University of Nevada, Reno

A Contribution to Variable Selection for the Cox Proportional Hazards Model with High-Dimensional Predictors

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Statistics and Data Science

by

Ryan Wu

Mihye Ahn, Ph.D. Thesis Advisor

Hojin Yang, Ph.D. Thesis Co-advisor

May, 2021

University of Nevada, Reno

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

RYAN WU

entitled

A Contribution to Variable Selection for the Cox Proportional Hazards Model with High-Dimensional Predictors

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Mihye Ahn, Ph.D. *Advisor*

Hojin Yang, Ph.D. *Co-advisor*

Yinghan Chen, Ph.D. *Committee Member*

Wei Yang, Ph.D., M.D. *Graduate School Representative*

David W. Zeh, Ph.D. Dean, *Graduate School*

May, 2021

Abstract

The aim of this thesis is to develop a variable selection framework with the spike-and-slab prior distribution via the hazard function of the Cox model. Specifically, we consider the transformation of the score and information functions for the partial likelihood function evaluated at the given data from the parameter space into the space generated by the logarithm of the hazard ratio. Thereby, we reduce the nonlinear complexity of the estimation equation for the Cox model and allow the utilization of a wider variety of stable variable selection methods. Then, we use a stochastic variable search Gibbs sampling approach via the spike-and-slab prior distribution to obtain the sparsity structure of the covariates associated with the survival outcome. To demonstrate the efficiency and accuracy of the proposed method in both low-dimensional and high-dimensional settings, we conduct numerical simulations to evaluate the finite-sample performance of the proposed method. Finally, we apply this novel framework within biological contexts on real world data sets such as primary biliary cirrhosis and lung adenocarcinoma data to find important variables associated with decreased survival in subjects with the aforementioned diseases.

Dedication

To **my parents, Renoir Wu and Rita Wu**

Acknowledgements

I want to first extend my deepest appreciation and gratitude to both of my thesis co-advisers. I want to thank Dr. Mihye Ahn, as she was the primary influence behind my choice to pursue my M.S. and has been providing me endless amounts of advice, guidance, and support since I started researching with her in my final undergraduate semester. I also want to thank Dr. Hojin Yang for the conception of this thesis and for his endless commitment and support as a co-adviser, even after his departure from UNR. I want to thank them both for their patience, guidance, time, and advice in bringing this thesis to its completion and for being wonderful co-advisers.

Additionally, I would like to thank the members of the examining committee, Dr. Yinghan Chen and Dr. Wei Yang for their time, valuable feedback, and patience.

I would also like to thank the Faculty and Staff at the Mathematics and Statistics Department here at UNR for creating an environment and curriculum where undergraduate and graduate students can succeed and grow as aspiring Statisticians.

Lastly, I would like to thank my family and friends who have always provided me endless love and support, and who ultimately shaped my studies at UNR to be a fulfilling and memorable experience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Some prominent cancers such as lung adenocarcinoma, a type of non-small cell lung cancer, have been widely studied in the 21st century, as many researchers have aimed to find how overexpression or downregulation in various genes can influence a person with lung adenocarcinoma's survival (Beer et al., 2002; DiFeo et al., 2008; Puzone et al., 2013; Chen et al., 2013). Finding genes associated with lung adenocarcinoma's survival is important because today, lung adenocarcinoma remains the most common lung cancer in the United States, representing about 40% of all lung cancers (Myers & Wallen, 2019). Additionally, this subtype of lung cancer is the most common lung cancer diagnosed to people who have never smoked, stressing the importance of detecting significant, associated genes. A typical approach is to conduct univariate analyses for each gene, find the relationship between each gene and survival times, and choose a few genes with strong signal. As an illustration, Beer et al. (2002) have used this marginal analysis based on Kaplan-Meier survival curves for over 7000 genes and selected a few genes associated with survival in lung adenocarcinoma patients.

Although these existing approaches are useful in expressing individual genes and their relations to lung adenocarcinoma, the genes identified may not sufficiently account for the biological mechanism. Additionally, these existing forms of univariate

analysis can not only be inefficient but can also cause complications due to multiple testing and constant adjustments to the significance level. Thus, model-based approaches such as a proportional hazards model would be needed for a more complete analysis and potentially more efficient, as the strength of model-based approaches lie in being able to efficiently and simultaneously detect significant genes while accounting for the joint effects among the covariates on the survival outcome.

The Cox proportional hazards model (Cox, 1972) specifies the association between survival time and a set of predictors via the utilization of a hazard function. In recent years, important applications that examine this association have risen in prominence in many fields of biomedical research such as clinical trials and gene studies (Singh et al., 2012; Güler, 2017). One of the primary advantages of the Cox proportional hazards model lies in its semi-parametric nature, which creates interpretability when accounting for the associations between proportional risk and the baseline hazard function. The proportional risk component allows for variable selection, as the hazard function is constructed by the predictors which most heavily influence the survival outcome of interest. In practice, researchers focused in clinical trials may be interested in how clinical treatments or patient attributes such as age and sex may affect their survival to a disease being studied. Similarly, researchers focused in genetic studies may be interested in performing variable selection by pinpointing which overexpressed or downregulated genes among thousands are associated with the survival time to a disease being studied.

