event of interest was first kidney transplant for patients who were currently on dialysis. Death, renal function recovery, and discontinuation of dialysis are competing risks. Subjects who are lost to follow up or had no event by the end of study period are considered as right censored. We randomly split the data into a training set ($n = 125,000$) and test set ($n = 100,000$). Table A4.4 in the Appendix A4.8.2 shows that the proportions of each event type are similar across the training and test sets.

The BAR method along with SCAD and MCP are used to fit the PSH model using the

training set. Similar to Section 4.2.3, we consider BAR(λ_n) with three implementations: without CYCBAR and the forward-backward scan; with CYCBAR and without the forward-backward scan; and with both CYCBAR and the forward-backward scan. In addition, we consider BAR$_{EBIC}$ as defined earlier in Table 4.1 with both CYCBAR and the forward-backward scan. SCAD and MCP were performed using the **crrp** R package (Fu et al., 2017) where its variance estimation component is removed to allow a fair comparison of their runtime with BAR only for parameter estimation. To assess the predictive performance of the selected models we calculate the concordance index (c-index) proposed by Wolbers et al. (2009) using the test set. Table 4.2 summarizes the computational time (in hours), the c-index, and the number of selected variables for each method.

We observe from Table 4.2 that CYCBAR took 36 hours to finish, a marked reduction in runtime over the original original BAR implementation which did not finish after 96 hours and was terminated. More impressively, BAR(λ_n) using the combination of CYCBAR and the forward-backward scan finished in 0.009 hours or 32 seconds, an enormous boost in speeding up the computation. Furthermore, using a predetermined tuning parameter, BAR$_{EBIC}$ using the combination of CYCBAR and the forward-backward scan finished in 0.002 hours or 7 seconds. Lastly, the current existing methods MCP and SCAD using **crrp** both took over a day to finish. Therefore, when using the combination of CYCBAR and the forward-backward scan, the speedups of BAR(λ_n) and BAR$_{EBIC}$ over the original BAR and other existing methods are enormous. For example, the speedups of BAR(λ_n) (32 seconds) and BAR$_{EBIC}$ (7 seconds) over MCP (28.6 hours) are 3,177 and 14,300 folds, respectively.

The predictive and selection performances of all methods are comparable with similar c index and model size (number of selected variables), which can be attributed to the massive sample size of both the training and test set and the fact that the considered methods are selection consistent and have oracle properties. Many of the selected variables by all four methods such as racial differences (Kasiske et al., 1991; Purnell et al., 2016, 2018), insurance type (Keith et al., 2008; Schold et al., 2011), neighborhood poverty (Patzer et al., 2009), and smoking (Stack et al., 2016) have been previously reported to have an impact on kidney transplantation.

Table 4.2: Analysis results of a USRDS data using BAR with four different implementations along with MCP and SCAD (BAR(λ_n): $\xi_n = \log(p_n)$ and λ_n selected through a grid search; BAR$_{EBIC}$: $\xi_n = \lambda_n = \log(p_n)$; cyc. = CYCBAR; lin. = forward-backward scan; Except for BAR$_{EBIC}$, BIC was used to select tuning parameters; *The original BAR(λ_n) without CYCBAR and forward-backward scan did not finish after 96 hours.)

	BAR(λ_n) (no cyc+no scan)	BAR(λ_n) (cyc+no scan)	BAR(λ_n) (cyc+scan)	BAR$_{EBIC}$ (cyc+scan)	MCP	SCAD
c-index	–	0.85	0.85	0.85	0.85	0.85
Model size	96*	42	42	40	48	49

4.4 Discussion

We have extended the BAR method for simultaneous parameter estimation and variable selection to the Fine and Gray (1999) PSH model for competing risks data and developed its large-sample properties. More importantly, to make the BAR method scalable to large-scale competing risks data, we have further developed 1) a novel coordinate-wise update (CYCBAR) algorithm to avoid carrying out multiple ridge regressions in the original BAR implementation and 2) a forward-backward scan algorithm to reduce the computational cost of the log-likelihood and its derivatives for the PSH model from the order of $O(n^2)$ to $O(n)$. While showing comparable selection and estimation performance, the BAR method for the PSH model using the two new algorithms has produced thousands to tens of thousands fold speedups over some current penalization methods for the PSH model in numerical studies. Furthermore, when the sample size is very large and the true model is sparse, further computational savings can be achieved by using BAR with the pre-determined $\lambda_n = \log(p_n)$ according to an extend BIC criterion to avoid costly data-driven tuning parameter selection.

An important domain of application of the developed scalable sparse regression method is large comparative effectiveness and drug safety studies using massive electronic health record (EHR) databases such as the Observational Health Data Sciences and Informatics (OHDSI) program (Hripcsak et al., 2015) (*https://ohdsi.org/*) and the U.S. FDA's Sentinel Initiative (*https://www.fda.gov/safety/fdassentinelinitiative/ucm2007250.htm*). These massive databases typically contain millions to hundreds of millions patient records with tens of thousands patient attributes, which are particularly useful for drug safety studies of a rare

event (such as an unexpected adverse event (AE) or severe adverse event (SAE)) to protect public health. As illustrated by the USRDS data example in Section 4.3, while existing methods for the PSH model is likely to grind to a halt, the developed scalable BAR method with high performance algorithms has made it possible to analyze these massive data in real time. To this end, we point out that for a large data with millions of patient records on tens of thousands covariates, it may not always be feasible to fit a model when the data is stored in the standard dense format due to the high memory requirement. On the other hand, these massive datasets are often sparse with only a small portion of covariates are being nonzeros for a given subject. We are currently working on implementing the developed BAR method for sHDMSS competing risks data by exploiting the sparsity in the data matrix as in Chapter 3.

Finally, we emphasize that the developed CYCBAR method in Section 4.1.3 and the forward-backward scan method of Lemma 4.1 in Section 4.1.4 are of interest on their own. The CYCBAR method can be applied directly to other models and data settings. It is also straightforward to apply the forward-backward scan method to accelerate other estimation methods for the PSH model including the unpenalized estimation method of Fine and Gray (1999) and other popular penalization methods such as LASSO, SCAD and MCP, which we have developed in Chapter 5.

Appendix to Chapter 4

A4.1 Regularity conditions

Define

$$S^{(k)}(\boldsymbol{\beta}, s) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(s) Y_i(s) \mathbf{z}_i^{\otimes k} \exp(\mathbf{z}_i'\boldsymbol{\beta}), \quad k = 0, 1, 2,$$

$$\mathbf{E}(\boldsymbol{\beta}, s) = S^{(1)}(\boldsymbol{\beta}, s)/S^{(0)}(\boldsymbol{\beta}, s),$$

and

$$V(\boldsymbol{\beta}, s) = S^{(2)}(\boldsymbol{\beta}, s)/S^{(0)}(\boldsymbol{\beta}, s) - \mathbf{E}(\boldsymbol{\beta}, s)^{\otimes 2},$$

where $\mathbf{x}^{\otimes k} = (1, \mathbf{x}, \mathbf{x}\mathbf{x}')$ for $k = 0, 1, 2$, respectively. Moreover, with $N_i(t) = I(T_i \leq t, \epsilon_i = 1)$ and $Y_i(t) = 1 - N_i(t-)$ define $M_i(\boldsymbol{\beta}, t) = \int_0^t dN_i(u) - \int_0^t Y_i(u) h_{10}(u) \exp(\mathbf{z}_i'\boldsymbol{\beta}) du$. Similarly, with defining $N_i^c(t) = I(C_i \leq t)$ and $H^c(t)$ being the cumulative hazard function by treating the censored observations as events, $M_i^c(t) = N_i^c(t) - \int_0^t I(X_i \geq u) dH^c(u)$. Let $||\cdot||_p$ be the ℓ_p-norm for vectors and the norm induced by the vector p_n-norm for matrices. The following technical conditions will be needed in our derivations for the statistical properties of the pshBAR estimator.

(C1) $\int_0^\tau h_{01}(t) dt < \infty$;

(C2) There exists some compact neighborhood, $\mathcal{B}_0$, of the true value $\boldsymbol{\beta}_0$ such that for $k = 0, 1, 2$, there exists a scalar, vector, and matrix function $s^{(k)}(\boldsymbol{\beta}, t)$ defined on $\mathcal{B}_0 \times [0, \tau]$ such that

$$\sup_{t \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}_0} \left\| S^{(k)}(\boldsymbol{\beta}, t) - s^{(k)}(\boldsymbol{\beta}, t) \right\|_2 = o_p(1), \quad \text{as } n \to \infty;$$

(C3) Let $s^{(1)}(\boldsymbol{\beta}, t) = \partial s^{(0)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}$ and $s^{(2)}(\boldsymbol{\beta}, t) = \partial s^{(1)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}$. For $k = 0, 1, 2$, the functions $s^{(k)}(\boldsymbol{\beta}, t)$ are continuous with respect to $\boldsymbol{\beta} \in \mathcal{B}_0$, uniformly in $t \in [0, \tau]$, and are bounded on $\boldsymbol{\beta}_0 \times [0, \tau]$; furthermore, $s^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from zero on $\mathcal{B}_0 \times [0, \tau]$;

(C4) Let $\mathbf{e}(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$, $v(\boldsymbol{\beta}, t) = s^{(2)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t) - \mathbf{e}(\boldsymbol{\beta}, t)^{\otimes 2}$, and $\Omega = \int_0^\tau v(\boldsymbol{\beta}_0, u) s^{(0)}(\boldsymbol{\beta}_0, u) h_{10}(u) du$. There exists some constants C_2 and C_3 such that

$$0 < C_2 < \text{eigen}_{\min}(\Omega) \leq \text{eigen}_{\max}(\Omega) < C_3 < \infty,$$

where for any real diagonalizable matrix $\mathbf{A}$, $\text{eigen}_{\min}(\mathbf{A})$ and $\text{eigen}_{\max}(\mathbf{A})$ represent its smallest and largest eigenvalues, respectively; furthermore, there also exists a matrix Γ such that $\|n^{-1} \sum_{i=1}^n \text{var}(\mathbf{U}_i) - \Gamma\|_2 \to 0$, where

$$\mathbf{U}_i = \int_0^\tau \{\mathbf{z}_i(u) - \mathbf{e}(\boldsymbol{\beta}_0, u)\} w_i(u) dM_i(\boldsymbol{\beta}_0, u) + \int_0^\tau \mathbf{q}(u)/\pi(u) dM_i^c(u),$$

for

$$w_i(t) = I(C_i \geq T_i \wedge t)G(t)/G(X_i \wedge t)$$
$$\mathbf{q}(u) = -\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{z}_i(t) - \mathbf{e}(\boldsymbol{\beta}_0, t)\}w_i(t)I(X_i < u \leq t)dM_i(\boldsymbol{\beta}_0, t)$$
$$\pi(u) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I(X_i \geq u)$$

(C5) There exists a constant C_6 such that $\sup_{1\leq i\leq n} E(U_{ij}^2 U_{il}^2) < C_6 < \infty$ for all $1 \leq j, l \leq p$, where U_{ij} is the j-th element of $\mathbf{U}_i$ defined in (C4);

(C6) As $n \to \infty$, $p_n^4/n \to 0$, $\lambda_n \to \infty$, $\xi_n \to \infty$, $\xi_n b_n/\sqrt{n} \to 0$, $p/(na_n^2) \to 0$, $\lambda_n b_n^3\sqrt{q_n}/\sqrt{n} \to 0$ and $\lambda_n\sqrt{q_n}/(a_n^3\sqrt{n}) \to 0$, where $a_n = \min_{j=1,\ldots,q_n}(|\beta_{0j}|)$ and $b_n = \max_{j=1,\ldots q_n}(|\beta_{0j}|)$.

The above conditions (C1)-(C5) are similar to those proposed by Cai et al. (2005) and Ahn et al. (2018) and have been discussed in Appendix A3.1.

Remark 4.4 *Ahn et al. (2018) showed that under Conditions (C1) - (C5) and* $p_n^4/n \to 0$

$$||\dot{l}(\boldsymbol{\beta}_0)||_2 = O_p(\sqrt{np_n}) \tag{4.12}$$

and

$$n^{-1}\ddot{l}(\boldsymbol{\beta}) = \Omega + o_p(1), \tag{4.13}$$

in probability, uniformly in $\boldsymbol{\beta} \in \mathcal{B}_0$.

The proof of Theorem 4.1 parallels the proof of Theorem 3.1 in Appendix A3.2.

A4.2 Proof of Lemmas for Theorem 4.1

Lemma 4.2 (Consistency of Ridge Estimator) *Let*

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} \xi_n \beta_j^2 \right\},$$

be the PSH ridge estimator defined in Equation (3). Assume that Conditions (C1) - (C6) hold. Then

$$||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p\left[\sqrt{p_n}\{n^{-1/2}(1 + \xi_n b_n/\sqrt{n})\}\right] = O_p(\sqrt{p_n/n}), \tag{4.14}$$

where b_n is an upper bound of the true nonzero $|\beta_{0j}|$'s defined in Condition (C6).

Proof. Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + \xi_n b_n/n)$ and $\ell(\boldsymbol{\beta}) = -2l(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p_n} \beta_j^2$. To prove Lemma 4.2, it is sufficient to show that for any $\epsilon > 0$, there exists a large enough constant K_0 such that

$$\text{pr}\left\{ \inf_{||\mathbf{u}||_2 = K_0} L(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) > L(\boldsymbol{\beta}_0) \right\} \geq 1 - \epsilon, \tag{4.15}$$

since (4.15) implies that there exists a local minimum, $\hat{\boldsymbol{\beta}}_{ridge}$, inside the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : ||\mathbf{u}||_2 \leq K_0\}$ such that $||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p(\alpha_n)$, with probability tending to one. To prove

(4.15), we first note

$$\begin{aligned}
\frac{1}{n}L(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-\frac{1}{n}L(\boldsymbol{\beta}_0) &= -\frac{1}{n}\{2l(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-\frac{1}{n}2l(\boldsymbol{\beta}_0)\}+\frac{\xi_n}{n}\sum_{j=1}^{p_n}\left\{(\beta_{0j}+\alpha_n u_j)^2-\beta_{0j}^2\right\} \\
&= -\frac{1}{n}\{2l(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-2l(\boldsymbol{\beta}_0)\}+\frac{\xi_n}{n}\sum_{j=1}^{p_n}\left(2\beta_{0j}\alpha_n u_j+\alpha_n^2u_j^2\right) \\
&\geq -\frac{1}{n}\{2l(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-2l(\boldsymbol{\beta}_0)\}+\frac{2\xi_n\alpha_n}{n}\sum_{j=1}^{p_n}\beta_{0j}u_j \\
&= -\frac{1}{n}\{2l(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-2l(\boldsymbol{\beta}_0)\}+\frac{2\xi_n\alpha_n}{n}\sum_{j=1}^{q_n}\beta_{0j}u_j \\
&\equiv W_1+W_2.
\end{aligned}$$

By Taylor expansion, we have

$$\begin{aligned}
W_1 &= -\frac{2}{n}\alpha_n\mathbf{u}'\dot{l}(\boldsymbol{\beta}_0)-\frac{1}{n}\alpha_n^2\mathbf{u}'\ddot{l}(\boldsymbol{\beta}^*)\mathbf{u} \\
&= W_{11}+W_{12},
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0+\alpha_n\mathbf{u}$, and $\dot{l}(\boldsymbol{\beta})$ and $\ddot{l}(\boldsymbol{\beta})$ denote the first and second derivatives of $l(\boldsymbol{\beta})$, respectively. By the Cauchy-Schwartz inequality,

$$W_{11}=-\frac{2}{n}\alpha_n\mathbf{u}'\dot{l}(\boldsymbol{\beta}_0)\leq\frac{2}{n}\alpha_n||\dot{l}(\boldsymbol{\beta}_0)||_2\cdot||\mathbf{u}||_2=\frac{2}{n}\alpha_nO_p(\sqrt{np_n})||\mathbf{u}||_2\leq O_p(\alpha_n^2)||\mathbf{u}||_2,$$

where the second equality is due to (4.12). By (4.13) we have

$$W_{12}=-\frac{1}{n}\alpha_n^2\mathbf{u}'\ddot{l}(\boldsymbol{\beta}^*)\mathbf{u}=\alpha_n^2\mathbf{u}'\Omega\mathbf{u}\{1+o_p(1)\}.$$

Since $\text{eigen}_{min}(\Omega)\geq C_2>0$ by Condition (C4), W_{12} dominates W_{11} uniformly in $||\mathbf{u}||_2=K_0$ for a sufficiently large K_0. Furthermore

$$W_2\leq\frac{2\xi_n\alpha_n}{n}|\boldsymbol{\beta}_{01}'\mathbf{u}|\leq\frac{2\sqrt{q_n}\xi_n\alpha_nb_n}{n}||\mathbf{u}||_2=O_p(\alpha_n^2)||\mathbf{u}||_2,$$

where the last step follows from the fact that $\sqrt{q_n}\xi_n b_n/n < \sqrt{p_n}(n^{-1/2} + \xi_n b_n/n) = \alpha_n$. Therefore for a sufficiently large K_0, we have that W_{12} dominates W_{11} and W_2 uniformly in $||\mathbf{u}||_2 = K_0$. Since W_{12} is positive, (4.15) holds and therefore $||\hat{\boldsymbol{\beta}}_{ridge} - \boldsymbol{\beta}_0||_2 = O_p(\alpha_n) = O_p\left[\sqrt{p_n}\{n^{-1/2}(1 + \xi_n b_n/\sqrt{n})\}\right] = O_p(\sqrt{p_n/n})$, where the last step follows from condition (C6). □

Remark 4.5 *Recall $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1'$ and $\boldsymbol{\beta}_2'$ correspond to the first q_n and remaining $p_n - q$ components of $\boldsymbol{\beta}$, respectively. Let*

$$Q_n(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2l(\boldsymbol{\theta}) + \lambda_n \boldsymbol{\theta}' D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.16}$$

where $D(\boldsymbol{\beta}) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_{q_n}^{-2}, \beta_{q_n+1}^{-2}, \ldots, \beta_{p_n}^{-2})$ and $l(\boldsymbol{\theta})$ is the p_n-dimensional log-pseudo likelihood of the reduced model. Let $\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ and $\ddot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ be the first and second derivatives of $Q(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ with respective to $\boldsymbol{\theta}$, respectively. Then

$$\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2\dot{l}(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.17}$$

$$\ddot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2\ddot{l}(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta}). \tag{4.18}$$