While various methods, extensions, and application have been proposed to estimate the hazard function of the Cox model in settings where the number of predictors is a small number (Lawless, 2011; Kalbfleisch & Prentice, 2011; Fleming & Harrington, 2011; Ibrahim et al., 2001; Fan & Jiang, 2009; Su et al., 2016), these approaches experience difficulties in constructing the hazard function from high-dimensional pre-

dictors. This shortcoming becomes problematic in applications such as gene expression profiling, where data is strictly high-dimensional. Some high-dimensional frequentist approaches in the Cox model framework that perform variable selection and can reduce dimensionality include lasso (Tibshirani, 1997), smoothly-clipped absolute deviation (Fan & Li, 2001), and adaptive lasso (Zhang & Lu, 2007). These methods have been successfully utilized in some instances of gene expression profiling (Xu, 2012), but may suffer in performance as a result of diverging spectra, noise accumulation, computational burden, and inferential uncertainty issues (Fan & Lv, 2010; Ahn et al., 2012; Yang et al., 2018).

Alternatively, the high-dimensional Bayesian approaches of variable selection have received much attention and a large number of methods including the works of Park & Casella (2008), Griffin & Brown (2011), and Leng et al. (2014) have shown that the aforementioned frequentist approaches can be potentially outperformed via Bayesian methods in terms of variable selection. The primary characteristic of these approaches is their usage of independent Laplace type (*i.e.*, double-exponential) distributions as the prior distribution, concentrating more mass near 0 and in the tails. This characteristic, thereby, yields estimates for the regression coefficients as a sparse structure. Moreover, Tang et al. (2017) introduced a double-exponential spike-and-slab prior distribution which has been successfully utilized to analyze genes associated with Dutch breast cancer data and myelodysplastic syndromes.

There have been several advantages of the spike-and-slab prior distribution defined by the mixture distribution of the normal distribution and degenerate distribution on a certain point (Mitchell & Beauchamp, 1988; Madigan & Raftery, 1994; George & McCulloch, 1997; Ishwaran & Rao, 2005; Yang et al., 2020). For instance, any prior distribution concentrating more mass near a point can be flexibly generated by adjusting the point. Additionally, it provides nonlinear shrinkage of the

regression coefficients, which results in smoothed/regularized estimates as well as fully Bayesian inference after fitting a Markov Chain Monte Carlo. Lastly, it would be possible to develop a unified framework to deal with variable selection yielding the sparsity structure of the coefficients for both low-dimensional and high-dimensional circumstances.

The aim of this thesis is to develop a variable selection framework with the spike-and-slab prior distribution and consider high-dimensional applications to specify the association between survival time and a set of predictors via the hazard function of the Cox model. Specifically, we consider the transformation of the score and first derivative of the partial likelihood function evaluated at the given data from the parameter space into the space generated by the logarithm of the hazard ratio. Thereby, we reduce the nonlinear complexity of the estimation equation for the Cox model and make it possible to utilize a wider variety of stable variable selection methods. Then, we consider using a stochastic variable search (SVS) Gibbs sampling approach via the spike-and-slab prior distribution in order to obtain the sparsity structure of the covariates associated with the survival outcome. By incorporating these two steps, a more established and potentially more stable form of sparse variable selection can be constructed without any loss of information. We show that this approach provides a model where it is easy to interpret the resulting sparsity structure, as our primary goal is to detect and differentiate significant covariates from white noise. We also conduct numerical simulations to evaluate the finite-sample performance of our method. Finally, we apply our proposed methodology to detect the sparsity structure for the lung adenocarcinoma data (Beer et al., 2002). Unlike previous analyses which focus primarily on utilizing univariate analytical methods to describe the associations between individual genes and lung adenocarcinoma, we will obtain a more efficient form of sparse variable selection to simultaneously select dozens of genes which are

associated with survival times.

Upon completion of this thesis, we will be able to establish a novel framework for survival analysis models which can accurately and efficiently select significant covariates associated with survival times in both high-dimensional and low-dimensional settings. The established framework will also be capable of outperforming existing variable selection techniques commonly utilized on real data sets found in clinical contexts and gene expression profiling studies.

This thesis is organized as follows. In Section 2, we discuss the background behind popular existing procedures within the survival analysis framework. In Section 3, we introduce the specifics behind our estimation procedure. In Section 4, we conduct simulation studies to evaluate the finite-sample performance of our method when compared to other sparse estimation techniques. In Section 5, we apply our method to our analysis of the primary biliary cirrhosis data collected by the Mayo Clinic and lung adenocarcinoma cancer data collected from Beer et al. (2002). We provide concluding remarks in Section 6.

Chapter 2

Literature Review on Survival Analysis

2.1 Introduction to Survival Analysis

Survival analysis is a field that primarily focuses on analyzing a group of subjects with the purpose of predicting the expected time until an event of interest will occur to each subject. This event of interest is often death in biological applications of survival analysis such as disease studies and gene expression profiling which are the primary applications we will explore in this thesis. These studies are often conducted over finite time periods which involve *censored* data. Censoring is a form of missing data problem in which time-to-event is not observed. In this thesis, we will focus on *right censored* data, where censoring is observed for reasons such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing the event.