Remark 4.6 *Let $\hat{\boldsymbol{\beta}}_{ridge,1}$ and $\hat{\boldsymbol{\beta}}_{ridge,2}$ denote the first q_n and the remaining $p_n - q_n$ components of $\hat{\boldsymbol{\beta}}_{ridge}$, respectively. Then, Lemma 4.2 and condition (C6) imply that for $j = 1, \ldots, q_n$ and sufficiently large n, $a_n/2 \leq |\hat{\beta}_{ridge,1j}| \leq 2b_n$, where $\hat{\beta}_{ridge,1j}$ is the jth component of $\hat{\boldsymbol{\beta}}_{ridge,1}$ and $||\hat{\boldsymbol{\beta}}_{ridge,2}||_2 = O(\sqrt{p_n/n})$.*

Remark 4.7 *Recall $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ where $\boldsymbol{\beta}_1'$ and $\boldsymbol{\beta}_2'$ correspond to the first q_n and remaining $p_n - q$ components of $\boldsymbol{\beta}$, respectively. Let*

$$Q_n(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2l(\boldsymbol{\theta}) + \lambda_n \boldsymbol{\theta}' D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.19}$$

where $D(\boldsymbol{\beta}) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_{q_n}^{-2}, \beta_{q_n+1}^{-2}, \ldots, \beta_{p_n}^{-2})$ and $l(\boldsymbol{\theta})$ is the p_n-dimensional log-pseudo likelihood of the reduced model. Let $\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ and $\ddot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ be the first and second

derivatives of $Q(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ *with respective to* $\boldsymbol{\theta}$*, respectively. Then*

$$\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2\dot{l}(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.20}$$

$$\ddot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2\ddot{l}(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta}). \tag{4.21}$$

Lemma 4.3 *Let* $M_n = \max\{2/a_n, 2b_n\}$*. Define* $\mathcal{H}_n \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' : |\boldsymbol{\beta}_1| = (|\beta_1|, \ldots, |\beta_{q_n}|)' \in [1/M_n, M_n]^{q_n}, 0 < \|\boldsymbol{\beta}_2\|_2 \leq \delta_n\sqrt{p_n/n}, \}$*, where* δ_n *is a sequence of positive real numbers satisfying* $\delta_n \to \infty$ *and* $p_n\delta_n^2/\lambda_n \to 0$*. For any given* $\boldsymbol{\beta} \in \mathcal{H}_n$*, define*

$$Q_n(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2l(\boldsymbol{\theta}) + \lambda_n \boldsymbol{\theta}' D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.22}$$

where $l(\boldsymbol{\theta})$ *is the* p_n*-dimensional log-pseudo likelihood and* $D(\boldsymbol{\beta}) = diag(\beta_1^{-2}, \ldots, \beta_{p_n}^{-2})$*. Let* $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})', g_2(\boldsymbol{\beta})')'$ *be a solution to* $\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = \mathbf{0}$*, where*

$$\dot{Q}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) = -2\dot{l}(\boldsymbol{\theta}) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta}, \tag{4.23}$$

is the derivative of $Q(\boldsymbol{\theta} \mid \boldsymbol{\beta})$ *with respective to* $\boldsymbol{\theta}$*. Assume that conditions (C1) - (C6) hold. Then, as* $n \to \infty$*, with probability tending to 1,*

(a) $\sup_{\boldsymbol{\beta} \in \mathcal{H}_n} \frac{\|g_2(\boldsymbol{\beta})\|_2}{\|\boldsymbol{\beta}_2\|_2} \leq \frac{1}{K_1}$, *for some constant* $K_1 > 1$*;*

(b) $|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}$*.*

Proof. By the first-order Taylor expansion and the definition of $g(\boldsymbol{\beta})$, we have

$$\dot{Q}(\boldsymbol{\beta}_0|\boldsymbol{\beta}) = \dot{Q}\{g(\boldsymbol{\beta}) \mid \boldsymbol{\beta}\} + \ddot{Q}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta})\{\boldsymbol{\beta}_0 - g(\boldsymbol{\beta})\} = \ddot{Q}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta})\{\boldsymbol{\beta}_0 - g(\boldsymbol{\beta})\}, \tag{4.24}$$

where $\boldsymbol{\beta}_0$ is the true parameter vector, and $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $g(\boldsymbol{\beta})$. Rearranging terms, we have

$$\ddot{Q}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta})g(\boldsymbol{\beta}) = -\dot{Q}(\boldsymbol{\beta}_0|\boldsymbol{\beta}) + \ddot{Q}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta})\boldsymbol{\beta}_0, \tag{4.25}$$

which can be rewritten as

$$\left\{-2\ddot{l}(\boldsymbol{\beta}^*) + 2\lambda_n D(\boldsymbol{\beta})\right\} g(\boldsymbol{\beta}) = -\left\{-2\dot{l}(\boldsymbol{\beta}_0) + 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\beta}_0\right\} + \left\{-2\ddot{l}(\boldsymbol{\beta}^*) + 2\lambda_n D(\boldsymbol{\beta})\right\}\boldsymbol{\beta}_0$$
$$= 2\dot{l}(\boldsymbol{\beta}_0) - 2\ddot{l}(\boldsymbol{\beta}^*)\boldsymbol{\beta}_0.$$

Write $H(\boldsymbol{\beta}) \equiv -n^{-1}\ddot{l}(\boldsymbol{\beta})$, we have

$$\left\{H(\boldsymbol{\beta}^*) + \frac{\lambda_n}{n} D(\boldsymbol{\beta})\right\} g(\boldsymbol{\beta}) = H(\boldsymbol{\beta}^*)\boldsymbol{\beta}_0 + \frac{1}{n}\dot{l}(\boldsymbol{\beta}_0), \tag{4.26}$$

which can be further written as

$$\{g(\boldsymbol{\beta}) - \boldsymbol{\beta}_0\} + \frac{\lambda_n}{n} H(\boldsymbol{\beta}^*)^{-1} D(\boldsymbol{\beta}) g(\boldsymbol{\beta}) = \frac{1}{n} H(\boldsymbol{\beta}^*)^{-1}\dot{l}(\boldsymbol{\beta}_0). \tag{4.27}$$

Now we partition $H(\boldsymbol{\beta}^*)^{-1}$ into

$$H(\boldsymbol{\beta}^*)^{-1} = \begin{bmatrix} A & B \\ B' & G \end{bmatrix}$$

and partition $D(\boldsymbol{\beta})$ into

$$D(\boldsymbol{\beta}) = \begin{bmatrix} D_1(\boldsymbol{\beta}_1) & \mathbf{0} \\ \mathbf{0}' & D_2(\boldsymbol{\beta}_2) \end{bmatrix}$$

where $D_1(\boldsymbol{\beta}_1) = \text{diag}(|\beta_1|^{-2}, ..., |\beta_{q_n}|^{-2})$ and $D_2(\boldsymbol{\beta}_2) = \text{diag}(|\beta_{q_n+1}|^{-2}, ..., |\beta_{p_n}|^{-2})$. Then (4.27) can be re-written as

$$\begin{pmatrix} g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01} \\ g_2(\boldsymbol{\beta}) \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \\ B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta}) \end{pmatrix} = \frac{1}{n} H(\boldsymbol{\beta}^*)^{-1}\dot{l}(\boldsymbol{\beta}_0). \tag{4.28}$$

Moreover, it follows from (4.12), (4.13), and condition (C5) that

$$\left\| n^{-1} H(\boldsymbol{\beta}^*)^{-1}\dot{l}(\boldsymbol{\beta}_0) \right\|_2 = O_p(\sqrt{p_n/n}). \tag{4.29}$$

Therefore,

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|g_2(\boldsymbol{\beta})+\frac{\lambda_n}{n}B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})+\frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2=O_p(\sqrt{p_n/n}). \tag{4.30}$$

Furthermore,

$$\begin{aligned}\|g(\boldsymbol{\beta})-\boldsymbol{\beta}_0\|_2 &= \left\|-\left\{H(\boldsymbol{\beta}^*)+\frac{\lambda_n}{n}D(\boldsymbol{\beta})\right\}^{-1}\left\{\frac{\lambda_n}{n}D(\boldsymbol{\beta})\boldsymbol{\beta}_0-\frac{1}{n}\dot{l}(\boldsymbol{\beta}_0)\right\}\right\|_2\\ &\le \left\|\{H(\boldsymbol{\beta}^*)\}^{-1}\left\{\frac{\lambda_n}{n}D(\boldsymbol{\beta})\boldsymbol{\beta}_0-\frac{1}{n}\dot{l}(\boldsymbol{\beta}_0)\right\}\right\|_2\\ &\le \left\|\{H(\boldsymbol{\beta}^*)\}^{-1}\right\|_2\cdot\left\{\left\|\frac{\lambda_n}{n}D_1(\boldsymbol{\beta}_1)\boldsymbol{\beta}_{01}\right\|_2+\left\|\frac{1}{n}\dot{l}(\boldsymbol{\beta}_0)\right\|_2\right\}\\ &= O_p(1)\left\{O(n^{-1}\lambda_nM_n^3\sqrt{q_n})+O_p(\sqrt{p_n/n})\right\}\\ &= O_p(\sqrt{p_n/n}),\end{aligned}$$

where the first equality follows from (4.26) and the fourth step follows from (4.13), condition (C3), $\|n^{-1}\lambda_nD_1(\boldsymbol{\beta}_1)\boldsymbol{\beta}_{01}\|_2=O(n^{-1}\lambda_nM_n^3\sqrt{q_n})$, and $\left\|n^{-1}\dot{l}(\boldsymbol{\beta}_0)\right\|_2=O_p(\sqrt{p_n/n})$, and the last step holds since $n^{-1}\lambda_nM_n^3\sqrt{q_n}=o(1/\sqrt{n})$ under condition (C6). Hence,

$$\|g(\boldsymbol{\beta})\|_2\le\|\boldsymbol{\beta}_0\|_2+\|g(\boldsymbol{\beta})-\boldsymbol{\beta}_0\|_2=O_p(M_n\sqrt{q_n}). \tag{4.31}$$

Also note that $\|B\|_2=O_p(1)$ since $\|BB'\|_2\le\|A^2+BB'\|_2+\|A^2\|_2\le 2\|A^2+BB'\|_2\le 2\|H(\boldsymbol{\beta}^*)^{-2}\|_2=O_p(1)$. This, combined with (4.31), implies that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|\frac{\lambda_n}{n}B'D_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})\right\|_2\le\frac{\lambda_n}{n}\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\|B\|_2\|D_1(\boldsymbol{\beta}_1)\|_2\|g_1(\boldsymbol{\beta})\|_2=O_p\left(\frac{\lambda_nM_n^3\sqrt{q_n}}{n}\right)=o(1/\sqrt{n}). \tag{4.32}$$

It then follows that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\left\|g_2(\boldsymbol{\beta})+\frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2\le O_p(\sqrt{p_n/n})+o(1/\sqrt{n})=O_p(\sqrt{p_n/n}).$$

Since G is positive definite and symmetric with probability tending to one, by the spectral decomposition theorem, $G=\sum_{i=1}^{p_n-q_n}r_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}'$, where r_{2i} and $\mathbf{u}_{2i}$ are the eigenvalues and

eigenvectors of G, respectively. Now with probability tending to one,

$$\begin{aligned}\frac{\lambda_n}{n}\left\|GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 &= \frac{\lambda_n}{n}\left\|\left(\sum_{i=1}^{p_n-q_n} r_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}'\right)D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \\ &\geq \frac{\lambda_n}{n}\left\|C_2\left(\sum_{i=1}^{p_n-q_n} \mathbf{u}_{2i}\mathbf{u}_{2i}'\right)D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \\ &\geq C_2\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2, \end{aligned} \tag{4.33}$$

where the first inequality is due to (4.13) and condition (C4) since we can assume that for all $i = 1, \ldots, p-q$, $r_{2i} \in (C_2, C_3)$ for some $C_2 < C_3 < \infty$ with probability tending to one. Therefore with probability tending to one,

$$C_2\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 - \left\|g_2(\boldsymbol{\beta})\right\|_2 \leq \left\|g_2(\boldsymbol{\beta}) + \frac{\lambda_n}{n}GD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \leq \delta_n\sqrt{p_n/n}, \tag{4.34}$$

where δ_n diverges to ∞. Let $\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2} = (g_2(\beta_{q_n+1})/\beta_{q_n+1}, \ldots, g_2(\beta_{p_n})/\beta_{p_n})'$. Because $||\boldsymbol{\beta}_2||_2 \leq \delta_n\sqrt{p_n/n}$, we have

$$C_2\left\|\frac{\lambda_n}{n}D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 = C_2\frac{\lambda_n}{n}\left\|D_2(\boldsymbol{\beta}_2)^{1/2}\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \geq C_2\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2, \tag{4.35}$$

and

$$\left\|g_2(\boldsymbol{\beta})\right\|_2 = \left\|D_2(\boldsymbol{\beta}_2)^{-1/2}\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \left\|D_2(\boldsymbol{\beta}_2)^{-1/2}\right\|_2 \cdot \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2. \tag{4.36}$$

Hence it follows from (4.34), (4.35), and (4.36) that with probability tending to one,

$$C_2\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 - \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \delta_n\sqrt{p_n/n}.$$

This implies that with probability tending to one,

$$\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \frac{1}{\lambda_n/(C_1 p\delta_n^2) - 1} < \frac{1}{K_1}, \tag{4.37}$$

for some constant $K_1 > 1$ provided that $\lambda_n/(p_n\delta_n^2) \to \infty$ as $n \to \infty$. Now from (4.37), we have

$$\|g_2(\boldsymbol{\beta})\|_2 \leq \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \max_{q+1\leq j\leq p} |\beta_j| \leq \left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \|\boldsymbol{\beta}_2\|_2 \leq \frac{1}{K_1}\|\boldsymbol{\beta}_2\|_2, \tag{4.38}$$

with probability tending to one. Thus

$$\text{pr}\left(\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \frac{\|g_2(\boldsymbol{\beta})\|_2}{\|\boldsymbol{\beta}_2\|_2} < \frac{1}{K_1}\right) \to 1 \quad \text{as } n \to \infty$$

and (a) is proved.

To prove part (b), we first note from (4.38) that as $n \to \infty$, $\text{pr}(\left\|\mathbf{m}_{g_2(\boldsymbol{\beta})/\boldsymbol{\beta}_2}\right\|_2 \leq \delta_n\sqrt{p_n/n}) \to 1$. Therefore it is sufficient to show that for any $\boldsymbol{\beta} \in \mathcal{H}_n$, $|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}$ with probability tending to 1. By (4.28) and (4.29), we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|(g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}) + \frac{\lambda_n}{n}AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta}) + \frac{\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 = O_p(\sqrt{p_n/n}). \tag{4.39}$$

Similar to (4.32), it can be shown that

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|\frac{\lambda_n}{n}AD_1(\boldsymbol{\beta}_1)g_1(\boldsymbol{\beta})\right\|_2 = O_p\left(\frac{\lambda_n M_n^3\sqrt{q_n}}{n}\right) = o_p(1/\sqrt{n}), \tag{4.40}$$

where the last equality holds trivially under condition (C6). Furthermore, with probability tending to one,

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \left\|\frac{\lambda_n}{n}BD_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\right\|_2 \leq \frac{\lambda_n}{n}\sup_{\boldsymbol{\beta}\in\mathcal{H}_n}\|B\|_2 \cdot \|D_2(\boldsymbol{\beta}_2)g_2(\boldsymbol{\beta})\|_2 \leq \frac{\lambda_n}{n}\sqrt{2K_3}\left(\delta_n\sqrt{\frac{p_n}{n}}\right)^2, \tag{4.41}$$

for some $K_3 > 0$, since $||g_2(\boldsymbol{\beta})|| \leq \delta_n\sqrt{p_n/n}$, $||B||_2 = O_p(1)$ and $\|D_2(\boldsymbol{\beta}_2)\|_2 \leq \delta_n\sqrt{p_n/n}$. Therefore, combing (4.39), (4.40) and (4.41) gives

$$\sup_{\boldsymbol{\beta}\in\mathcal{H}_n} \|g_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{01}\|_2 \leq \frac{\lambda_n}{n}\sqrt{2K_3}\left(\delta_n\sqrt{\frac{p_n}{n}}\right)^2 + \frac{\delta_n\sqrt{p_n}}{\sqrt{n}},$$

with probability tending to one. Because $\lambda_n/n \to 0$ and $\delta_n\sqrt{p_n/n} = \sqrt{p_n\delta_n^2/\lambda_n}\sqrt{\lambda_n/n} \to 0$ as $n \to \infty$, we have $\mathrm{pr}(|g_1(\boldsymbol{\beta})| \in [1/M_n, M_n]^{q_n}) \to 1$. This completes the proof of part (b). $\square$

Lemma 4.4 *Let $\boldsymbol{\beta}_1$ be the first q_n components of $\boldsymbol{\beta}$. Define $f(\boldsymbol{\beta}_1) = \arg\min_{\boldsymbol{\theta}_1}\{Q_{n1}(\boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_1)\}$, where $Q_{n1}(\boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_1) = -2l_{n1}(\boldsymbol{\theta}_1) + \lambda_n\boldsymbol{\theta}_1' D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1$, is a weighted ℓ_2-penalized -2log-pseudo likelihood for the oracle model of model size q_n, and $D_1(\boldsymbol{\beta}_1) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_{q_n}^{-2})$. Assume that conditions (C1) - (C6) hold. Then with probability tending to one,*

(a) $f(\boldsymbol{\beta}_1)$ is a contraction mapping from $[1/M_n, M_n]^{q_n}$ to itself;

(b) $\sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) \to N(0,1)$, for any q_n-dimensional vector $\mathbf{d}_n$ such that $\mathbf{d}_n'\mathbf{d}_n = 1$ and where $\hat{\boldsymbol{\beta}}_1^\circ$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$ and Σ_{11} and Ω_{11} are the first $q_n \times q_n$ submatrices of Σ and Ω, respectively.

Proof: (a) First we show that $f(\cdot)$ is a mapping from $[1/M_n, M_n]^{q_n}$ to itself with probability tending to one. Again through a first order Taylor expansion, we have

$$\{f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01}\} + \frac{\lambda_n}{n}H_1(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1) = \frac{1}{n}H_1(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_1(\boldsymbol{\beta}_{01}), \tag{4.42}$$

where $H_1(\boldsymbol{\beta}_1^*) = -n^{-1}\ddot{l}_1(\boldsymbol{\beta}_1^*)$ exists and is invertible for $\boldsymbol{\beta}_1^*$ between $\boldsymbol{\beta}_{01}$ and $f(\boldsymbol{\beta}_1)$. We have

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_n,M_n]^{q_n}} \left\| f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01} + \frac{\lambda_n}{n}H_1(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1)\right\|_2 = O_p(\sqrt{q_n/n}),$$

where the right-hand side follows in the same fashion as (4.32). Similar to (4.32) we have

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_0,M_0]^{q_n}} \left\| \frac{\lambda_n}{n}H_1(\boldsymbol{\beta}_1^*)^{-1}D_1(\boldsymbol{\beta}_1)f(\boldsymbol{\beta}_1)\right\|_2 = O_p\left(\frac{\lambda_n M_n^3}{\sqrt{n}}\sqrt{\frac{q_n}{n}}\right) = o_p\left(1/\sqrt{n}\right).$$

Therefore, with probability tending to one

$$\sup_{|\boldsymbol{\beta}_1|\in[1/M_n,M_n]^{q_n}} \|f(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_{01}\|_2 \le \delta_n\sqrt{q_n/n}, \tag{4.43}$$

where δ_n is a sequence of real numbers diverging to ∞ and satisfies $\delta_n\sqrt{p_n/n} \to 0$. As a result, we have

$$\mathrm{pr}(f(\boldsymbol{\beta}_1) \in [1/M_n, M_n]^{q_n}) \to 1$$

as $n \to \infty$. Hence $f(\cdot)$ is a mapping from the region $[1/M_n, M_n]^{q_n}$ to itself. To prove that $f(\cdot)$ is a contraction mapping, we need to further show that

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2 = o_p(1). \tag{4.44}$$

Since $f(\boldsymbol{\beta}_1)$ is a solution to $\dot{Q}_1(\boldsymbol{\theta}_1 \mid \boldsymbol{\beta}_1) = 0$, we have

$$-\frac{1}{n}\dot{l}_1(f(\boldsymbol{\beta}_1)) = -\frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) f(\boldsymbol{\beta}_1). \tag{4.45}$$

Taking the derivative of (4.45) with respect to $\boldsymbol{\beta}_1'$ and rearranging terms, we obtain

$$\left\{ H_1(f(\boldsymbol{\beta}_1)) + \frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) \right\} \dot{f}(\boldsymbol{\beta}_1) = \frac{2\lambda_n}{n} diag\{f_1(\boldsymbol{\beta}_1)/\beta_1^3, \ldots, f_{q_n}(\boldsymbol{\beta}_1)/\beta_{q_n}^3\}. \tag{4.46}$$

With probability tending to one, we have

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \frac{2\lambda_n}{n} \left\| diag\{f_1(\boldsymbol{\beta}_1)/\beta_1^3, \ldots, f_{q_n}(\boldsymbol{\beta}_1)/\beta_{q_n}^3\} \right\|_2 = O_p\left(\frac{\lambda_n M_n^4}{n}\right) = o_p(1),$$

where the last step follows from condition (C6). This, combined with (4.46) implies that

$$\sup_{|\boldsymbol{\beta}_1| \in [1/M_n, M_n]^{q_n}} \left\| \left\{ H_1(f(\boldsymbol{\beta}_1)) + \frac{\lambda_n}{n} D_1(\boldsymbol{\beta}_1) \right\} \dot{f}(\boldsymbol{\beta}_1) \right\|_2 = o_p(1). \tag{4.47}$$

Now, it can be shown that probability tending to one,

$$\left\| H_1(f(\boldsymbol{\beta}_1)) \dot{f}(\boldsymbol{\beta}_1) \right\|_2 \geq \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2 \cdot \left\| H_1(f(\boldsymbol{\beta}_1))^{-1} \right\|_2^{-1} \geq \frac{1}{K_2} \left\| \dot{f}(\boldsymbol{\beta}_1) \right\|_2,$$

for some $K_2 > 0$, and that

$$\frac{\lambda_n}{n}\left\|D_1(\boldsymbol{\beta}_1)\dot{f}(\boldsymbol{\beta}_1)\right\|_2 \geq \frac{\lambda_n}{n}\left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2 \left\|D_1(\boldsymbol{\beta}_1)^{-1}\right\|_2^{-1} \geq \frac{\lambda_n}{n}\frac{1}{M_n^2}\left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2.$$

Therefore, combining the above two inequalities with (4.46) and (4.47) gives

$$\left(\frac{1}{K_2} - \frac{\lambda_n}{nM_n^2}\right) \sup_{|\boldsymbol{\beta}_1|\in[1/M_n, M_n]^{q_n}} \left\|\dot{f}(\boldsymbol{\beta}_1)\right\|_2 = o_p(1).$$

This, together with the fact that $\frac{\lambda_n}{n}\frac{1}{M_n^2} = o(1)$, implies that (4.44) holds. Therefore, with probability tending to one, $f(\cdot)$ is a contraction mapping and consequently has a unique fixed point, say $\hat{\boldsymbol{\beta}}_1^\circ$, such that $\hat{\boldsymbol{\beta}}_1^\circ = f(\hat{\boldsymbol{\beta}}_1^\circ)$.

We next prove part (b). By (4.42) we have

$$f(\boldsymbol{\beta}_1) = \left\{H_1(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\boldsymbol{\beta}_1)\right\}^{-1}\left\{H_1(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01} + \frac{1}{n}\dot{l}_1(\boldsymbol{\beta}_{01})\right\}.$$

Now,

$$\begin{aligned}\sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) &= \sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}\left[\left\{H_1(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1} H_1(\boldsymbol{\beta}_1^*) - I_{q_n}\right]\boldsymbol{\beta}_{01} \\ &\quad + \sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}\left[\left\{H_1(\boldsymbol{\beta}_1^*) + \frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_1(\boldsymbol{\beta}_{01})\right] \\ &= I_1 + I_2. \end{aligned} \tag{4.48}$$

Note that for any two conformable invertible matrices Φ and Ψ, we have

$$(\Phi + \Psi)^{-1} = \Phi^{-1} - \Phi^{-1}\Psi(\Phi + \Psi)^{-1},$$

Thus we can rewrite I_1 as

$$
\begin{aligned}
I_1 &= \sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}\left[\left\{H_1(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}H_1(\boldsymbol{\beta}_1^*)-I_{q_n}\right]\boldsymbol{\beta}_{01}\\
&= -\frac{\lambda_n}{\sqrt{n}}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\left\{H_1(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}H_1(\boldsymbol{\beta}_1^*)\boldsymbol{\beta}_{01}.
\end{aligned}
$$

Moreoever

$$
\begin{aligned}
\|I_1\|_2 &\le \frac{\lambda_n}{\sqrt{n}}\left\|\Gamma_{11}^{-1/2}\Omega_{11}\right\|_2\left\|H_1(\boldsymbol{\beta}_1^*)^{-1}\right\|_2\left\|D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\|_2\left\|\left\{H_1(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\right\|_2\|H_1(\boldsymbol{\beta}_1^*)\|_2\|\boldsymbol{\beta}_{01}\|_2\\
&= \frac{\lambda_n}{\sqrt{n}}\cdot O(1)\cdot O_p(1)\cdot M_n^2\cdot O_p(1)\cdot O_p(1)\cdot M_n\sqrt{q_n}\\
&= O_p(\lambda_n M_n^3\sqrt{q_n}/\sqrt{n}) = o_p(1), \qquad (4.49)
\end{aligned}
$$

where the first equality follows from (4.13) and condition (C4), and the last equality is a consequence of condition (C6). Similarly, we can rewrite I_2 as

$$
\begin{aligned}
I_2 &= \sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}\left[\left\{H_1(\boldsymbol{\beta}_1^*)+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_1(\boldsymbol{\beta}_{01})\right]\\
&= \mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}\frac{1}{\sqrt{n}}\dot{l}_1(\boldsymbol{\beta}_{01})\\
&\quad -\frac{\lambda_n}{\sqrt{n}}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\left\{H_1(\boldsymbol{\beta}_1^*)^{-1}+\frac{\lambda_n}{n}D_1(\hat{\boldsymbol{\beta}}_1^\circ)\right\}^{-1}\frac{1}{n}\dot{l}_1(\boldsymbol{\beta}_{01})\\
&= \mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}\frac{1}{\sqrt{n}}\dot{l}_{n1}(\boldsymbol{\beta}_{01})+o_p(1). \qquad (4.50)
\end{aligned}
$$

We now establish the asymptotic normality of $n^{-1/2}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_1(\boldsymbol{\beta}_{01})$ which will be derived in a similar manner to the proof of Theorem 2 in (Cai et al., 2005). By (4.13),

(4.43), and the continuity of Ω, we can deduce that $H_1(\boldsymbol{\beta}^*) = \Omega_{11} + o_p(1)$. This implies that

$$\begin{aligned} I_2 &= n^{-1/2} \sum_{i=1}^{n} \mathbf{d}_n' \Gamma_{11}^{-1/2} \Omega_{11} H_1(\boldsymbol{\beta}_1^*)^{-1} \mathbf{U}_{i1} + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \mathbf{d}_n' \Gamma_{11}^{-1/2} \mathbf{U}_{i1} + \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{d}_n' \Gamma_{11}^{-1/2} \Omega_{11} \mathbf{U}_{i1} \right\} o_p(1) + o_p(1) \\ &= I_{21} + I_{22} \cdot o_p(1) + o_p(1), \end{aligned} \tag{4.51}$$

where $\mathbf{U}_{i1}$ consists of the first q_n components of $\mathbf{U}_i$. Letting $Y_{ni} = n^{-1/2} \mathbf{d}_n' \Gamma_{11}^{-1/2} \mathbf{U}_{i1}$, then by condition (C5)

$$\begin{aligned} s_n^2 = \sum_{i=1}^{n} \mathrm{var}(Y_{ni}) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_n' \Gamma_{11}^{-1/2} \mathrm{var}(\mathbf{U}_{i1}) \Gamma_{11}^{-1/2} \mathbf{d}_n \\ &= \mathbf{d}_n' \Gamma_{11}^{-1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathrm{var}(\mathbf{U}_{i1}) \right\} \Gamma_{11}^{-1/2} \mathbf{d}_n \to 1. \end{aligned}$$

To prove the asymptotic normality of I_{21}, we need to verify the Lindeberg condition: for all $\epsilon > 0$,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} E\{Y_{ni}^2 I(|Y_{ni}| \geq \epsilon s_n)\} \to 0, \tag{4.52}$$

as $n \to \infty$. Note that

$$\begin{aligned} \sum_{i=1}^{n} E(Y_{ni}^4) &= n^{-2} \sum_{i=1}^{n} E\left[\left\{ \mathbf{d}_n' \Gamma_{11}^{-1/2} \mathbf{U}_{i1} \right\}^4 \right] \\ &\leq n^{-2} \sum_{i=1}^{n} E\left[||\mathbf{d}_n||_2^4 \cdot ||\Gamma_{11}^{-1/2}||_2^4 \cdot ||\mathbf{U}_{i1}||_2^4 \right] \\ &= n^{-2} \mathrm{eigen}_{\max}^2\{\Gamma_{11}^{-1}\} \sum_{i=1}^{n} E(||\mathbf{U}_{i1}||_2^4) \\ &= n^{-2} \mathrm{eigen}_{\max}^2\{\Gamma_{11}^{-1}\} \sum_{i=1}^{n} \sum_{j=1}^{p_n} \sum_{k=1}^{p_n} E(U_{ij}^2 U_{ik}^2) \\ &= O(p^2/n), \end{aligned} \tag{4.53}$$

where the first inequality is due to Cauchy-Schwarz, the second equality is due to $||\mathbf{d}_n||_2 = 1$ and the last step follows from conditions (C4) and (C5). Therefore for any $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{s_n^2}\sum_{i=1}^{n} E\left\{Y_{ni}^2 I(|Y_{ni}| > \epsilon s_n)\right\} &\leq \frac{1}{s_n^2}\sum_{i=1}^{n}\left\{E(Y_{ni}^4)\right\}^{1/2}\left[E\left\{I(|Y_{ni}| > \epsilon s_n)\right\}^2\right]^{1/2} \\
&\leq \frac{1}{s_n^2}\left\{\sum_{i=1}^{n} E(Y_{ni}^4)\right\}^{1/2} \cdot \left\{\sum_{i=1}^{n} \mathrm{pr}(|Y_{ni}| > \epsilon s_n)\right\}^{1/2} \\
&\leq \frac{1}{s_n^2}\left\{\sum_{i=1}^{n} E(Y_{ni}^4)\right\}^{1/2} \cdot \left\{\sum_{i=1}^{n} \frac{\mathrm{var}(Y_{ni})}{\epsilon^2 s_n^2}\right\}^{1/2} \\
&= \frac{1}{s_n^2}\left\{O(p^2/n)\right\}^{1/2}\frac{1}{\epsilon} \to 0.
\end{aligned}$$

Thus, (4.52) is satisfied and by the Lindeberg-Feller central limit theorem and Slutsky's theorem

$$I_{21} = s_n\left(\frac{1}{s_n}\sum_{i=1}^{n} Y_{ni}\right) \to N(0,1). \tag{4.54}$$

Similarly one can show that $I_{22} = O_p(1)$ and by Slutsky's theorem,

$$\begin{aligned}
n^{-1/2}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}H_1(\boldsymbol{\beta}_1^*)^{-1}\dot{l}_1(\boldsymbol{\beta}_{01}) &= n^{-1/2}\sum_{i=1}^{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\mathbf{U}_{i1} \\
&\quad + \left\{n^{-1/2}\sum_{i=1}^{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}\mathbf{U}_{i1}\right\}o_p(1) + o_p(1) \\
&= I_{21} + I_{22}\cdot o_p(1) + o_p(1) \\
&\to N(0,1).
\end{aligned}$$

Hence, combining (4.48), (4.49), (4.51), and (4.54) gives

$$\sqrt{n}\mathbf{d}_n'\Gamma_{11}^{-1/2}\Omega_{11}(\hat{\boldsymbol{\beta}}_1^\circ - \boldsymbol{\beta}_{01}) \to N(0,1),$$

which proves part (b). □

A4.3 Proof of Theorem 4.1

Part (a) of the theorem follows immediately from part (a) of Lemma 3 in the Supplementary Material. Part (b) of the theorem will follow from part (b) Lemma 4 in the Supplementary Material and the following

$$\Pr\left(\lim_{k\to\infty}\left\|g_1(\boldsymbol{\beta}^{(k)})-\hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2=0\right)\to 1, \tag{4.55}$$

where $\hat{\boldsymbol{\beta}}_1^{\circ}$ is the fixed point of $f(\boldsymbol{\beta}_1)$ defined in Lemma 4 in the Supplementary Material. Note that $g(\boldsymbol{\beta})$ is a solution to

$$-\frac{1}{n}D(\boldsymbol{\beta})^{-1}\dot{l}(\boldsymbol{\theta})+\frac{1}{n}\lambda_n\boldsymbol{\theta}=\mathbf{0}, \tag{4.56}$$

where $D(\boldsymbol{\beta})^{-1}=diag\{\beta_1^2,\ldots,\beta_{q_n}^2,\beta_{q_n+1}^2,\ldots,\beta_{p_n}^2\}$. It is easy to see from (4.56) that

$$\lim_{\boldsymbol{\beta}_2\to 0}g_2(\boldsymbol{\beta})=\mathbf{0}_{p_n-q_n}.$$

This, combined with (4.56), implies that for any $\boldsymbol{\beta}_1$

$$\lim_{\boldsymbol{\beta}_2\to 0}g_1(\boldsymbol{\beta})=f(\boldsymbol{\beta}_1).$$

Hence, $g(\cdot)$ is continuous and thus uniform continuous on the compact set $\boldsymbol{\beta}\in\mathcal{H}_n$. Hence as $k\to\infty$,

$$\omega_k\equiv\sup_{|g_1(\boldsymbol{\beta})|\in[1/M_n,M_n]^{q_n}}\left\|g_1(\boldsymbol{\beta}_1,\hat{\boldsymbol{\beta}}_2^{(k)})-f(\boldsymbol{\beta}_1)\right\|_2\to 0, \tag{4.57}$$

with probability tending to one. Furthermore,

$$\left\|\hat{\boldsymbol{\beta}}_1^{(k+1)}-\hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2\le\left\|g_1(\hat{\boldsymbol{\beta}}^{(k)})-f(\hat{\boldsymbol{\beta}}_1^{(k)})\right\|_2+\left\|f(\hat{\boldsymbol{\beta}}_1^{(k)})-\hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2\le\omega_k+\frac{1}{K_4}\left\|\hat{\boldsymbol{\beta}}_1^{(k)}-\hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2, \tag{4.58}$$

for some $K_4 > 1$, where the last inequality follows from (4.44) and the definition of ω_k. Denote by $a_k = \left\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2$, we can rewrite (4.58) as

$$a_{k+1} \leq \frac{1}{K_4} a_k + \omega_k.$$

By (4.57), for any $\epsilon > 0$, there exists an $N > 0$ such that $\omega_k < \epsilon$ for all $k > N$. Therefore for $k > N$,

$$\begin{aligned}
a_{k+1} &\leq \frac{1}{K_4} a_k + \omega_k \\
&\leq \frac{a_{k-1}}{K_4^2} + \frac{\omega_{k-1}}{K_4} + \omega_k \\
&\leq \frac{a_1}{K_4^k} + \frac{\omega_1}{K_4^{k-1}} + \cdots + \frac{\omega_N}{K_2^{k-N}} + \left(\frac{\omega_{N+1}}{K_4^{k-N-1}} + \cdots + \frac{\omega_{k-1}}{K_4} + \omega_k \right) \\
&\leq (a_1 + \omega_1 + ... + \omega_N) \frac{1}{K_4^{k-N}} + \frac{1 - (1/K_4)^{k-N}}{1 - 1/K_4} \epsilon \to 0, \quad \text{as } k \to \infty,
\end{aligned}$$

with probability tending to one. Therefore,

$$\Pr\left(\lim_{k \to \infty} \left\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^{\circ}\right\|_2 = \mathbf{0} \right) = 1$$

with probability tending to one, or equivalently

$$\Pr(\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^{\circ}) = 1 \tag{4.59}$$

with probability tending to one. This proves (4.55) and thus complete the proof of the theorem. □

A4.4 Proof of Theorem 4.2.