Within survival analysis, an important universal concept is defined as the *hazard ratio*. The hazard ratio measures a subject's probability of instantaneously experiencing the event of interest at time t, given that they have survived to time t and

have not been censored. The hazard ratio at any given time t can be computed with the hazard function:

$$\lambda(t) = \lim_{h \to 0} \frac{P[t \leq T < t + h | T \geq t]}{h} \tag{2.1}$$

The hazard function from (2.1) can then be integrated to obtain the survival function:

$$S(t) = \exp\Big\{ - \int_0^t \lambda(u) du \Big\} \tag{2.2}$$

In survival analysis, one may try to assume that the hazard function in (2.1), and by extension, the survival function in (2.2) originate from a common underlying distribution, such as the Weibull or exponential distributions. However, this assumption is rather strong and difficult to make in practice, often making more flexible approaches which do not assume an underlying hazard distribution more practical choices. In the next two sections, we examine the Kaplan-Meier estimator and the Cox proportional hazards model (Cox, 1972) which do not assume any underlying distribution for the hazard function.

2.2 Kaplan-Meier Estimator

While proportional hazards models are generally utilized to perform multivariate analyses with survival data, it is important to discuss the most popular univariate option utilized in the survival analysis framework. Kaplan-Meier estimators (Kaplan & Meier, 1958), which are conceptually identical to the complement of empirical distribution functions when censoring is not present, are utilized to compare survival curves among different univariate groups.

The Kaplan-Meier estimator for $S(t) = P(T \geq t)$ is defined as:

$$\widehat{S}_{KM}(t) = \prod_{x \leq t} \Big(1 - \frac{d(x)}{n(x)}\Big), \tag{2.3}$$

where $d(x)$ represents the amount of deaths at time x, and $n(x)$ represents the number of individuals at risk just prior to time x.

The Kaplan-Meier estimator can be applied to different levels within a single variable, producing plots that can display how the survival curves for individuals differ based on their respective levels. For instance, one may be able to produce Kaplan-Meier plots based on sex to assess if one sex has increased survival over the other for a particular disease. Generally speaking, one can visually assess the plots and determine if there is a significant difference between the survival curves. Software within R is capable of displaying Kaplan-Meier curves alongside confidence intervals and log-rank test statistics.

The standard error for the Kaplan-Meier estimator at time t, which is also known as Greenwood's formula (Greenwood, 1926), is defined as follows:

$$SE[\widehat{S}_{KM}(t)] = \widehat{S}_{KM}(t) \sqrt{\sum_{x \leq t} \frac{d(x)}{n(x)[n(x) - d(x)]}} \tag{2.4}$$

Utilizing the definition of the survival function in (2.3) and standard error in (2.4), one can incorporate confidence intervals for all of the survival curves for all levels of a given variable. Often, when two Kaplan-Meier curves have overlapping confidence intervals across the majority of the interval, there is a lack of evidence to suggest that there is a statistically significant difference in survival times between the two groups.

The log-rank test (Peto & Peto, 1972) formally tests the hypotheses $H_0 : S_1(t) =$

$S_2(t)$ for all t versus $H_1 : S_1(t) \neq S_2(t)$ for some t with the following χ^2 test statistic:

$$\chi^2 = \sum_{j=1}^{2} \frac{\left(\sum_t O_{jt} - \sum_t E_{jt}\right)^2}{\sum_t E_{jt}}, \tag{2.5}$$

where $\sum_t O_{jt}$ represents the sum of the observed number of events in the j^{th} group over time and $\sum_t E_{jt}$ represents the sum of the expected number of events in the j^{th} group over time.

The main drawbacks of this univariate analysis are most apparent in high-dimensional settings which are common within various biological applications such as gene expression studies. In these contexts, often thousands of genes must be investigated individually, making the process inefficient. Additionally, complications can easily arise due to multiple testing and constant adjustments to the significance level. Thus, model-based approaches such as the Cox proportional hazards model would be needed for a more complete and efficient analysis, a topic which is covered in the following section.

2.3 The Cox Proportional Hazards Model

Let the random variables T and C be the survival time and censoring time, respectively. The observed time variable is given by the minimum time denoted by $\widetilde{T} = \min\{T, C\}$. Let $\delta = I\{T \leq C\}$ be the censoring indicator that takes the value of 1 when the observed time experienced the event of interest, and takes the value of 0 when the observed time was censored. In most clinical and gene expression profiling applications, the event of interest is generally death, while censoring occurs when an individual is no longer at risk of dying from the disease. Let a random vector $X = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ be a set of covariates. Let T_i be the true survival time of the ith individual and assume that it takes its value in $\mathcal{T} = [0, \tau]$ for some positive

constant τ. We consider a sample of n subjects given by $\{(X_i, \widetilde{T}_i, \delta_i) : i = 1, \ldots, n\}$, where X_i denotes a vector of covariate of the ith individual and $\delta_i = I\{T_i \leq C_i\}$ denotes the censoring indicator of the ith individual.