Under Conditions (C1) - (C6) in Appendix A4.1, by Theorem 4.1 we have that $\hat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}^{(k)}$, where

$$\hat{\boldsymbol{\beta}}^{(k+1)} = g(\hat{\boldsymbol{\beta}}^{(k)}) = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p_n} \frac{I(\beta_j \neq 0)\beta_j^2}{\left(\hat{\beta}_j^{(k)}\right)^2} \right\}.$$

Note that

$$D(\hat{\boldsymbol{\beta}}^{(k)})^{-1} \dot{l}_n(\hat{\boldsymbol{\beta}}^{(k+1)}) = \lambda_n \hat{\boldsymbol{\beta}}^{(k+1)}.$$

Therefore for any $l = i, j$ where $\hat{\beta}_i \neq 0$, $\hat{\beta}_j \neq 0$,

$$\hat{\beta}_l^{(k+1)} = \frac{(\hat{\beta}_l^{(k)})^2}{\lambda_n} \dot{l}_{nl}(\hat{\boldsymbol{\beta}}^{(k+1)}).$$

Letting $k \to \infty$, (4.60), we have

$$\hat{\beta}_l^{-1} = \frac{1}{\lambda_n} \dot{l}_{nl}(\hat{\boldsymbol{\beta}}).$$

Letting $\boldsymbol{\eta} = Z\boldsymbol{\beta}$ we can rewrite the score function

$$\zeta(\eta_i) = \frac{\partial}{\partial \eta_i} l_n(\boldsymbol{\eta}) = \int_0^\tau \hat{w}_i(s) dN_i(s) + \int_0^\tau \frac{\hat{w}_i^2(s) Y_i(s) \exp(\hat{\eta}_i)}{\sum_{j=1}^n \hat{w}_j(s) Y_j(s) \exp(\hat{\eta}_j)} d\bar{N}(s) \quad i = 1, \ldots, n.$$

Recall that $\hat{w}_i(s) Y_i(s) \in [0, 1]$ for all $i = 1, \ldots n$. Then

$$|\zeta(\hat{\eta}_i)| \leq |N_i(1)| + \left| \int_0^\tau \frac{\hat{w}_i^2(s) Y_i(s) \exp(\hat{\eta}_i)}{\sum_{j=1}^n \hat{w}_j(s) Y_j(s) \exp(\hat{\eta}_j)} d\bar{N}(s) \right| \leq 1 + e_n \quad i = 1, \ldots, n,$$

where $e_n = \sum_{i=1}^n I(\epsilon_i = 1)$. Hence

$$\|\zeta(\hat{\boldsymbol{\eta}})\|_2 \leq \|\mathbf{1} + e_n \mathbf{1}\|_2 = \sqrt{n(1 + e_n)^2}.$$

Let $\mathbf{z}_{[,i]}$ denote the ith column of Z. Since Z is assumed to be standardized, $\mathbf{z}_{[,i]}' \mathbf{z}_{[,i]} = n - 1$ and $\mathbf{z}_{[,i]}' \mathbf{z}_{[,j]} = (n-1) r_{ij}$, for all $i \neq j$ and where r_{ij} is the sample correlation between $\mathbf{z}_{[,i]}$

and $\mathbf{z}_{[,j]}$. Since

$$\hat{\beta}_i^{-1} = \frac{1}{\lambda_n}\mathbf{z}'_{[,i]}\zeta(\hat{\boldsymbol{\eta}}) \quad \text{and} \quad \hat{\beta}_j^{-1} = \frac{1}{\lambda_n}\mathbf{z}'_{[,j]}\zeta(\hat{\boldsymbol{\eta}}),$$

we have

$$\begin{aligned}
\left|\hat{\beta}_i^{-1} - \hat{\beta}_j^{-1}\right| &= \left|\frac{1}{\lambda_n}\mathbf{z}'_{[,i]}\zeta(\hat{\boldsymbol{\eta}}) - \frac{1}{\lambda_n}\mathbf{z}'_{[,j]}\zeta(\hat{\boldsymbol{\eta}})\right| \\
&= \left|\frac{1}{\lambda_n}(\mathbf{z}_{[,i]} - \mathbf{z}_{[,j]})'\zeta(\hat{\boldsymbol{\eta}})\right| \\
&\leq \frac{1}{\lambda_n}\left\|(\mathbf{z}_{[,i]} - \mathbf{z}_{[,j]})\right\| \|\zeta(\hat{\boldsymbol{\eta}})\| \\
&\leq \frac{1}{\lambda_n}\sqrt{2\{(n-1) - (n-1)r_{ij}\}}\sqrt{n(1+e_n)^2}
\end{aligned}$$

for any $\hat{\beta}_i \neq 0$ and $\hat{\beta}_j \neq 0$. $\square$

A4.5 Proof of Theorem 4.3

Because $\hat{\boldsymbol{\beta}}$ is a fixed point of $g(\cdot)$ or $\hat{\boldsymbol{\beta}} = g(\hat{\boldsymbol{\beta}})$, we have for $j = 1, \ldots, p$

$$\{\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda_n D(\hat{\boldsymbol{\beta}})\} \left(\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ 0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ 0 \end{bmatrix} \right) = \tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{4.60}$$

Alternative, one can rewrite (4.60) as

$$\{D(\hat{\boldsymbol{\beta}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda_n \mathbf{I}_p\} \left(\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ 0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ 0 \end{bmatrix} \right) = D(\hat{\boldsymbol{\beta}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{4.61}$$

By extracting the jth element of (4.61), we have

$$\tilde{\mathbf{x}}_j' \sum_{i \neq j} \tilde{\mathbf{x}}_i \hat{\beta}_i^3 + \lambda_n \cdot 0 + \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j \cdot \hat{\beta}_j^3 + \lambda_n \hat{\beta}_j = \tilde{\mathbf{x}}_j'\tilde{\mathbf{y}}\hat{\beta}_j^2, \tag{4.62}$$

Letting $b_j^* = \tilde{\mathbf{x}}_j'(\tilde{\mathbf{y}} - \sum_{i \neq j} \tilde{\mathbf{x}}_i \hat{\beta}_i)$, simple algebra will allow us to rewrite (4.62) as

$$\hat{\beta}_j(\tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j \cdot \hat{\beta}_j^2 - b_j^* \hat{\beta}_j + \lambda_n) = 0, \tag{4.63}$$

which yields

$$\hat{\beta}_j = \begin{cases} 0, & \text{if } |b_j^*| < 2\sqrt{\tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j \lambda_n} \\ \frac{b_j^* + sign(b_j^*)\sqrt{(b_j^*)^2 - 4\tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j\lambda_n}}{2\tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j}, & \text{otherwise.} \end{cases} \tag{4.64}$$

for $j = 1, ..., p$. $\square$

A4.6 Proof of Lemma 4.1

Recall that, for the PSH model, $\tilde{w}_{ik} = \hat{G}(X_i)/\hat{G}(X_k \wedge X_i)$. Because $R_i = \{y : (X_y \geq X_i) \cup (X_y \leq X_i \cap \epsilon_y = 2)\}$, $k \in R_i$ implies that either $k \in \{y : (X_y \geq X_i)\}$ or $k \in \{y : (X_y \leq X_i \cap \epsilon_y = 2)\}$. If $k \in \{y : (X_y \geq X_i)\}$, then $\tilde{w}_{ik} = \hat{G}(X_i)/\hat{G}(X_i) = 1$. If $k \in \{y : (X_y \leq X_i \cap \epsilon_y = 2)\}$, then $\tilde{w}_{ik} = \hat{G}(X_i)/\hat{G}(X_k)$. Therefore

$$\begin{aligned}\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k) &= \sum_{k \in R_i(1)} \tilde{w}_{ik} \exp(\eta_k) + \sum_{k \in R_i(2)} \tilde{w}_{ik} \exp(\eta_y) \\ &= \sum_{k \in R_i(1)} \exp(\eta_k) + \hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k) / \hat{G}(X_k),\end{aligned}$$

where $R_i(1) = \{y : (X_y \geq X_i)\}$ and $R_i(2) = \{y : (X_y < X_i \cap \epsilon_y = 2)\}$.

A4.7 BAR implementation via CCD

Algorithm 2: The BAR algorithm using cyclic coordinate descent optimization

```
Set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{ridge}$;
for $k = 1, 2, \ldots$ do
    $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}^{(k-1)}$;
    for $s = 1, 2, \ldots$ do
        # Enter cyclic coordinate descent
        for $j = 1, \ldots p$ do
            Calculate $c_{1j} = \dot{l}_j(\boldsymbol{\beta}^{(s-1)})$ and $c_{2j} = -\ddot{l}_{jj}(\boldsymbol{\beta}^{(s-1)})$;
            $\boldsymbol{\beta}_j^{(s)} = (c_{2j}\beta_j^{(s-1)} + c_{1j})/\{c_{2j} + \lambda_n/(\hat{\boldsymbol{\beta}}_j^{(k-1)})^2\}$;
        end
        if $\left\|\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s-1)}\right\| < tol_1$ then
            $\hat{\boldsymbol{\beta}}^{(k)} = \boldsymbol{\beta}^{(s)}$ and break;
        end
    end
    if $\left\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\right\| < tol_2$ then
        $\hat{\boldsymbol{\beta}}_{BAR} = \hat{\boldsymbol{\beta}}^{(k)}$ and break ;
    end
end
$\hat{\boldsymbol{\beta}}_{BAR} = \hat{\boldsymbol{\beta}}_{BAR} \times I(|\hat{\boldsymbol{\beta}}_{BAR}| > \epsilon^*)$ # Induce sparsity;
```

A4.8 Additional figures and tables

A4.8.1 Figures

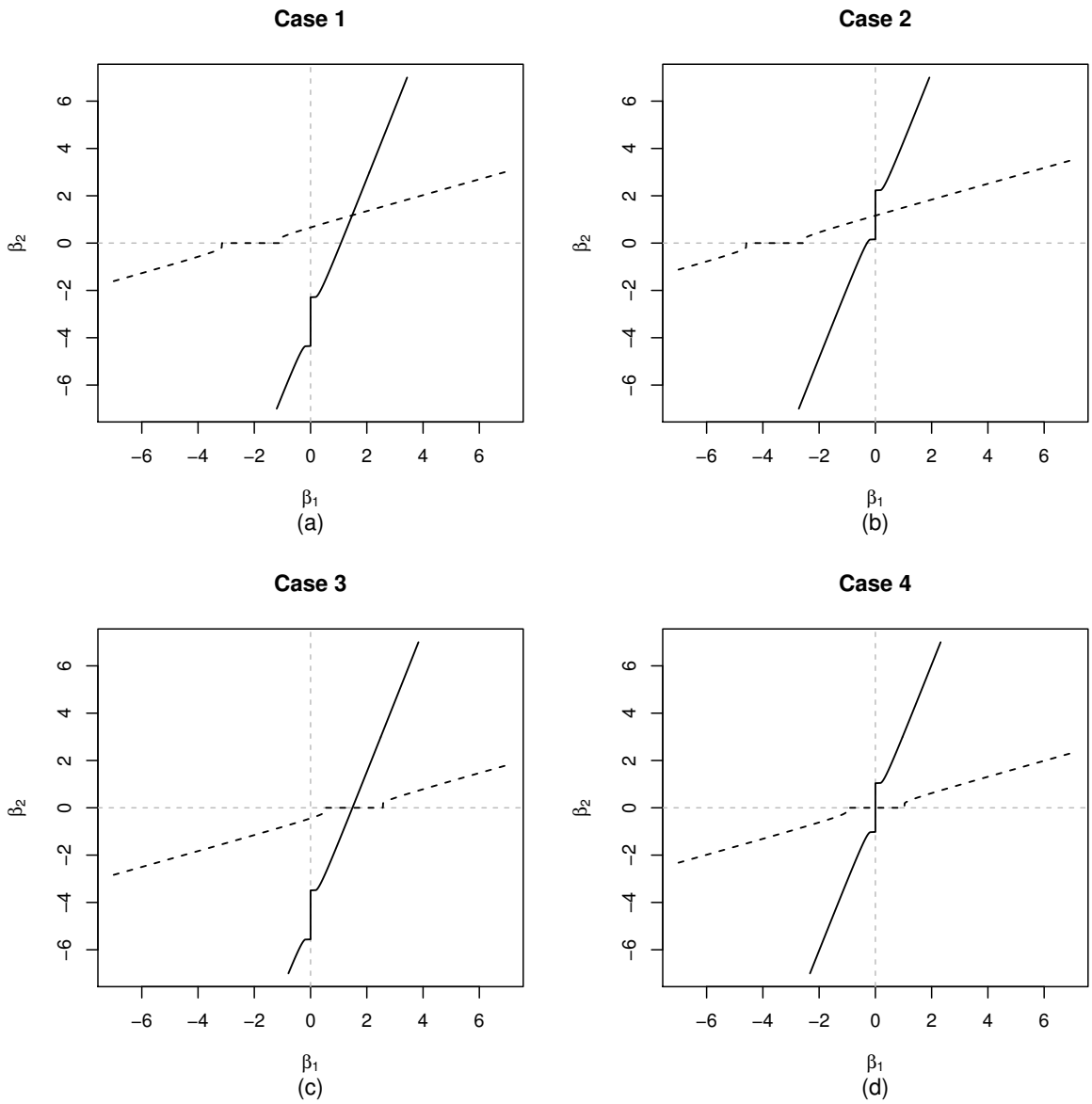

Figure A4.1: Graphs of $\beta_1 = g_1(\beta_2)$ (solid line) and $\beta_2 = g_2(\beta_1)$ (dotted line) under selected scenarios, which by Theorem 2, intersect at the fixed-point of $g(\beta_1, \beta_2)$.

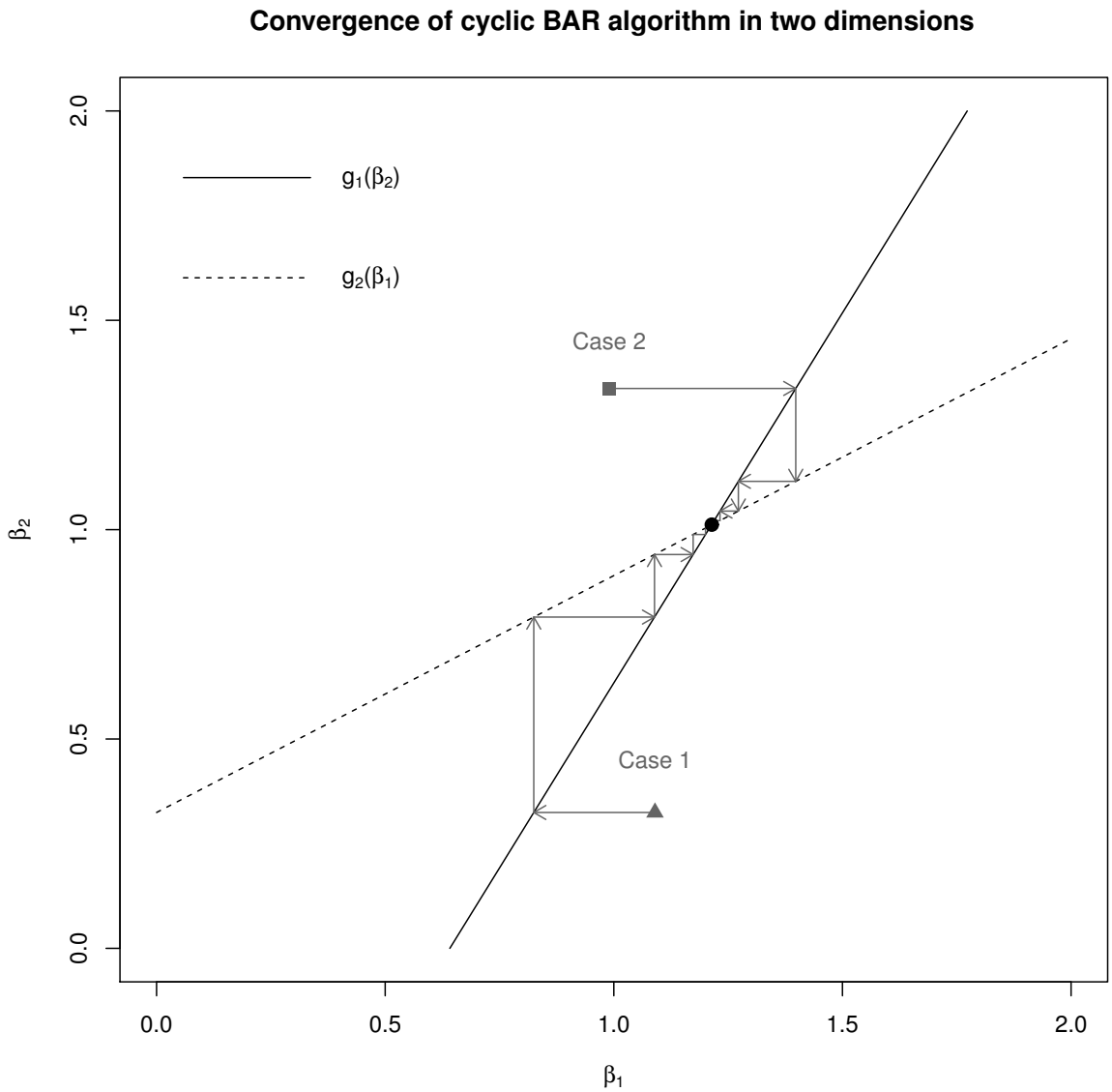

Figure A4.2: An illustration of the CYCBAR algorithm in a zoomed in picture of Figure S1(a). The BAR estimator is the fixed point of $g(\beta_1, \beta_2)$, which, by Theorem 2, is the intersection of $\beta_1 = g_1(\beta_2)$ and $\beta_2 = g_2(\beta_1)$.

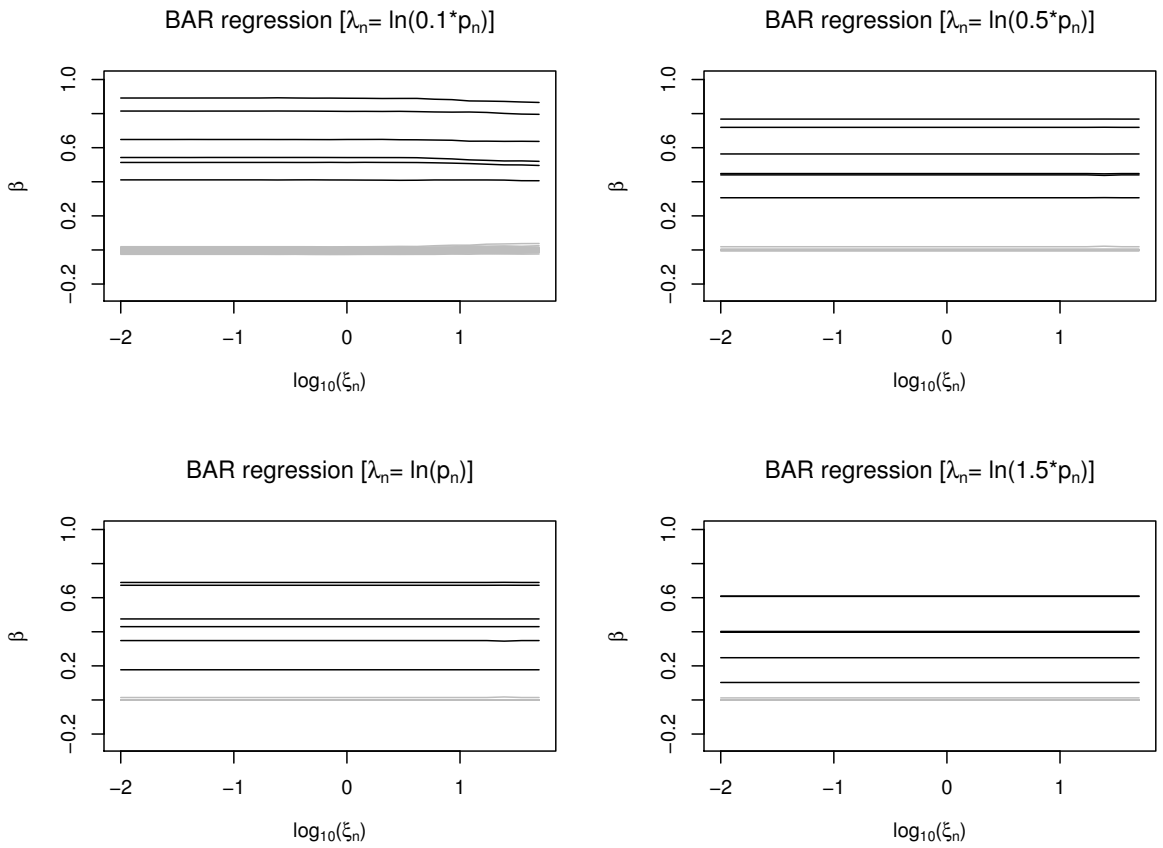

Figure A4.3: Path plot for BAR regression with varying ξ_n and several fixed values of λ_n where $n = 300$ and $p_n = 40$. The path plots are averaged over 100 simulations.

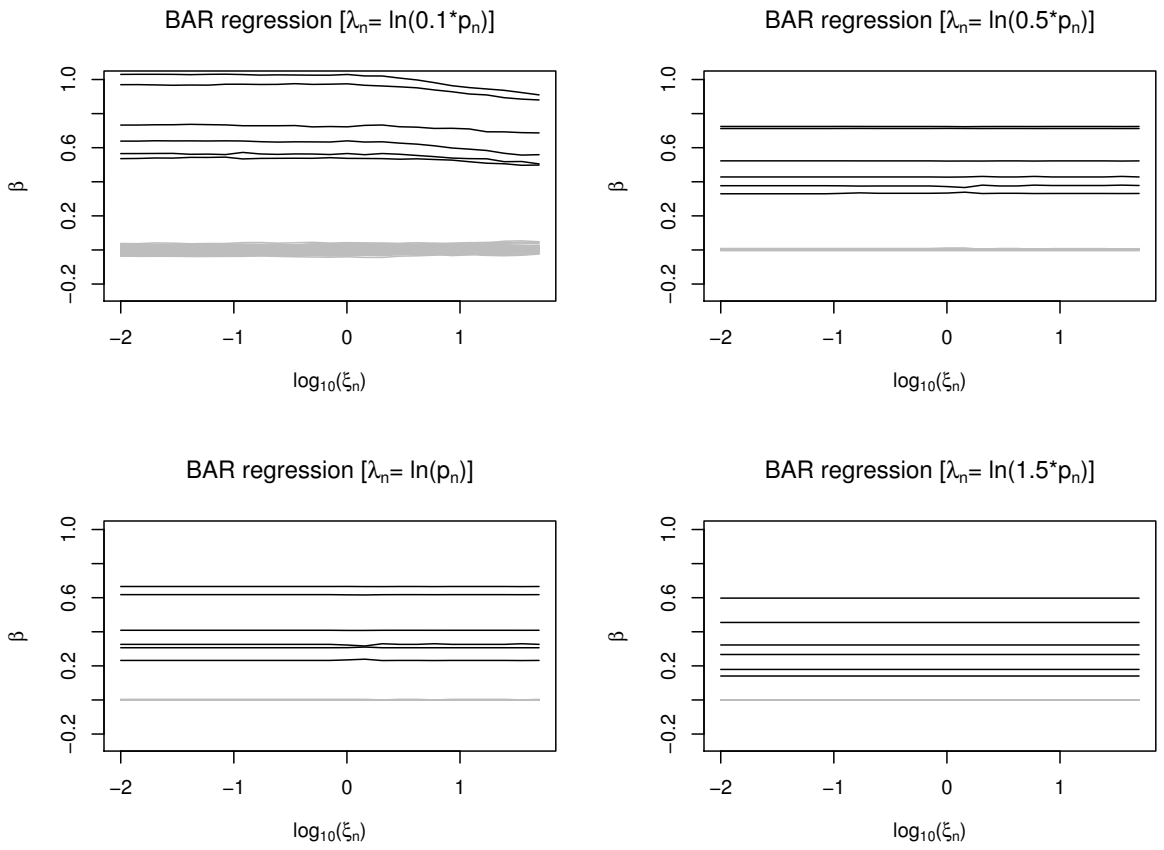

Figure A4.4: Path plot for BAR regression with varying ξ_n and several fixed values of λ_n where $n = 300$ and $p_n = 100$. The path plots are averaged over 100 simulations.

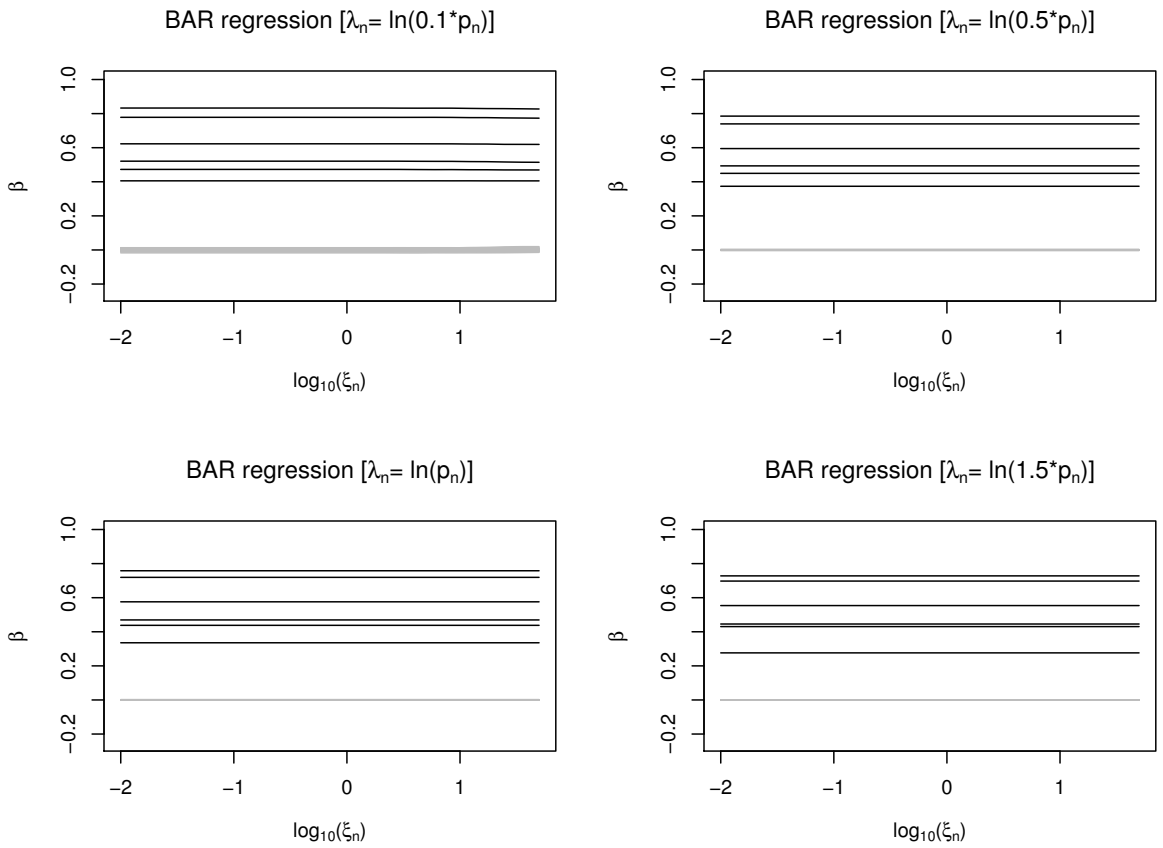

Figure A4.5: Path plot for BAR regression with varying ξ_n and several fixed values of λ_n where $n = 700$ and $p_n = 40$. The path plots are averaged over 100 simulations.

A4.8.2 Tables

Table A4.1: Additional simulation results for model comparison. Based on 100 replications with $\rho = 0.5$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^*, \mathbf{0}_{p_n-10})$ where $\boldsymbol{\beta}^* = (0.40, 0.45, 0, 0.50, 0, 0.60, 0.75, 0, 0, 0.80)$, censoring rate $\approx 33\%$ and type 1 event rate $\approx 41\%$.

	$n = 300; p = 100$				$n = 700; p = 100$			
Method	MSB	FN	FP	SM	MSB	FN	FP	SM
ORACLE	0.09	0.00	0.00	1.00	0.04	0.00	0.00	1.00
BAR(ξ_n, λ_n)	0.31	0.40	1.87	0.85	0.06	0.01	0.89	0.94
BAR(λ_n)	0.32	0.49	1.70	0.85	0.06	0.01	0.86	0.94
BAR$_{EBIC}$	0.51	1.68	0.03	0.84	0.10	0.25	0.00	0.98
LASSO	0.44	0.10	2.82	0.83	0.21	0.00	2.49	0.85
ALASSO	0.39	0.75	2.00	0.81	0.09	0.00	0.73	0.95
SCAD	0.43	0.33	2.73	0.82	0.12	0.02	1.39	0.91
MCP	0.37	0.56	1.89	0.84	0.08	0.08	0.65	0.95

Table A4.2: Additional simulation results for model comparison. Based on 100 replications with $\rho = 0.5$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^*, \mathbf{0}_{p_n-10})$ where $\boldsymbol{\beta}^* = (0.40, 0.45, 0, 0.50, 0, 0.60, 0.75, 0, 0, 0.80)$, censoring rate $\approx 33\%$ and type 1 event rate $\approx 32\%(\pi = 0.4)$ and $\approx 43\%(\pi = 0.75)$.

	$n = 700; p = 100; \pi = 0.4$				$n = 700; p = 100; \pi = 0.75$			
Method	MSB	FN	FP	SM	MSB	FN	FP	SM
ORACLE	0.04	0.00	0.00	1.00	0.03	0.00	0.00	1.00
BAR(ξ_n, λ_n)	0.08	0.03	0.88	0.94	0.04	0.00	0.65	0.96
BAR(λ_n)	0.08	0.04	0.84	0.94	0.05	0.00	0.67	0.95
BAR$_{EBIC}$	0.12	0.36	0.00	0.97	0.05	0.07	0.00	0.99
LASSO	0.23	0.00	2.63	0.85	0.18	0.00	2.51	0.86
ALASSO	0.11	0.06	0.94	0.93	0.06	0.00	0.58	0.96
SCAD	0.15	0.08	1.44	0.90	0.09	0.00	0.96	0.93
MCP	0.11	0.13	0.73	0.94	0.06	0.02	0.35	0.97

Table A4.3: Additional simulation results for model comparison. Based on 100 replications with $\rho = 0.5$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \mathbf{0}_{p_n-30})$ where $\boldsymbol{\beta}^* = (0.40, 0.45, 0, 0.50, 0, 0.60, 0.75, 0, 0, 0.80)$, censoring rate $\approx 33\%$ and type 1 event rate $\approx 41\%$.

	$n = 300; p = 100$				$n = 700; p = 100$			
Method	MSB	FN	FP	SM	MSB	FN	FP	SM
ORACLE	0.40	0.00	0.00	1.00	0.13	0.00	0.00	1.00
BAR(ξ_n, λ_n)	0.84	0.24	3.60	0.91	0.16	0.01	1.38	0.96
BAR(λ_n)	0.79	0.36	3.27	0.91	0.16	0.01	1.33	0.97
BAR$_{EBIC}$	4.41	7.77	0.01	0.75	0.33	0.08	0.00	1.00
LASSO	2.32	0.05	11.02	0.79	1.27	0.00	11.92	0.78
ALASSO	1.21	0.57	6.35	0.85	0.32	0.00	2.40	0.94
SCAD	0.98	0.19	7.00	0.85	0.16	0.01	1.54	0.96
MCP	1.03	0.33	3.59	0.91	0.15	0.02	0.73	0.98

Table A4.4: Additional information about the USRDS subset. Summary of event count (%) observed for the training ($n = 125,000$) and test ($n = 100,000$) sets for the USRDS subset. (Disc: Discontinued dialysis; Recov: Renal function recovery; RC: Right censored including loss-to-follow up and end of study time.)

Set	Transplant	Death	Disc.	Recov.	RC	Total
Training	11,943(10%)	60,175 (48%)	8,160 (6%)	7,555 (6%)	37,167 (30%)	125,000 (100%)
Test	9,642 (10%)	47,830 (48%)	6,459 (7%)	6,057 (6%)	30,012 (29%)	100,000 (100%)

CHAPTER 5

Fast and scalable Fine-Gray regression and cumulative incidence function estimation

This chapter extends the forward-backward scan introduced in Section 4.1.4 for parameter and cumulative incidence function estimation of unpenalized and penalized Fine-Gray regression. The chapter is organized as follows. We briefly review the basic definition of the Fine-Gray proportional subdistribution hazards model, the CIF, and penalized Fine-Gray regression. We introduce our forward-backward scan algorithm in Section 5.2. Then, in Section 5.3 we describe the main functionalities of the **fastcmprsk** package that we developed for R which utilizes the aforementioned algorithm, which include unpenalized and penalized parameter estimation and CIF estimation. We perform simulation studies in Section 5.4 to compare the performance of our proposed method to some of their popular competitors. Section 5.5 provides an illustration on real data using a subset of the United States Renal Data Systems. The **fastcmprsk** package is available at `https:github.com/erickawaguchi/fastcmprsk`.

5.1 Data structure and model

Recall that for subject $i = 1, \ldots, n$, let T_i, C_i, and ϵ_i be the event time, possible right-censoring time, and cause (event type), respectively. Without loss of generality assume there are two event types $\epsilon \in \{1, 2\}$ where $\epsilon = 1$ is the event of interest (or primary event) and $\epsilon = 2$ is the competing risk. With the presence of right-censoring, we generally observe $X_i = T_i \wedge C_i$, $\delta_i = I(T_i \leq C_i)$, where $a \wedge b = \min(a, b)$ and $I(\cdot)$ is the indicator function. Letting $\mathbf{z}_i$ be a p-dimensional vector of time-independent subject-specific covariates, competing risks data consist of the following independent and identically distributed quadruplets

$\{(X_i, \delta_i, \delta_i\epsilon_i, \mathbf{z}_i)\}_{i=1}^n$. Assume that there also exists a τ such that 1) for some arbitrary time t, $t \in [0, \tau]$; 2) $\Pr(T_i > \tau) > 0$ and $\Pr(C_i > \tau) > 0$ for all $i = 1, \ldots, n$, and that for simplicity, no ties are observed.

The CIF for the primary event conditional on the covariates $\mathbf{z} = (z_1, \ldots, z_p)$ is $F_1(t; \mathbf{z}) = \Pr(T \leq t, \epsilon = 1|\mathbf{z})$. To model the covariate effects on $F_1(t; \mathbf{z})$, Fine and Gray (1999) introduced the now well-appreciated proportional subdistribution hazards (PSH) model:

$$h_1(t|\mathbf{z}) = h_{10}(t) \exp(\mathbf{z}'\boldsymbol{\beta}), \tag{5.1}$$

where

$$\begin{aligned} h_1(t|\mathbf{z}) &= \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T \leq t + \Delta t, \epsilon = 1 | T \geq t \cup (T \leq t \cap \epsilon \neq 1), \mathbf{z}\}}{\Delta t} \\ &= -\frac{d}{dt} \log\{1 - F_1(t; \mathbf{z})\} \end{aligned}$$

is a subdistribution hazard (Gray, 1988), $h_{10}(t)$ is a completely unspecified baseline subdistribution hazard, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. As Fine and Gray (1999) mentioned, the risk set associated with $h_1(t; \mathbf{z})$ is somewhat counterfactual as it includes subjects who are still at risk ($T \geq t$) and those who have already observed the competing risk prior to time t ($T \leq t \cap \epsilon \neq 1$). However, this construction is useful for direct modeling of the CIF.

5.1.1 Parameter estimation for unpenalized Fine-Gray regression

Parameter estimation and large-sample inference of the PSH model follows from the log-pseudo likelihood:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty \left[\boldsymbol{\beta}'\mathbf{z}_i - \ln \left\{ \sum_k \hat{w}_k(u) Y_k(u) \exp\left(\mathbf{z}_k'\boldsymbol{\beta}\right) \right\} \right] \hat{w}_i(u) dN_i(u), \tag{5.2}$$

where $N_i(t) = I(X_i \leq t, \epsilon_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, and $\hat{w}_i(t)$ is a time-dependent weight based on the inverse probability of censoring weighting (IPCW) technique (Robins and Rot-

nitzky, 1992). To parallel Fine and Gray (1999), we define the IPCW for subject i at time t as $\hat{w}_i(t) = I(C_i \geq T_i \wedge t)\hat{G}(t)/\hat{G}(X_i \wedge t)$, where $G(t) = \Pr(C \geq t)$ is the survival function of the censoring variable C and $\hat{G}(t)$ is the Kaplan-Meier estimate for $G(t)$. However, we can generalize the IPCW to allow for dependence between C and $\mathbf{z}$.

Let $\hat{\boldsymbol{\beta}}_{mple} = \arg\min_{\boldsymbol{\beta}}\{-l(\boldsymbol{\beta})\}$ be the maximum pseudo likelihood estimator of $\boldsymbol{\beta}$. Fine and Gray (1999) investigate the large-sample properties of $\hat{\boldsymbol{\beta}}_{mple}$ and prove that, under certain regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{mple} - \boldsymbol{\beta}_0) \to N(0, \boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{-1}), \tag{5.3}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$ is the limit of the negative of the partial derivative matrix of the score function evaluated at $\boldsymbol{\beta}_0$, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the limiting distribution of the score function. The package **cmprsk** implements this estimation procedure.