The Cox proportional hazards model, which is semi-parametric in nature, does not depend on an underlying baseline hazard function, and is utilized to model the relationship between survival time and a vector of covariates via the following hazard function:

$$\lambda(t|X_i) = \lambda_0(t)e^{x_{i1}\beta_1 + \cdots + x_{ip}\beta_p} = \lambda_0(t)e^{X_i^T\beta}, \tag{2.6}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is a vector of parameters to be estimated and $\lambda_0(t)$ is the baseline hazard function, for which the baseline hazard function in (2.6) is the underlying hazard for any individual with $X_i = (0, \ldots, 0)^T$.

Let $N_i(t) = I\{\widetilde{T}_i \leq t, \delta_i = 0\}$ be the counting process and $Y_i(t) = I\{\widetilde{T}_i \geq t\}$ be the at-risk process for the ith individual for any $t \in [0, \tau]$. Without loss of generality, suppose that there are no ties in observed event times. The maximum likelihood estimator for β can be found through maximizing the partial log-likelihood function (Lawless, 2011; Kalbfleisch & Prentice, 2011; Fleming & Harrington, 2011) given by

$$\mathcal{L}(\tau, \beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[X_i^T\beta - \log\left\{ \sum_{j=1}^{n} Y_j(s) \exp(X_j^T\beta) \right\} \right] dN_i(s). \tag{2.7}$$

It should be noted that the primary limitation of the Cox model framework is that it does not work with high-dimensional data. In this scenario, it is common to rely on alternative frequentist estimators such as the ridge and lasso estimators covered in Section 2.6.

2.4 Stepwise Variable Selection

We briefly explore some of the most common stepwise variable selection techniques that are utilized to select significant covariates within the Cox model framework as constructed in (2.7). Specifically, we explore variable selection methods that arise from employing the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). Other techniques which consider stepwise variable selection via hazard ratio confidence intervals and p-values have been considered from a theoretical perspective previously, but have consequently been ruled out due to limitations with multiple comparisons in models with a large number of covariates and inherent flaws with small p-values that result from analyses on datasets with large sample sizes.

Within statistical packages such as R and SAS, it is possible to utilize functions that employ algorithms which perform stepwise variable selection for low-dimensional models with AIC and BIC. These algorithms are notoriously slow and may require the exploration of as many as 2^p models to find the optimal model. Nevertheless, these algorithms serve as a classic means of performing variable selection in basic survival models, and should be discussed for foundational purposes.

The AIC in survival analysis models is defined in a very similar manner as it is in the linear regression framework with the following function:

$$AIC = -2\log(\mathcal{L}(\tau, \beta)) + 2p, \tag{2.8}$$

where $\mathcal{L}(\tau, \beta)$ is a likelihood function and p is the number of covariates.

The formula for AIC in (2.8) is often not recommended with the usage of large data sets, as its penalty term does not take the sample size n or number of uncensored events d into consideration. The penalty term $-2p$ is modified in the construction of

the formula for the modified BIC below:

$$BIC = -2\log(\mathcal{L}(\tau, \beta)) + p\log(d), \tag{2.9}$$

where d is the number of uncensored events. The formula in (2.9) was proposed by Volinsky & Raftery (2000) as a modification to the original BIC formula utilizing n in the penalty term.

While stepwise methods are fairly simplistic from a theoretical perspective, it is important to note that its flaws severely limit the amount of survival datasets in which it can adequately examine. A lack of compatibility with high-dimensional datasets and low computational efficiency with datasets with a large amount of parameters still hinder the performance of stepwise methods reliant on AIC and BIC, whereas other stepwise methods which are reliant on p-values for covariates and hazard ratio confidence intervals are theoretically inferior to competing sparse variable selection methods in large low-dimensional datasets.

2.5 Sparse Estimation with Minimizing Approximated Information Criterion

An alternative to stepwise variable selection in low-dimensional settings comes with the utilization of sparse estimation with Minimizing approximated Information Criterion (MIC) as proposed by Nabi & Su (2017). This method considers the utilization of the following objective function:

$$MIC = -2\log(\mathcal{L}(\tau, \beta)) + \log(d)\sum_{j=1}^{p} w(\gamma_j),$$

where $w(\gamma)$ is a function of $\beta = \gamma w(\gamma) = \gamma \tanh(\gamma)$.

This particular choice in β is due to $\gamma \tanh(\gamma)$ being a unit dent function which is smooth everywhere except for $\beta = 0$, providing the sparsity structure in the selection of the covariates. This particular choice of penalty function will provide a much more efficient means of selecting covariates than in stepwise methods. To estimate β, generally the partial likelihood estimator in (2.7) is utilized as an initial starting point, prior to the employment of simulated annealing (Bélisle, 1992) and quasi-Newton BFGS methods (Broyden, 1967; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) for finding the MIC estimator. It should be noted, however, that the MIC estimator shares the primary flaw found in stepwise regression of being incompatible with high-dimensional data.