5.1.2 Estimating the cumulative incidence function

An alternative interpretation of the coefficients from the Fine-Gray model is to model their effect on the CIF. Using a Breslow-type estimator (Breslow, 1974), we can obtain a consistent estimate for $H_{10}(t) = \int_0^t h_{10}(s)ds$ through

$$\hat{H}_{10}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{1}{\hat{S}^{(0)}(\hat{\boldsymbol{\beta}}, u)}\hat{w}_i(u)dN_i(u),$$

where $\hat{S}^{(0)}(\hat{\boldsymbol{\beta}}, u) = n^{-1}\sum_{i=1}^{n}\hat{w}_i(u)Y_i(u)\exp(\mathbf{z}_i'\hat{\boldsymbol{\beta}})$. The predicted CIF, conditional on $\mathbf{z} = \mathbf{z}_0$, is then

$$\hat{F}_1(t; \mathbf{z}_0) = 1 - \exp\left\{\int_0^t \exp(\mathbf{z}_0'\hat{\boldsymbol{\beta}})d\hat{H}_{10}(u)\right\}.$$

We refer the readers to Appendix B of Fine and Gray (1999) for the large-sample properties of $\hat{F}_1(t; \mathbf{z}_0)$. The quantities needed to estimate $\int_0^t d\hat{H}_{10}(u)$ are already precomputed when

estimating $\hat{\boldsymbol{\beta}}$. Fine and Gray (1999) proposed a resampling approach to calculate confidence intervals and confidence bands for $\hat{F}_1(t; \mathbf{z}_0)$.

5.1.3 Penalized Fine-Gray regression for variable selection

Oftentimes, reserachers are interested in identifying which covariates have an effect on the CIF. Penalization methods (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Zhang et al., 2010) offer a popular way to perform variable selection and parameter estimation simultaneously through minimizing the objective function

$$Q(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{5.4}$$

where $l(\boldsymbol{\beta})$ is defined in (5.2), $p_\lambda(|\beta_j|)$ is a penalty function where the sparsity of the model is controlled by the non-negative tuning parameter λ. Fu et al. (2017) recently extend several popular variable selection procedures - LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), and MCP (Zhang, 2010) - to the Fine-Gray model, explore its asymptotic properties under fixed model dimension, and develop the R package **crrp** (Fu, 2016) for implementation. Parameter estimation in the **crrp** package employs a cyclic coordinate algorithm.

The sparsity of the model depends heavily on the choice of the tuning parameters. Practically, finding a suitable (or optimal) tuning parameter involves applying a penalization method over a sequence of possible candidate values of λ and finding the λ that minimizes some metric such as the Bayesian information criterion (Schwarz, 1978) or generalized cross validation measure (Craven and Wahba, 1978). A more thorough discussion on tuning parameter selection can partially be found in Wang et al. (2007); Zhang et al. (2010); Wang and Zhu (2011); Fan and Tang (2013); Fu et al. (2017); Ni and Cai (2018).

5.2 Forward-backward scan for parameter estimation

This section discusses a novel forward-backward scan algorithm that reduces the computational complexity associated with parameter estimation from $O(n^2)$ to $O(n)$. Commonly-used optimization routines generally require the calculation of the log-pseudo likelihood (5.2), the score function

$$\dot{l}_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) z_{ij} - \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) \frac{\sum_{k \in R_i} z_{kj} \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)}, \tag{5.5}$$

and, in some cases, the Hessian diagonals

$$\ddot{l}_{jj}(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) \left[\frac{\sum_{k \in R_i} z_{kj}^2 \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)} - \left\{ \frac{\sum_{k \in R_i} z_{kj} \tilde{w}_{ik} \exp(\eta_k)}{\sum_{k \in R_i} \tilde{w}_{ik} \exp(\eta_k)} \right\}^2 \right], \tag{5.6}$$

where

$$\tilde{w}_{ik} = \hat{w}_k(X_i) = \hat{G}(X_i)/\hat{G}(X_i \wedge X_k), \quad k \in R_i,$$

$R_i = \{y : (X_y \geq X_i) \cup (X_y \leq X_i \cap \epsilon_y = 2)\}$ and $\eta_k = \mathbf{z}_k' \boldsymbol{\beta}$ for use within cyclic coordinate descent. Direct calculations using the above formulas will need $O(n^2)$ operations due to the double summations, that becomes computationally taxing for large n. Below we will show how to calculate the double summation linearly, allowing us to compute (5.2), (5.5), and (5.6) in $O(n)$ time.

Before proceeding with the algorithm, we first define what we mean by a forward and backward scan. A forward (prefix) scan maps $\{a_1, a_2, \ldots, a_n\} \mapsto \{a_1, a_1 + a_2, \ldots, \sum_{i=1}^{n} a_i\}$; whereas a backward (prefix) scan maps to $\{\sum_{i=1}^{n} a_i, \sum_{i=2}^{n} a_i, \ldots, a_1\}$. First, note that R_i partitions into two disjoint subsets: $R_i(1) = \{y : X_y \geq X_i\}$ and $R_i(2) = \{y : (X_y \leq X_i \cap \epsilon_y = 2)\}$. Here $R_i(1)$ is the set of observations that have an observed event time after X_i and $R_i(2)$ is the set of observations that have observed the competing event before time X_i. Further, $\tilde{w}_{ik} = 1$ if $k \in R_i(1)$ and $\tilde{w}_{ik} = \hat{G}(X_i)/\hat{G}(X_k)$, if $k \in R_i(2)$. Since $R_i(1)$ and $R_i(2)$ are disjoint, we can write the double summation of, for example, the score function

(5.5) as

$$\sum_{i=1}^{n} I(\delta_i \epsilon_i = 1) \frac{\sum_{k \in R_i(1)} z_{kj} \exp(\eta_k) + \hat{G}(X_i) \sum_{k \in R_i(2)} z_{kj} \exp(\eta_k)/\hat{G}(X_k)}{\sum_{k \in R_i(1)} \exp(\eta_k) + \hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k)/\hat{G}(X_k)}. \tag{5.7}$$

We will first tackle the denominator term $\sum_{k \in R_i(1)} \exp(\eta_k) + \hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k)/\hat{G}(X_k)$. If we arrange the observed event times in decreasing order, we see that $\sum_{k \in R_i(1)} \exp(\eta_k)$ is a series of cumulative sums. For example, given $X_i > X_{i'}$, the set $R_{i'}(1)$ consists of the observations from $R_i(1)$ and the set of observations $\{y : X_y \in [X_{i'}, X_i)\}$, therefore $\sum_{k \in R_{i'}(1)} \exp(\eta_k) = \sum_{k \in R_i(1)} \exp(\eta_k) + \sum_{k \in \{y : X_y \in [X_{i'}, X_i)\}} \exp(\eta_k)$ and thus calculating $\sum_{k \in R_i(1)} \exp(\eta_k)$ for all $i = 1, \ldots, n$ requires $O(n)$ calculations in total. However, $\hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k)/\hat{G}(X_k)$ does not monotonically increase as the event times decrease. Instead, we observe that $\hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k)/\hat{G}(X_k)$ is a series of cumulative sums as the event times increase. Thus calculating the denominator term will requires two scans: one forward scan going forward from largest observed event time to smallest to calculate $\sum_{k \in R_i(1)} \exp(\eta_k)$ and one backward scan from smallest observed event time to largest to calculate $\hat{G}(X_i) \sum_{k \in R_i(2)} \exp(\eta_k)/\hat{G}(X_k)$. Likewise, we calculate both $\sum_{k \in R_i} z_{kj} \exp(\eta_k)$ and $\sum_{k \in R_i} z_{kj}^2 \exp(\eta_k)$ in linear time since the terms z_{kj} and z_{kj}^2 are multiplicative constants that do not affect the cumulative structures of the summations. As a consequence, the ratio in the double summation is available in $O(n)$ time.

Furthermore, the outer summation of subjects who observe the event of interest is also a cumulative sum since, provided that $X_i > X_{i'}$ and both $\delta_i = 1$ and $\delta_{i'} = 1$,

$$\sum_{l=1}^{i} I(\delta_l \epsilon_l = 1) \frac{\sum_{k \in R_l} z_{kj} \exp(\eta_k)}{\sum_{k \in R_l} \exp(\eta_k)} = \sum_{l=1}^{i'} I(\delta_l \epsilon_l = 1) \frac{\sum_{k \in R_l} z_{kj} \exp(\eta_k)}{\sum_{k \in R_l} \exp(\eta_k)} \tag{5.8}$$

$$+ I(\delta_i \epsilon_i = 1) \frac{\sum_{k \in R_i} z_{kj} \exp(\eta_k)}{\sum_{k \in R_i} \exp(\eta_k)}, \tag{5.9}$$

that also only requires $O(n)$ calculations since the ratios are precomputed in $O(n)$ calculations and thus the score function (5.5) can be calculated in linear time. Similarly, both the log-pseudo likelihood (5.2) and the diagonal elements of the Hessian (5.6) are also calculated

Table 5.1: Currently available functions in **fastcmprsk** (as of May 15, 2019).

Function name	Basic description
Modeling functions	
fastCrr	Fits unpenalized Fine-Gray regression
fastCrrp	Fits penalized Fine-Gray regression
Utilities	
summary	Returns ANOVA table from **fastCrr** output
predict	Estimates CIF given a vector of covariates
plot	Plots output (object dependent)
varianceControl	Options for bootstrap variance
simulateTwoCauseFineGrayModel	Simulates two-cause competing risks data

linearly.

5.3 The fastcmprsk package

We utilize this forward-backward scan algorithm for both penalized and unpenalized parameter estimation for the Fine-Gray model in linear time. Furthermore, we also develop scalable methods to estimate the predicted CIF and its corresponding confidence interval/band. For convenience to researchers and readers, we further include a function to simulate two-cause competing risks data. Table 5.1 provides a quick summary of the currently available functions provided in **fastcmprsk**. We briefly detail the use of these functions below.

5.3.1 Simulating competing risks data

Researchers can simulate two-cause competing risks data using the `simulateTwoCauseFineGrayModel` function in **fastcmprsk**. The data generation scheme follows a similar design to that of Fine and Gray (1999) and Fu et al. (2017). Given a design matrix $\mathbf{Z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$, let the cumulative incidence function for cause 1 (the event of interest) be defined as $F_1(t; \mathbf{z}_i) = \Pr(T_i \leq t, \epsilon_i = 1 | \mathbf{z}_i) = 1 - [1 - \pi\{1 - \exp(-t)\}]^{\exp(\mathbf{z}_i'\boldsymbol{\beta}_1)}$, which is a unit exponential mixture with mass $1 - \pi$ at ∞ when $\mathbf{z}_i = \mathbf{0}$ and where π controls the cause 1 event rate. The cumulative incidence function for cause 2 is obtained by setting $\Pr(\epsilon_i = 2 | \mathbf{z}_i) = 1 - \Pr(\epsilon_i = 1 | \mathbf{z}_i)$ and then using an exponential

distribution with rate $\exp(\mathbf{z}_i'\boldsymbol{\beta}_2)$ for the conditional cumulative incidence function $\Pr(T_i \leq t|\epsilon_i = 2, \mathbf{z}_i)$. Censoring times are independently generated from a uniform distribution $U(u_{\min}, u_{\max})$ where $u_{\min}$ and $u_{\max}$ control the censoring percentage. Appendix A5.1 provides more details on the data generation process. Below is a toy example of simulating competing risks data where $n = 500$, $\boldsymbol{\beta}_1 = (0.40, -0.40, 0, -0.50, 0, 0.60, 0.75, 0, 0, -0.80)$, $\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_1$, $u_{\min} = 0$, $u_{\max} = 1$, $\pi = 0.5$, and where $\mathbf{Z}$ is simulated from a multivariate standard normal distribution with unit variance. This simulated dataset will be used to illustrate the use of the different modeling functions within **fastcmprsk**.

```
R> library(fastcmprsk)
R> set.seed(2019)
R> nobs <- 500
R> beta1 <- c(0.40, -0.40,  0, -0.50,  0,  0.60,  0.75,  0,  0, -0.80)
R> beta2 <- -beta1
R> Z <- matrix(rnorm(nobs * length(beta1)), nrow = nobs)
R> dat <- simulateTwoCauseFineGrayModel(nobs, beta1, beta2,
+ Z, u.min = 0, u.max = 1, p = 0.5)
R> table(dat$fstatus) # Event counts

  0   1   2
241 118 141

R> head(dat$ftime) # First 6 observed survival times

[1] 0.098345608 0.008722629 0.208321175 0.017656904 0.495185038 0.222799124
```

5.3.2 Unpenalized parameter estimation and inference

We first illustrate the coefficient estimation from (5.1) using the Fine-Gray log-pseudo likelihood. The `fastCrr` function estimates these parameters using our forward-backward scan

algorithm and is functionally similar to the `crr` function from the **cmprsk** package.

```
# cmprsk package
R> fit1 <- cmprsk::crr(dat$ftime, dat$fstatus, Z, failcode = 1, cencode = 0,
+                      variance = FALSE)

# fastcmprsk package
R> fit2 <- fastcmprsk::fastCrr(dat$ftime, dat$fstatus, Z,
+                              failcode = 1, cencode = 0,
+       variance = FALSE, returnDataFrame = TRUE)

R> max(abs(fit1$coef - fit2$coef))

[1] 8.534242e-08
```

As expected, the `fastCrr` function calculates nearly identical parameter estimates to the `crr` function. We compare the runtime performance between these two methods in Section 5.4.1.

We now show how to obtain the variance-covariance matrix for the parameter estimates. The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ can not be directly estimated using the `fastCrr` function. First, the asymptotic expression requires estimating both $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$, which can not be trivially calculated in linear time. Second, for large-scale data where both n and p can be large, matrix calculations, storage, and inversion can be computationally prohibitive. Instead, we propose to estimate the variance-covariance matrix using the bootstrap (Efron, 1979). Let $\tilde{\boldsymbol{\beta}}^{(1)}, \ldots \tilde{\boldsymbol{\beta}}^{(B)}$ be bootstrapped parameter estimates obtained by resampling subjects with replacement from the original data B times. Unless otherwise noted, the size of each resample is the same as the original data. For $j = 1, \ldots, p$ and $k = 1, \ldots, p$, we can estimate the covariance between $\hat{\beta}_j$ and $\hat{\beta}_k$ by

$$\widehat{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \frac{1}{B-1} \sum_{b=1}^{B} (\tilde{\beta}_j^{(b)} - \bar{\beta}_j)(\tilde{\beta}_k^{(b)} - \bar{\beta}_k), \tag{5.10}$$

where $\bar{\boldsymbol{\beta}}_j = \frac{1}{B}\sum_{b=1}^{B}\tilde{\boldsymbol{\beta}}_j^{(b)}$. Therefore, with $\hat{\sigma}_j^2 = \widehat{Cov}(\hat{\beta}_j, \hat{\beta}_j)$, a $(1-\alpha)\times 100\%$ confidence interval for $\boldsymbol{\beta}_j$ is given by

$$\hat{\boldsymbol{\beta}}_j \pm z_{1-\alpha/2}\hat{\sigma}_j, \tag{5.11}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha)\times 100th$ percentile of the standard normal distribution. Since parameter estimation for the Fine-Gray model can be done in linear time using our forward-backward scan algorithm, the collection of parameter estimates obtained by bootstrapping can also be obtained linearly. The `varianceControl` function controls the parameters used for bootstrapping, that one then passes into the `var.control` argument in `fastCrr`.

```
R> vc   <- varianceControl(B = 100, seed = 2019)
R> fit3 <- fastcmprsk::fastCrr(dat$ftime, dat$fstatus, Z,
+                              failcode = 1, cencode = 0, variance = TRUE,
+                              var.control = vc, returnDataFrame = TRUE)
# returnDataFrame = TRUE is necessary for CIF estimation (next section)

R> sqrt(diag(fit3$var))

 [1] 0.099 0.096 0.099 0.104 0.099 0.113 0.103 0.097 0.104 0.135

R> summary(fit3, conf.int = FALSE, digits = 2)

Fine-Gray Regression via fastcmprsk package.

Call:
fastcmprsk::fastCrr(dat$ftime, dat$fstatus, Z, failcode = 1,
    cencode = 0, variance = TRUE, var.control = vc, returnDataFrame = TRUE)

fastCrr converged in 24 iterations.
```

```
          coef exp(coef) se(coef)       z p-value
 [1,]  0.19228     1.212   0.0993  1.9358 5.3e-02
 [2,] -0.38640     0.679   0.0963 -4.0142 6.0e-05
 [3,]  0.01816     1.018   0.0988  0.1838 8.5e-01
 [4,] -0.39769     0.672   0.1042 -3.8169 1.4e-04
 [5,]  0.10571     1.111   0.0986  1.0724 2.8e-01
 [6,]  0.57494     1.777   0.1130  5.0895 3.6e-07
 [7,]  0.77884     2.179   0.1032  7.5478 4.4e-14
 [8,] -0.00611     0.994   0.0972 -0.0628 9.5e-01
 [9,] -0.06571     0.936   0.1040 -0.6315 5.3e-01
[10,] -0.99687     0.369   0.1346 -7.4054 1.3e-13

Pseudo Log-likelihood = -590
Null Pseudo Log-likelihood = -675
```

5.3.3 Cumulative incidence function and interval/band estimation

The CIF is also available in linear time in the **fastcmprsk** package. Fine and Gray (1999) propose a Monte Carlo simulation method for interval and band estimation. We implement a slightly different approach using bootstrapping for interval and band estimation in our package. Let $\tilde{F}_1^{(1)}(t;\mathbf{z}_0),\ldots,\tilde{F}_1^{(B)}(t;\mathbf{z}_0)$ be the bootstrapped predicted CIF obtained by resampling subjects with replacement from the original data B times and let $m(\cdot)$ be a known, monotone, and continuous transformation. In our current implementation we let $m(x)=\log\{-\log(x)\}$; however, we plan on incorporating other transformations in our future implementation. We first estimate the variance function $\sigma^2(t;\mathbf{z}_0)$ of the transformed CIF through

$$\hat{\sigma}^2(t;\mathbf{z}_0)=\frac{1}{B}\sum_{b=1}^{B}\left[m\{\tilde{F}_1^{(b)}(t;\mathbf{z}_0)\}-\bar{m}\{\tilde{F}_1(t;\mathbf{z}_0)\}\right]^2, \tag{5.12}$$

where $\bar{m}\{\tilde{F}_1(t;\mathbf{z}_0)\} = \frac{1}{B}\sum_{b=1}^{B} m\{\tilde{F}_1^{(b)}(t;\mathbf{z}_0)\}$. Using the functional delta method, we can now construct $(1-\alpha)\times 100\%$ confidence intervals for $F_1(t;\mathbf{z}_0)$ by

$$m^{-1}\left[m\{\hat{F}_1(t;\mathbf{z}_0)\} \pm z_{1-\alpha/2}\hat{\sigma}(t;\mathbf{z}_0)\right]. \tag{5.13}$$

Next we propose a symmetric global confidence band for the estimated CIF $\hat{F}_1(t;\mathbf{z}_0)$, $t \in [t_L, t_U]$ via bootstrap. We first determine a critical region $C_{1-\alpha}(\mathbf{z}_0)$ such that

$$\Pr\left\{\sup_{t\in[t_L,t_U]} \frac{|m\{\hat{F}_1(t;\mathbf{z}_0)\} - m\{F_1(t;\mathbf{z}_0)\}|}{\sqrt{\widehat{Var}[m\{\hat{F}_1(t;\mathbf{z}_0)\}]}} \leq C_{1-\alpha}(\mathbf{z}_0)\right\} = 1-\alpha. \tag{5.14}$$

While Equation (5.12) estimates $\widehat{Var}[m\{\hat{F}_1(t;\mathbf{z}_0)\}]$ we still need to find $C_{1-\alpha}(\mathbf{z}_0)$ by the bootstrap $(1-\alpha)^{th}$ percentile of the distribution of the supremum in the equation above. The algorithm is as follows:

1. Resample subjects with replacement from the original data B times and estimate $\tilde{F}_1^{(b)}(t;\mathbf{z}_0)$ for $b = 1,\ldots,B$ and $\hat{\sigma}^2(t;\mathbf{z}_0)$ using (5.12).

2. For the bth bootstrap sample , $b \in \{1,\ldots,B\}$, calculate

$$C^{(b)} = \sup_{t\in[t_L,t_U]} \frac{|m\{\tilde{F}_1^{(b)}(t;\mathbf{z}_0)\} - m\{\hat{F}_1(t;\mathbf{z}_0)\}|}{\hat{\sigma}(t;\mathbf{z}_0)}.$$

3. Estimate $C_{1-\alpha}(\mathbf{z}_0)$ from the sample $(1-\alpha)^{th}$ percentile of the B values of $C^{(b)}$, denoted by $\hat{C}_{1-\alpha}(\mathbf{z}_0)$.

Finally, the $(1-\alpha)\times 100\%$ confidence band for $F_1(t;\mathbf{z}_0)$, $t \in [t_L, t_U]$ is given by

$$m^{-1}\left[m\{\hat{F}_1(t;\mathbf{z}_0)\} \pm \hat{C}_{1-\alpha}(\mathbf{z}_0)\hat{\sigma}(t;\mathbf{z}_0)\right]. \tag{5.15}$$

One can perform CIF estimation and interval/band estimation using the `predict` function.

```
R> set.seed(2019)
```

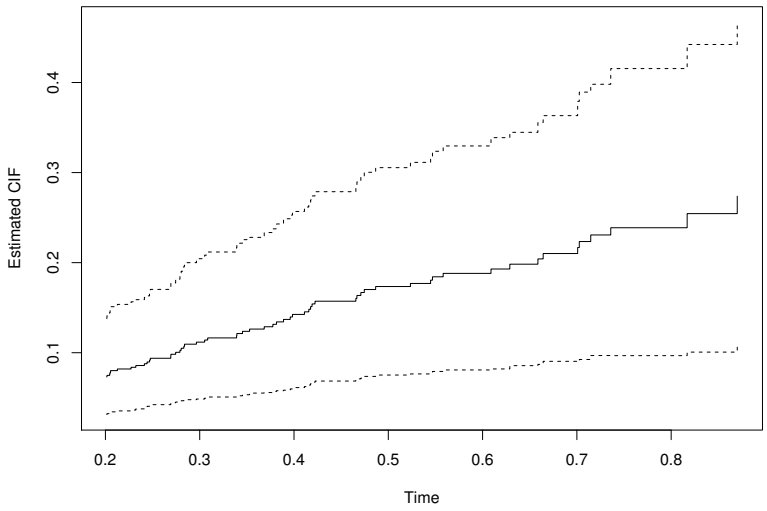

Figure 5.1: CIF estimate and corresponding 95% confidence intervals between $t_L = 0.2$ and $t_U = 0.9$.

```
R> z0 <- rnorm(10) # New covariate entries to predict
R> cif.point <- predict(fit2, cov = z0, getBootstrapVariance = TRUE,
+                       type = "interval", B = 100, seed = 2019,
+                       tL = 0.2, tU = 0.9)

R> plot(cif.point) # Figure 4.1
```

5.3.4 Penalized Fine-Gray regression via forward-backward scan

We extend our forward-backward scan approach for for penalized Fine-Gray regression as described in Section 5.1.3. The `fastCrrp` function performs LASSO, SCAD, MCP, and ridge (Hoerl and Kennard, 1970) penalization. The advantage of implementing this algorithm for penalized Fine-Gray regression is two fold. Since the cyclic coordinate descent algorithm used in the `crrp` function calculates the gradient and Hessian diagonals in $O(pn^2)$ time, as opposed to $O(pn)$ using our approach, we expect to see drastic differences in runtime for large sample sizes. Second, as mentioned earlier, researchers generally tune the strength of penalization through multiple model fits over a grid of candidate tuning parameter values. Thus the

difference in runtime between both methods grows larger as the number of candidate values increases. Below provides an example of performing LASSO-penalized Fine-Gray regression using 25 candidate values for λ. The syntax for `fastCrrp` is nearly identical to the syntax for `crrp`.

```
R> library(crrp)
R> lam.path <- 10^seq(log10(0.1), log10(0.001), length = 25)

R> # crrp package
R> fit.crrp <- crrp::crrp(dat$ftime, dat$fstatus, Z, penalty = "LASSO",
+                         lambda = lam.path, eps = 1E-6)

R> # fastcmprsk package
R> fit.fcrrp <- fastcmprsk::fastCrrp(dat$ftime, dat$fstatus, Z,
+                                    penalty = "LASSO", lambda = lam.path)

R> max(abs(fit.fcrrp$coef - fit.crrp$beta))

[1] 1.110223e-15

R> plot(fit.fcrrp) # Figure 4.2
```

5.4 Simulation studies

This section provides a more comprehensive illustration of the computational performance of the **fastcmprsk** package over two popular competing packages **cmprsk** and **crrp**. We simulate datasets under various sample sizes and fix the number of covariates $p = 100$. We generate the design matrix, $\mathbf{Z}$ from a p-dimensional standard normal distribution with mean zero, unit variance, and pairwise correlation $\text{corr}(z_i, z_j) = \rho^{|i-j|}$, where $\rho = 0.5$ simulates moderate correlation. For Section 5.4.1, the vector of regression parameters for cause 1, the cause of

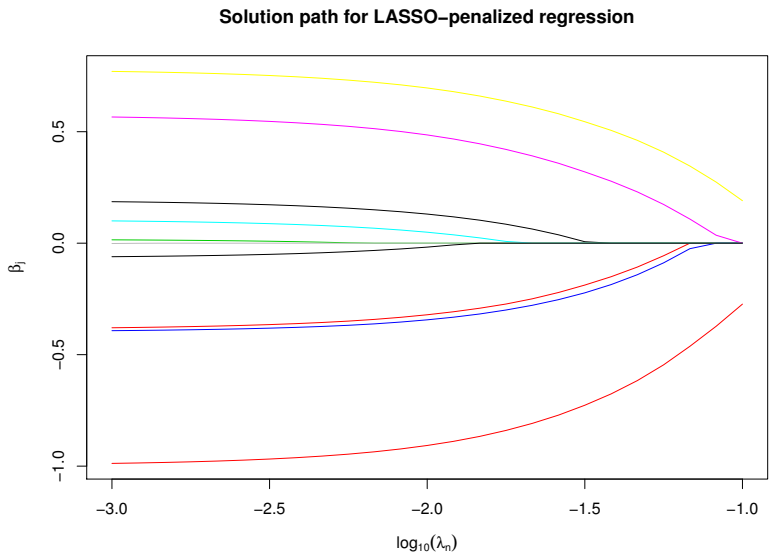

Figure 5.2: Path plot for LASSO-penalized Fine-Gray regression in our toy example.

interest, is $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^*, \ldots, \boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^* = (0.40, -0.40, 0, -0.50, 0, 0.60, 0.75, 0, 0, -0.80)$. For Section 5.4.2, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^*, \mathbf{0}_{p-10})$. We let $\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_1$. We set $\pi = 0.5$, which corresponds to a cause 1 event rate of approximately 41%. The average censoring percentage for our simulations varies between $30-35\%$. We use `simulateTwoCauseFineGrayModel` to simulate these data and average results over 100 Monte Carlo replicates. We report timing on a system with an Intel Core i5 2.9 GHz processor and 16GB of memory.

5.4.1 Comparison to the crr package

In this section, we compare the runtime and estimation performance of the `fastCrr` function to `crr`. We vary n from 1000 to 4000 and run `fastCrr` and `crr` both with and without variance estimation. We take 100 bootstrap samples to obtain the bootstrap standard errors with `fastCrr`.

Figure 5.3 illustrates the runtime performance (in seconds) between both `fastCrr` (dashed lines) and `crr` (solid lines) as n increases. It is clear that the performance of the `crr` methods increases quadratically while the `fastCrr` methods remain approximately linear. This leads to substantial improvement in computational performance for large sample sizes. Sec-

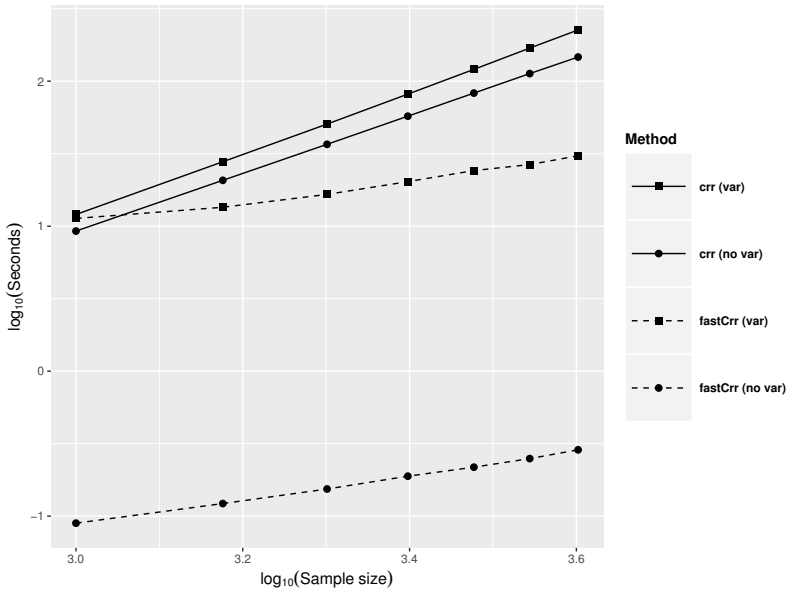

Figure 5.3: Runtime comparison between fastCrr and crr with and without variance estimation.

Table 5.2: Coverage probability (and standard errors) of 95% confidence intervals for $\boldsymbol{\beta}_{11} = 0.4$. Standard errors for fastCrr are obtained using 100 bootstrap samples.

	$n = 1000$	2000	3000	4000
crr	0.93 (0.03)	0.90 (0.03)	0.93 (0.03)	0.95 (0.02)
fastCrr	1.00 (0.00)	0.96 (0.02)	0.96 (0.02)	0.96 (0.02)

ond, the forward-backward scan allows us to efficiently compute variance estimates through bootstrapping. We see that bootstrapping for smaller sample sizes may not result in computational gains; however, notable differences are observed for larger sample sizes.

To assess the performance of the bootstrap procedure for variance estimation, Table 5.2 shows the coverage probability (and standard errors) of the 95% confidence intervals for $\boldsymbol{\beta}_{11} = 0.4$. We see confidence intervals are generally wider for the bootstrap approach but are close to the nominal 95% level.

5.4.2 Comparison to the crrp package

As mentioned in Section 5.1.3, Fu et al. (2017) provide an R package **crrp** for performing penalized Fine-Gray regression using the LASSO, SCAD, and MCP penalties. We com-

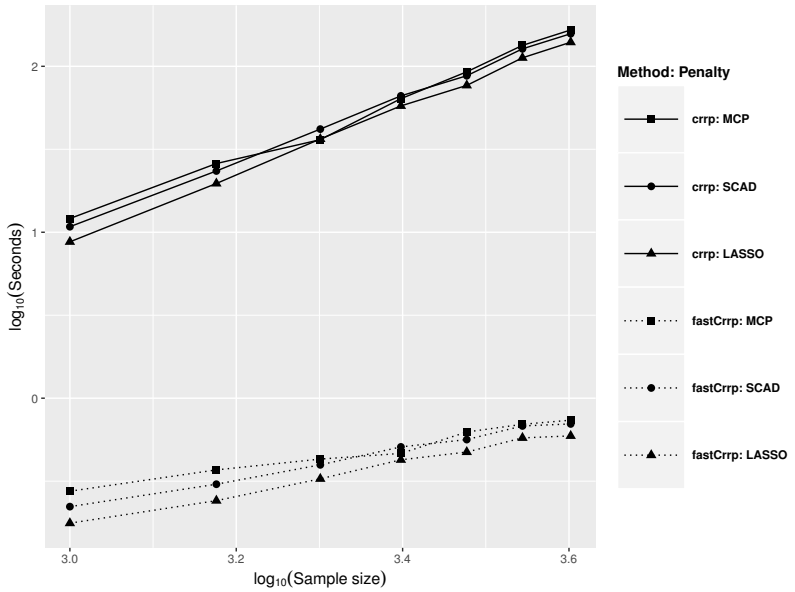

Figure 5.4: Runtime comparison between the **crrp** and **fastcmprsk** implementations of LASSO, SCAD, and MCP penalization. Solid and dashed lines represent the **crrp** and **fastcmprsk** implementation, respectively. Square, circle, and triangle symbols denote the penalties MCP, SCAD, and LASSO, respectively.

pare the runtime between `fastCrrp` with the implementation in the **crrp** package. To level comparisons, we modify the source code in `crrp` so that the function only calculates the coefficient estimates and BIC score. We vary $n = 1000, 1500, \ldots, 4000$, fix $p = 100$, and employ a 25-value grid search for the tuning parameter. Figure 5.4 illustrates the computational advantage the `fastCrrp` function has over `crrp`.

The computational performance of `crrp` (solid lines) increases quadratically while `fasrCrrp` (dashed lines) increases linearly, resulting in a 200 to 300-fold speed up in runtime when $n = 4000$. This, along with the previous section, strongly suggests that for large-scale competing risks datasets, analyses that may take several hours or days to perform using currently implemented methods are available within seconds or minutes using our forward-backward scan algorithm. We illustrate this in our real data analysis in the following section.

5.5 End-stage renal disease

We analyze data collected from the United States Renal Data System, a national data system funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that collects information about end-stage renal disease in the United States. Patients with end-stage renal disease are known to have a shorter life expectancy compared to their disease-free peers (USRDS Annual Report 2017) and kidney transplantation provides better health outcomes for patients with end-stage renal disease (Wolfe et al., 1999; Purnell et al., 2016). However patients may observe competing events such as death or renal function recovery or may wish to discontinue dialysis for quality of life purposes before transplant.

We extract a subset of the United States Renal Data System that spans a 10-year study time between January 2005 to June 2015 and a subsample to 125,000 subjects. We consider 63 demographic and clinical covariates. The event of interest is first kidney transplant for patients who were currently on dialysis. Death, renal function recovery, and discontinuation of dialysis are competing risks. Subjects who are lost to follow up or had no event by the end of study period are considered as right censored.

Table 5.3 shows the runtime results between both the `crr` and `fastCrr` implementations for unpenalized Fine-Gray regression. Using the `crr` function, parameter estimation without variance estimation took 1.3 hours to finish and with variance estimation took 26.7 hours to complete. The `fastCrr` function performed the same tasks within seconds, resulting in an over 1000-fold speedup for parameter estimation and an over 390-fold speedup for parameter and variance estimation. With respect to estimation, both approaches return nearly identical parameter estimates (maximum absolute difference of 3.29×10^{-7}).

To compare variance estimation, Figure 5.5 plots the 95% confidence intervals for the first six covariates: age at dialysis, sex, and presence of diabetes, hypertension, atherosclerotic heart disease, and cardiac failure. Both procedures return similar confidence intervals for all six covariates and we also observe similar results for the covariates not included in the figure.

Finally, we apply the LASSO, SCAD, and MCP variable selection routines to the dataset.

Table 5.3: Timing comparison using a subset of the USRDS dataset. The first two rows correspond to unpenalized Fine-Gray regression with and without variance estimation using crr and fastCrr. The last three rows correspond to penalized Fine-Gray regression using crrp and fastCrrp.

	Timing comparison (seconds)	
Unpenalized	crr	fastCrr
w.o. variance	4,544	4
w. variance	96,120	246
Penalized	crrp	fastCrrp
LASSO	86,304	32
SCAD	92,591	35
MCP	102,585	33

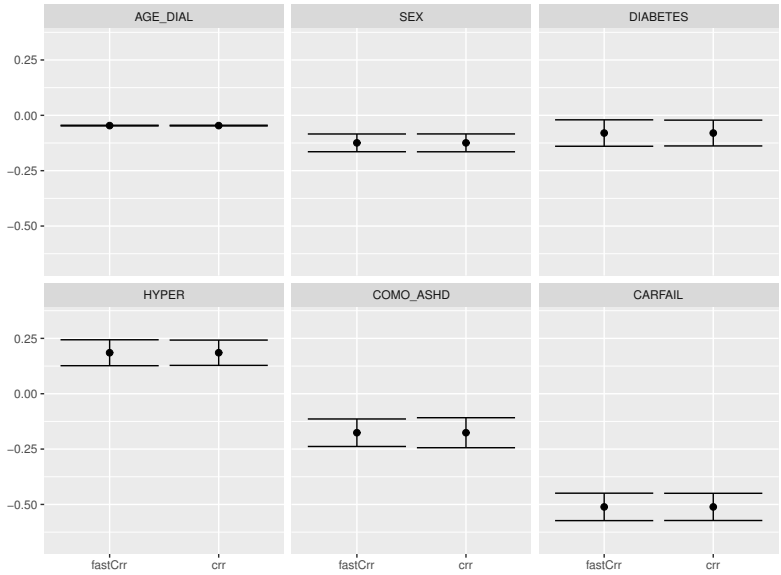

Figure 5.5: Point estimate and 95% confidence intervals reported by fastCrr (using 100 bootstrap samples) and crr.