2.6 Cox Ridge and Lasso Estimators

As noted in the previous sections, the primary limitation of the Cox model framework and variable selection methods such as stepwise selection and sparse approximation with MIC are that they are incompatible with high-dimensional data. In these scenarios, other estimators of β can be found through regularization methods such as the lasso (Tibshirani, 1996) and ridge (Hoerl & Kennard, 1970b,a) estimators. Generally, the regularized estimator, which utilizes the penalty term to the partial log-likelihood function, can be obtained by maximizing the following regularized partial log-likelihood:

$$\mathcal{R}(\tau, \beta) = \mathcal{L}(\tau, \beta) - \lambda \mathcal{P}(\beta), \tag{2.10}$$

where $\mathcal{P}(\beta)$ denotes the penalty function of β and $\lambda > 0$ is a regularization parameter.

When imposing the L_2 penalty, *i.e.*, $\mathcal{P}(\beta) = \sum_{j=1}^{p} \beta_j^2$ on the partial log-likelihood in (2.7), maximizing the objective function in (2.10) yields the ridge estimator, for

which the L_2 penalty term serves in encouraging smoothness and avoiding problems with overfitting. Similarly, when imposing the L_1 penalty, $\mathcal{P}(\beta) = \sum_{j=1}^{p} |\beta_j|$ on (2.7), maximizing (2.10) yields the lasso estimator, which serves to directly detect the sparsity structure within the Cox model. Many forms of literature associated with variable selection assume the sparsity condition that there are only a few covariates truly related to the outcome, whereas all other covariates serve as noise which have no real effect on the outcome. Following this assumption, we will develop the Cox proportional hazards model framework identifying the sparsity structure of the covariates, based on the Bayesian approach given by the spike-and-slab prior distribution proposed by Tang et al. (2017).

2.7 Bayesian Hierarchical Cox Model

Prior to discussing the proposed framework, we discuss the Bayesian Hierarchical Cox model proposed by Tang et al. (2017). This method utilizes the well-known concept that the lasso estimator can be expressed as a hierarchical model with double-exponential prior as follows:

$$\beta_j | s \sim DE(\beta_j | 0, s) = \frac{1}{2s} \exp(-|\beta_j|/s), \tag{2.11}$$

where $s = 1/\lambda$. The double-exponential prior in (2.11) can utilize a spike-and-slab structure to produce the following spike-and-slab mixture double-exponential prior:

$$\beta_j | \gamma_j, s_0, s_1 \sim (1 - \gamma_j) DE(\beta_j | 0, s_0) + \gamma_j DE(\beta_j | 0, s_1) \tag{2.12}$$

Equivalently, (2.12) can be rewritten as:

$$\beta_j|\gamma_j, s_0, s_1 \sim DE(\beta_j|0, S_j) = \frac{1}{2S_j}\exp(-|\beta_j|/S_j), \tag{2.13}$$

where γ_j is an indicator variable and $S_j = (1-\gamma_j)s_0 + \gamma_j s_1$ with $s_1 > s_0 > 0$. Here s_0 is chosen to be incredibly small and serves as the "spike" component of the spike-and-slab, whereas s_1 is chosen to be relatively large and serves as the "slab" component.

In order to obtain the parameter estimates, the Bayesian Hierarchical Cox Model employs the EM coordinate descent algorithm which works as follows:

- **Step 1**: Choose a starting β^0 and θ^0 value which generally correspond to 0 and 0.5, respectively.

- **Step 2**: Update γ_j, the indicator parameter by its posterior expectation p_j:

$$p_j = \frac{DE(\beta_j|0, s_1)\theta}{DE(\beta_j|0, s_0)(1-\theta) + DE(\beta_j|0, s_1)\theta}$$

- **Step 3**: Update β using the cyclic coordinate descent algorithm which maximizes

$$Q_1(\beta) = \mathcal{L}(\tau, \beta) - \sum_{j=1}^{J} S_j^{-1}|\beta_j|, \tag{2.14}$$

  where $S_j^{-1} = (1-p_j)/s_0 + p_j/s_1$.

- **Step 4**: Update θ which uses the function:

$$Q_2(\theta) = \sum_{j=1}^{J}(\gamma_j \log \overline{p}_j + (1-\gamma_j)\log(1-\overline{p}_j))$$

- **Step 5**: Repeat Steps 2–4 until the algorithm converges. Convergence is based off of $|d^{(t)} - d^{(t-1)}|/(0.1 - |d^{(t)}|) < \epsilon$, where $d^{(t)}$ is the estimate of the deviance at iteration t and ϵ is an appropriately small value such as 0.00001.

The optimization function in (2.14) will provide the sparsity structure in the Bayesian Hierarchical Cox framework. Utilizing a similar approach to the existing Bayesian Hierarchical Cox model, we establish our novel framework in the next chapter.

Chapter 3

Spike-and-Slab Type Variable Selection

3.1 Estimation of Proposed Method

In this section, we develop our proposed framework by transforming the score defined on $\mathcal{T} \times \mathbb{R}^p$ into the function defined on $\mathcal{T} \times \mathbb{R}^n$ and the negative information defined on $\mathcal{T} \times \mathbb{R}^p$ into the function defined on $\mathcal{T} \times \mathbb{R}^n$. Define $Z_i = X_i^T \beta$ as the linear predictor of the ith observation for $i = 1, \ldots, n$. We consider the score function for (2.7) with respect to the linear predictors $Z = (Z_1, \ldots, Z_n)^T$, whose jth component is specifically given by

$$\frac{\partial \mathcal{L}(\tau, Z)}{\partial Z_j} = \int_0^\tau \left\{ I\{\widetilde{T}_j \leq s\} - \frac{Y_j(s) \exp(Z_j)}{\sum_{i=1}^n Y_i(s) \exp(Z_i)} \right\} dN_j(s). \tag{3.1}$$

The original score function with p-dimensional length for (2.7) can be written as

$$\frac{\partial \mathcal{L}(\tau, \beta)}{\partial \beta} = \sum_{i=1}^n \int_0^\tau \left\{ X_i - \frac{\sum_{j=1}^n Y_j(s) \exp(X_j^T \beta) X_j}{\sum_{j=1}^n Y_j(s) \exp(X_j^T \beta)} \right\} dN_i(s). \tag{3.2}$$

Then, the relation between (3.1) and (3.2) can be represented as

$$\frac{\partial \mathcal{L}(\tau,\beta)}{\partial \beta} = \sum_{j=1}^{n} \frac{\partial Z_j}{\partial \beta} \frac{\partial \mathcal{L}(\tau, Z(\beta))}{\partial Z_j},$$

where $\frac{\partial Z_j}{\partial \beta}$ is a $p \times 1$ vector with its k-th entry being $\frac{\partial Z_j}{\partial \beta}(k) = \frac{\partial Z_j}{\partial \beta_k}$ for $k = 1, \ldots, p$.

Define

$$\frac{\partial \mathcal{L}(\tau, Z)}{\partial Z} = \Big(\frac{\partial \mathcal{L}(\tau, Z)}{\partial Z_1}, \ldots, \frac{\partial \mathcal{L}(\tau, Z)}{\partial Z_n} \Big)^T \quad \text{and} \quad \frac{\partial Z}{\partial \beta} = \Big(\frac{\partial Z_1}{\partial \beta}, \ldots, \frac{\partial Z_n}{\partial \beta} \Big) \tag{3.3}$$

as an $n \times 1$ vector and a $p \times n$ matrix with its (k,l)th entry being $\frac{\partial Z}{\partial \beta}(k,l) = \frac{\partial Z_l}{\partial \beta_k}$, respectively.

The information function of the linear predictors can similarly be constructed and derived from the partial log-likelihood function in (2.7). Thus we have

$$\frac{\partial^2 \mathcal{L}(\tau, Z)}{\partial Z_j^2} = -\int_0^\tau \left\{ \frac{Y_j(s)\exp(Z_j)}{\sum_{i=1}^n Y_i(s)\exp(Z_i)} - \frac{Y_j(s)\exp(2Z_j)}{\{\sum_{i=1}^n Y_i(s)\exp(Z_i)\}^2} \right\} dN_j(s).$$

and

$$\frac{\partial^2 \mathcal{L}(\tau, Z)}{\partial Z_j \partial Z_k} = \int_0^\tau \left\{ \frac{Y_j(s)\exp(Z_j) Y_k(s)\exp(Z_k)}{\{\sum_{i=1}^n Y_i(s)\exp(Z_i)\}^2} \right\} dN_j(s)$$

for $j \neq k$ and $j = 1, \ldots, n$. The derivative of the score function with the $p \times p$ matrix, denoted by $\frac{\partial^2 \mathcal{L}(\tau,\beta)}{\partial \beta \partial \beta^T}$, can be approximated as

$$\frac{\partial^2 \mathcal{L}(\tau,\beta)}{\partial \beta \partial \beta^T} \approx \frac{\partial Z}{\partial \beta} \frac{\partial^2 \mathcal{L}(\tau, Z(\beta))}{\partial Z \partial Z^T} \frac{\partial Z^T}{\partial \beta^T} \approx \sum_{j=1}^{n} \frac{\partial Z_j}{\partial \beta} \frac{\partial^2 \mathcal{L}(\tau, Z(\beta))}{\partial Z_j^2} \frac{\partial Z_j}{\partial \beta^T}. \tag{3.4}$$

Notice that the Fisher scoring method (McCullagh & Nelder, 1989) allows the first approximation and the argument introduced in Hastie & Tibshirani (1990) allows the second approximation.

Now we utilize these transformed components including (3.3) and (3.4) in the framework of the partial log-likelihood given by (2.7). Considering the quadratic approximation of the partial log-likelihood function around β, we derive an algebraic form for the maximum likelihood estimator based on the transformed functions. For the fixed τ, the usual Taylor expansion of the estimation equation about β yields

$$\mathcal{L}(\tau,\beta) \approx \mathcal{L}(\tau,\widehat{\beta}) + (Z-\widehat{Z})^T \frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z} + \frac{1}{2}(Z-\widehat{Z})^T \frac{\partial^2 \mathcal{L}(\tau,\widehat{Z})}{\partial Z \partial Z^T}(Z-\widehat{Z}), \tag{3.5}$$

where $\widehat{Z} = (\widehat{Z}_1,\ldots,\widehat{Z}_n)^T$, $\widehat{Z}_i = X_i^T\widehat{\beta}$, and $\widehat{\beta}$ denotes some value close to β.