Following Section 5.4.2, we use a grid of 25 candidate tuning parameters. The final model for each penalization method is chosen by selecting the tuning parameter that minimizes the BIC score. The runtime results can be found in last three rows of Table 5.3 which shows that the current implementations for variable selection are drastically slower than our package (an over 2000-fold difference in runtime). To assess the performance of each method, we consider a test set of 100,000 additional subjects and asses prediction performance through the concordance index (Wolbers et al., 2009). The predictive performance of all three methods are comparable

with similar concordance index values (≈ 0.85) that we attribute to the massive sample size of both the training and test set. As expected, both MCP and SCAD produce similar-sized models (48 variables for MCP and 49 variables for SCAD) due, in part, to their oracle behavior while LASSO selects 62 variables and are a superset of the variables selected by both MCP and SCAD. The variables selected by MCP are also all contained in the SCAD model.

In conclusion, our forward-backward scan algorithm results in a significant reduction in runtime for unpenalized and penalized Fine-Gray regression for large-scale competing risks data. Analyses using current packages may take hours or even over a day to finish; whereas the **fastcmprsk** package completes the same tasks within seconds or minutes.

5.6 Discussion

The **fastcmprsk** package provides a set of scalable tools for the analysis of large-scale competing risks data by developing an approach to linearize the computational complexity required to estimate the parameters of the Fine-Gray proportional subdistribution hazards model. The package implements both penalized and unpenalized Fine-Gray regression. We can conveniently extend our forward-backward algorithm to other applications such as stratified and clustered Fine-Gray regression. Calculating standard errors for both the parameter estimates and the CIF involves bootstrapping. We may further speed up standard error estimation through parallelization using, for example, the **doParallel** (Calaway et al., 2018) package.

Lastly, our current implementation assumes that covariates are densely observed across subjects. This is problematic in the sparse high-dimensional massive sample size (sHDMSS) domain (Mittal et al., 2014) where the number of subjects and sparsely-represented covariates easily exceed tens of thousands. These sort of data are typical in large comparative effectiveness and drug safety studies using massive administrative claims and electronic health record (EHR) databases and typically contain millions to hundreds of millions of patient records with tens of thousands patient attributes, which such settings are particularly useful for drug safety studies of a rare event such as unexpected adverse events (Schuemie et al.,

2018) to protect public health. We are currently extending our algorithm to this domain in a sequel paper.

Appendix to Chapter 5

A5.1 Data generation scheme

We describe the data generation process for the `simulateTwoCauseFineGrayModel` function. Let n, p, $\mathbf{Z}_{n\times p}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $u_{\min}$, $u_{\max}$ and π be specified. We first generate independent Bernoulli random variables to simulate the cause indicator ϵ for each subject. That is, $\epsilon_i \sim 1 + Bern\{(1-p)^{\exp(\mathbf{z}_i'\boldsymbol{\beta}_1)}\}$ for $i = 1, \ldots, n$. Then, conditional on the cause, event times are simulated from

$$\Pr(T_i \leq t | \epsilon_i = 1, \mathbf{z}_i) = \frac{1 - [1 - \pi\{1 - \exp(-t)\}]^{\exp(\mathbf{z}_i'\boldsymbol{\beta}_1)}}{1 - (1-\pi)^{\exp(\mathbf{z}_i'\boldsymbol{\beta}_1)}}$$

$$\Pr(T_i \leq t | \epsilon_i = 2, \mathbf{z}_i) = 1 - \exp\{-t\exp(\mathbf{z}_i'\boldsymbol{\beta}_2)\},$$

and $C_i \sim U(u_{\min}, u_{\max})$. Therefore, for $i = 1, \ldots, n$, we can obtain the following quadruplet $\{(X_i, \delta_i, \delta_i\epsilon_i, \mathbf{z}_i)\}$ where $X_i = T_i \wedge C_i$, and $\delta_i = I(X_i \leq C_i)$. Below is an excerpt of the code used in `simulateTwoCauseFineGrayModel` to simulate the observed event times, cause and censoring indicators.

```
#START CODE
...
...
# nobs, Z, p = pi, u.min, u.max, beta1 and beta2 are already defined.
# Simulate cause indicators here using a Bernoulli random variable
c.ind <- 1 + rbinom(nobs, 1, prob = (1 - p)^exp(Z %*% beta1))

ftime <- numeric(nobs)
```

```
eta1 <- Z[c.ind == 1, ] %*% beta1 #linear predictor for cause on interest
eta2 <- Z[c.ind == 2, ] %*% beta2 #linear predictor for competing risk

# Conditional on cause indicators, we simulate the model.
u1 <- runif(length(eta1))
t1 <- -log(1 - (1 - (1 - u1 * (1 - (1 - p)^exp(eta1)))^(1 / exp(eta1))) / p)
t2 <- rexp(length(eta2), rate = exp(eta2))
ci <- runif(nobs, min = u.min, max = u.max) # simulate censoring times

ftime[c.ind == 1] <- t1
ftime[c.ind == 2] <- t2
ftime <- pmin(ftime, ci) # X = min(T, C)
fstatus <- ifelse(ftime == ci, 0, 1) # 0 if censored, 1 if event
fstatus <- fstatus * c.ind  # 1 if cause 1, 2 if cause 2
...
...
```

CHAPTER 6

Concluding remarks and future research

In this dissertation we have 1) extended the BAR methodology to the Cox proportional hazards model and the Fine-Gray proportional subdistribution hazards model; 2) implemented an efficient algorithm to allow BAR regression for right-censored sHDMSS time-to-event data; 3) developed a cyclic coordinate-wise BAR algorithm that can compute the BAR estimator without carrying out iterative reweighted ℓ_2-penalizations and apply it to BAR for competing risks data; and 4) linearize parameter estimation for unpenalized and penalized Fine-Gray models.

In Chapter 3, we developed a new sparse Cox regression method by iteratively performing reweighted ℓ_2-penalized Cox regression where the penalty is adaptively reweighted to approximate the ℓ_0 penalty. The resulting estimator, known as BAR, can be viewed as a special local ℓ_0-penalized Cox regression method and is shown to enjoy properties of both ℓ_0- and ℓ_2-penalized Cox regression: it is selection consistent, oracle for parameter estimation, stable, and has a grouping property for highly-correlated covariates. Further, we illustrate through empirical studies that the BAR estimator has comparable or better performance for variable selection and parameter estimation as compared to current penalized Cox regression methods and, most importantly, can directly fit sHDMSS time-to-event data. In Chapter 4, we extended the BAR method for simultaneous parameter estimation and variable selection to the Fine-Gray model for competing risks data. More importantly, to make the BAR method scalable to large data, we have further developed 1) a novel coordinate-wise update (CYCBAR) algorithm to avoid carrying out multiple ridge regressions in the original BAR implementation and 2) a forward-backward scan algorithm to reduce the computational cost of the log-likelihood and its derivatives for the Fine-Gray model from the order of $O(n^2)$ to

$O(n)$. While showing comparable selection and estimation performance, the BAR method for the Fine-Gray model using the two new algorithms has produced thousands to tens of thousands fold speedups over some current penalization methods for the PSH model in numerical studies. In Chapter 5, we shifted our focus to implementing the forward-backward scan algorithm to parameter and cumulative incidence function estimation for unpenalized Fine-Gray regression and variable selection using LASSO, SCAD, and MCP penalizations. Our implementation, **fastcmprsk**, results in a significant reduction in runtime (up to 3000-fold) when compared to some of its popular competitors. Analyses using current packages may take hours or even over a day to finish; whereas our proposed implementation completes the same tasks within seconds or minutes.

There are several future directions where we can extend the research presented in this dissertation.

In my current work, we have shown the model selection consistency of BAR under diverging dimension. Expanding the BAR methodology to ultrahigh-dimensional time-to-event, where the number of covariates far exceeds the sample size, is of particular interest. One possibility is to replace the initial ridge estimator with a "truncated ridge" estimator in the algorithm to allow proper convergence rates needed for the initial estimator. The theoretical derivations will require more care and attention since standard techniques do not apply in ultrahigh-dimensional covariate spaces. Second, my current work focuses on estimation and model selection for both right-censored and competing risks data. In the variable selection community, there has been a surge of interest in post-selection inference. We can extend BAR to post-selection inference, allowing one to develop confidence intervals and perform hypothesis tests after penalization.

The **fastcmprsk** package can be further developed. We can conveniently extend our forward-backward algorithm to other applications such as stratified and clustered Fine-Gray regression. Calculating standard errors for both the parameter estimates and the CIF involves bootstrapping. We may further speed up standard error estimation through parallelization or estimate parameters for different strata or clusters in parallel. Further, this package currently only handles covariates that are densely observed across subjects. This

is problematic in the sparse high-dimensional massive sample size (sHDMSS) domain where the number of subjects and sparsely-represented covariates easily exceed tens of thousands. Extending this approach to handle sHDMSS data is critical in analyzing large comparative effectiveness and drug safety studies that use massive administrative claims and electronic health record (EHR) databases.

Bibliography

Ahn, K. W., A. Banerjee, N. Sahr, and S. Kim (2018). Group and within-group variable selection for competing risks data. *Lifetime Data Analysis 24*(3), 407–424.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

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.

Austin, P. C. and J. P. Fine (2017). Practical recommendations for reporting fine-g ray model analyses for competing risk data. *Statistics in Medicine 36*(27), 4391–4400.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics 5*(1), 232–253.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics 24*(6), 2350–2383.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*(1), 89–99.

Cai, J., J. Fan, R. Li, and H. Zhou (2005). Variable selection for multivariate failure time data. *Biometrika 92*(2), 303–316.

Calaway, R., S. Weston, and D. Tenenbaum (2018). **doParallel***: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.14.

Candes, E. J., M. B. Wakin, and S. P. Boyd (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications 14*(5), 877–905.

Chartrand, R. and W. Yin (2008). Iterative reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Cho, H. and A. Qu (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica 26*, 901–927.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 34*(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

Dai, L., K. Chen, Z. Sun, Z. Liu, and G. Li (2018). Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis 168*, 334–351.

Daubechies, I., R. DeVore, M. Fornasier, and C. S. Güntürk (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics 63*(1), 1–38.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Fan, J., Y. Feng, and Y. Wu (2010). High-dimensional variable selection for cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, Volume 6, pp. 70–86. Institute of Mathematical Statistics.

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.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 531–552.

Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*(446), 496–509.

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.

Frommlet, F. and G. Nuel (2016). An adaptive ridge procedure for l0 regularization. *PLoS ONE 11*(2), e0148620.

Fu, Z. (2016). **crrp***: Penalized Variable Selection in Competing Risks Regression*. R package version 1.0.

Fu, Z., C. R. Parikh, and B. Zhou (2017). Penalized variable selection in competing risks regression. *Lifetime Data Analysis 23*(3), 353–376.

Gao, X. and R. J. Carroll (2017). Data integration with high dimensionality. *Biometrika 104*(2), 251–272.

Gasso, G., A. Rakotomamonjy, and S. Canu (2009). Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing 57*(12), 4686–4698.

Genkin, A., D. D. Lewis, and D. Madigan (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics 49*(3), 291–304.

Gorodnitsky, I. F. and B. D. Rao (1997). Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing 45*(3), 600–616.

Gorst-Rasmussen, A. and T. Scheike (2012). Coordinate descent methods for the penalized semiparametric additive hazards model. *Journal of Statistical Software 47*(9), 1–17.

Gorst-Rasmussen, A. and T. Scheike (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(2), 217–245.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics 16*(3), 1141–1154.

Ha, I. D., M. Lee, S. Oh, J.-H. Jeong, R. Sylvester, and Y. Lee (2014). Variable selection in subdistribution hazard frailty models with competing risks data. *Statistics in Medicine 33*(26), 4590–4604.

Harrell, F. E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association 247*(18), 2543–2546.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*(4), 361–387.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hou, J., A. Paravati, J. Hou, R. Xu, and J. Murphy (2018). High-dimensional variable selection and prediction under competing risks with application to seer-medicare linked data. *Statistics in Medicine 37*(24), 3486–3502.

Hripcsak, G., J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. Van Der Lei, N. Pratt, G. N. Noren, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan (2015). Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics 216*, 574–578.

Johnson, B. A., Q. Long, Y. Huang, K. Chansky, and M. Redman (2012). Log-penalized least squares, iteratively reweighted lasso, and variable selection for censored lifetime medical cost. Technical report, Emory University.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Kasiske, B. L., J. F. Neylan III, R. R. Riggio, G. M. Danovitch, L. Kahana, S. R. Alexander, and M. G. White (1991). The effect of race on access and outcome in transplantation. *New England Journal of Medicine 324*(5), 302–307.

Keith, D., V. B. Ashby, F. K. Port, and A. B. Leichtman (2008). Insurance type and minority status associated with large disparities in prelisting dialysis among candidates for kidney transplantation. *Clinical Journal of the American Society of Nephrology 3*(2), 463–470.

Kosorok, M. R. and S. Ma (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *The Annals of Statistics 35*(4), 1456–1468.

Lawson, C. L. (1961). *Contributions to the theory of linear least maximum approximation.* University of California.

Liu, Z. and G. Li (2016). Efficient regularized regression with penalty for variable selection and network construction. *Computational and Mathematical Methods in Medicine 2016.*

Mittal, S., D. Madigan, R. S. Burd, and M. A. Suchard (2014). High-dimensional, massive sample-size cox proportional hazards regression for survival analysis. *Biostatistics 15*(2), 207–221.

Mittal, S., D. Madigan, J. Q. Cheng, and R. S. Burd (2013). Large-scale parametric survival analysis. *Statistics in Medicine 32*(23), 3955–3971.

Ni, A. and J. Cai (2018). Tuning parameter selection in cox proportional hazards model with a diverging number of parameters. *Scandinavian Journal of Statistics 45*(3), 557–570.

Ni, A., J. Cai, and D. Zeng (2016). Variable selection for case-cohort studies with failure time outcome. *Biometrika 103*(3), 547–562.

Osborne, M. R. (1985). *Finite algorithms in optimization and data analysis.* John Wiley & Sons, Inc.

Patzer, R. E., S. Amaral, H. Wasse, N. Volkova, D. Kleinbaum, and W. M. McClellan (2009). Neighborhood poverty and racial disparities in kidney transplant waitlisting. *Journal of the American Society of Nephrology 20*(6), 1333–1340.

Pintilie, M. (2006). *Competing Risks: A Practical Perspective*. John Wiley & Sons.

Prentice, R., J. Kalbfleisch, A. Peterson Jr, N. Flournoy, V. Farewell, and N. Breslow (1978). The analysis of failure times in the presence of competing risks. *Biometrics 34*(4), 541–554.

Purnell, T. S., X. Luo, L. A. Cooper, A. B. Massie, L. M. Kucirka, M. L. Henderson, E. J. Gordon, D. C. Crews, L. E. Boulware, and D. L. Segev (2018). Association of race and ethnicity with live donor kidney transplantation in the united states from 1995 to 2014. *Journal of the American Medical Association 319*(1), 49–61.

Purnell, T. S., X. Luo, L. M. Kucirka, L. A. Cooper, D. C. Crews, A. B. Massie, L. E. Boulware, and D. L. Segev (2016). Reduced racial disparity in kidney transplant outcomes in the united states from 1990 to 2012. *Journal of the American Society of Nephrology 27*(8), 2511–2518.

Putter, H., M. Fiocco, and R. Geskus (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine 26*(11), 2389–2430.

R Core Development Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pp. 297–331. Springer.

Schold, J. D., J. A. Gregg, J. S. Harman, A. G. Hall, P. R. Patton, and H.-U. Meier-Kriesche (2011). Barriers to evaluation and wait listing for kidney transplantation. *Clinical Journal of the American Society of Nephrology 6*(7), 1760–1767.

Schuemie, M. J., P. B. Ryan, D. Hripcsak, George Madigan, and M. A. Suchard (2017). Honest learning for the healthcare system: large-scale evidence from real-world data. *Science Under review*.

Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376*(2128), 20170356.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*(497), 223–232.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39*(5), 1.

Song, R., W. Lu, S. Ma, and X. Jessie Jeng (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika 101*(4), 799–814.

Stack, A. G., D. Yermak, D. G. Roche, J. P. Ferguson, M. Elsayed, W. Mohammed, L. F. Casserly, S. R. Walsh, and C. J. Cronin (2016). Differential impact of smoking on mortality and kidney transplantation among adult men and women undergoing dialysis. *BMC Nephrology 17*(95), 1–12.

Su, X., C. S. Wijayasinghe, J. Fan, and Y. Zhang (2016). Sparse estimation of cox proportional hazards models via approximated information criteria. *Biometrics 72*(3), 751–759.

Suchard, M. A., S. E. Simpson, I. Zorych, P. Ryan, and D. Madigan (2013). Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation 23*(1), 10.

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

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

Verweij, P. J. and H. C. Van Houwelingen (1993). Cross-validation in survival analysis. *Statistics in Medicine 12*(24), 2305–2314.

Verweij, P. J. and H. C. Van Houwelingen (1994). Penalized likelihood in cox regression. *Statistics in Medicine 13*(23-24), 2427–2436.

Volinsky, C. T. and A. E. Raftery (2000). Bayesian information criterion for censored survival models. *Biometrics 56*(1), 256–262.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Wang, T. and L. Zhu (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis 102*(7), 1141–1151.

Wipf, D. and S. Nagarajan (2010). Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing 4*(2), 317–329.

Wolbers, M., M. T. Koller, J. C. Witteman, and E. W. Steyerberg (2009). Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology 20*(4), 555–561.

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.

Wu, T. T. and K. Lange (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics 4*(2), 224–244.

Wu, Y. (2012). Elastic net for cox's proportional hazards model with a solution path algorithm. *Statistica Sinica 22*, 271–294.

Yang, G., Y. Yu, R. Li, and A. Buu (2016). Feature screening in ultrahigh dimensional cox's model. *Statistica Sinica 26*, 881–901.

Yang, Y. and H. Zou (2012). A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and Its Interface 6*(2), 167–173.

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

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

Zhang, T. and F. J. Oles (2001). Text categorization based on regularized linear classification methods. *Information Retrieval 4*(1), 5–31.

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

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*(489), 312–323.

Zhao, H., D. Sun, G. Li, and J. Sun (2018). Variable selection for recurrent event data with broken adaptive ridge regression. *Canadian Journal of Statistics 46*(3), 416–428.

Zhao, H., D. Sun, G. Li, and J. Sun (2019). Simultaneous estimation and variable selection for incomplete event history studies. *Journal of Multivariate Analysis 171*, 350–361.

Zhao, H., Q. Wu, G. Li, and J. Sun (2019). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association 0*(0), 1–13.

Zhao, S. D. and Y. Li (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis 105*(1), 397–411.

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.