Then the right-hand side in (3.5) is equal to

$$\frac{1}{2}(\widehat{Z} - \frac{\partial Z^2}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}\frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z} - Z)^T \frac{\partial^2 \mathcal{L}(\tau,\widehat{Z})}{\partial Z^2}(\widehat{Z} - \frac{\partial Z^2}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}\frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z} - Z) \tag{3.6}$$

$$+ \mathcal{L}(\tau,\widehat{\beta}) - \frac{1}{2}(\frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z})^T \frac{\partial Z^2}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}\frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z}, \tag{3.7}$$

where $\frac{\partial Z^2}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}$ is the inverse of $\frac{\partial^2 \mathcal{L}(\tau,\widehat{Z})}{\partial Z^2}$ such that $\frac{\partial Z^2}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}\frac{\partial^2 \mathcal{L}(\tau,\widehat{Z})}{\partial Z^2}$ to be an identity matrix.

In practice, we recommend utilizing the ridge estimator for β in high-dimensional settings, while the partial likelihood estimator is used in low-dimensional settings. Note that the second term given in (3.7) is not dependent on Z including β within the approximated partial log-likelihood function given in (3.5). Thus, (3.6) can be rewritten in the following algebraic form:

$$\frac{1}{2}(F - \mathbf{X}\beta)^T \mathbf{W}(F - \mathbf{X}\beta), \tag{3.8}$$

where $F = \widehat{Z} - \frac{\partial Z \partial Z^T}{\partial^2 \mathcal{L}(\tau,\widehat{Z})}\frac{\partial \mathcal{L}(\tau,\widehat{Z})}{\partial Z}$, $\mathbf{X} = (X_1,\ldots,X_n)^T$, and $\mathbf{W} = \frac{\partial^2 \mathcal{L}(\tau,\widehat{Z})}{\partial Z \partial Z^T}$ are an $n \times 1$ response vector, an $n \times p$ design matrix, and an $n \times n$ weight matrix, respectively. For notational convenience, we denote $F_0^* = \mathbf{W}^{1/2}F$ and $\mathbf{X}_0^* = \mathbf{W}^{1/2}\mathbf{X}$, where $F_0^* =$

$(f_{01}^*, \ldots, f_{0n}^*)^T$ and $\mathbf{X}_0^* = (X_{01}^*, \ldots, X_{0n}^*)^T$.

3.2 Stochastic Variable Search Gibbs Sampling

After the transformation inducing (3.8) is done via the proposed approximation, any existing sparse method of variable selection within a linear regression framework can be utilized. While there are a wide variety of sparse methods to choose from, we consider using the spike-and-slab prior distribution that employs stochastic variable search (SVS) Gibbs sampling (Ishwaran & Rao, 2005). Specifically, we consider the following model with the following prior distributions:

$$\begin{aligned}
(f_{0i}^* | X_{0i}^*, \beta, \sigma^2) &\overset{\text{ind}}{\sim} N({X_{0i}^*}^T \beta,\, \sigma^2 n), \quad i = 1, \ldots, n \\
(\beta_k | \mathcal{J}_k, \tau_k^2) &\overset{\text{ind}}{\sim} N(0, \mathcal{J}_k \tau_k^2), \quad k = 1, \ldots, p \\
(\mathcal{J}_k | \nu_0, w) &\overset{\text{i.i.d.}}{\sim} (1-w)\delta_{\nu_0}(\cdot) + w\delta_1(\cdot) \\
(\tau_k^{-2} | a_1, a_2) &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_1, a_2) \\
w &\sim \text{Uniform}[0, 1] \\
\sigma^{-2} &\sim \text{Gamma}(b_1, b_2),
\end{aligned} \tag{3.9}$$

where $\delta_1(\cdot)$ is used to denote a degenerate point mass distribution concentrated at the value 1, and ν_0 denotes a small value near zero. Additionally, $\mathcal{J}_k$ denotes the latent indicator variable, which takes $\mathcal{J}_k = 1$ if the kth covariate is classified in the nonzero group, or $\mathcal{J}_k = \nu_0$ if the kth covariate is classified in the zero group for each k. Our model allows $\mathcal{J}_k \tau_k^2$ as the conditional variance of the prior distribution for the kth parameter, *i.e.* β_k, where τ_k^2 is a small positive value. While $\mathcal{J}_k$ is treated as an independent Bernoulli random variable with parameter w, where w is a complexity parameter controlling the probability $0 < w < 1$ as mentioned in Ishwaran & Rao

(2005), we will use the uniform prior as an indifference distribution. Thereby, the variance $\mathcal{J}_k\tau_k^2$ yields a continuous bimodal distribution with a spike at ν_0 and a right-continuous tail, where we denote $\gamma_k = \mathcal{J}_k\tau_k^2$ for brevity. This is important as the spike allows the posterior to shrink insignificant parameters towards 0 while the right-continuous tail can identify nonzero parameters. Additional details on the derivation and reasoning behind this structure can be found in Ishwaran & Rao (2005).

We fit the Bayesian model in (3.9) using the SVS Gibbs sampler. The SVS Gibbs sampler works with the transformed design matrix $\mathbf{X}_0^*$ and transformed response vector F_0^* to obtain the posterior sample $(\beta, \boldsymbol{J}, \boldsymbol{\tau}, w, \sigma^2|F_0^*)$, where $\boldsymbol{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_p)^T$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^T$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$. We present the computational algorithm in Ishwaran & Rao (2005), which provides the posterior distributions as follows:

- **Step 1**: Simulate the conditional distribution for β:

$$(\beta|\boldsymbol{\gamma}, \sigma^2, F_0^*) \sim N(\mu, \sigma^2\Sigma),$$

  where $\mu = \Sigma\mathbf{X}_0^{*T}F_0^*$, $\Sigma = (\mathbf{X}_0^{*T}\mathbf{X}_0^* + \sigma^2 n\boldsymbol{\Gamma}^{-1})^{-1}$, and $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_p)$.

- **Step 2**: Simulate $\mathcal{J}_k$ from its conditional distribution:

$$(\mathcal{J}_k|\beta_k, \tau_k, w) \overset{\text{ind}}{\sim} \frac{w_{1,k}}{w_{1,k} + w_{2,k}}\delta_{\nu_0}(\cdot) + \frac{w_{2,k}}{w_{1,k} + w_{2,k}}\delta_1(\cdot),$$

  where $w_{1,k}=(1-w)\nu_0^{-1/2}\exp(-\frac{\beta_k^2}{2\nu_0\tau_k^2})$ and $w_{2,k} = w\exp(-\frac{\beta_k^2}{2\tau_k^2})$ for $k = 1, \ldots, p$.

- **Step 3**: Simulate τ_k^{-2} from its conditional distribution:

$$(\tau_k^{-2}|\beta_k, \mathcal{J}_k) \overset{\text{ind}}{\sim} \text{Gamma}\Big(a_1 + \frac{1}{2}, a_2 + \frac{\beta_k^2}{2\mathcal{J}_k}\Big) \quad \text{for} \quad k = 1, \ldots, p$$

- **Step 4**: Simulate w from its conditional distribution:

$$(w|\boldsymbol{\gamma}) \sim \text{Beta}\Big(1 + \sum_{k=1}^{p} I(\gamma_k = 1), 1 + \sum_{k=1}^{p} I(\gamma_k = \nu_0)\Big)$$

- **Step 5**: Simulate σ^{-2} from its conditional distribution,

$$(\sigma^{-2}|\beta, F_0^*) \sim \text{Gamma}\Big(b_1 + \frac{n}{2}, b_2 + \frac{1}{2n}||F_0^* - \mathbf{X}_0^*\beta||^2\Big)$$

- **Step 6**: Set $\gamma_k = \mathcal{J}_k \tau_k^2$ for $k = 1, \ldots, p$.

Once we obtain the transformed design matrix and response vector, we can approach the problem as a Bayesian regression problem and develop it with the spike-slab prior being able to identify the sparse structure of the covariates. There are several possible approaches using Markov Chain Monte Carlo (MCMC) from the complete conditional distributions aforementioned. One possible approach is to obtain the MCMC samples as part of the spike-slab package in R that have been developed in recent years (Ishwaran et al., 2010). We implement this framework to produce the posterior samples. Model results can be somewhat sensitive to hyperparameter values, so we fit our model under the prior distributions including two gamma distributions with hyperparameters that are chosen as $a_1 = 5, a_2 = 20$ and $b_1 = b_2 = 0.0001$, respectively, and the point mass that is chosen as $\nu_0 = 0.005$.

Chapter 4

Simulation Study

4.1 Simulation Setting

In this section, we conducted a simulation study to examine the performance of our proposed method with simulated time-to-event data. We considered data sets generated from both low-dimensional and high-dimensional simulation settings. For the low-dimensional settings, we conducted scenarios with sample sizes of $n = 1000$ and $n = 3000$ with three different censoring rates of 0.2, 0.3, and 0.4. The number of covariates was set to be $p = 100$. For the high-dimensional settings, identical sample sizes of n with the identical corresponding censoring rates were considered, but the number of covariates was set to be $p = 4000$. In all simulation settings, four significant parameters were used, with $\beta_1 = 0.8$, $\beta_2 = 0.9$, $\beta_3 = -0.8$, $\beta_4 = -0.9$, and $\beta_5 = \cdots = \beta_p = 0$. Hence, the true set of the indices associated with the nonzero covariates can be denoted as $\mathcal{M} = \{1 \leq j \leq p \,|\beta_j \neq 0\} = \{1, 2, 3, 4\}$.

All of the covariates x_{ij} are independently generated from $N(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We independently generated u_i from Uniform$[0, 1]$ and then gener-

ated the true survival time based on the proportional hazards model given by

$$T_i = \frac{-\log u_i}{\lambda \exp\{\sum_{j=1}^{p} x_{ij}\beta_j\}}, \tag{4.1}$$

where $\lambda > 0$ denotes the baseline hazard function set to be 1 in this simulation. Note that the data generating process in (4.1) is equivalent to an exponential model. After drawing the censoring time C_i from Exponential(1), the observed time $\widetilde{T}_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I\{T_i \leq C_i\}$ for $i = 1, \ldots, n$, we generated a simulated data set of n subjects given by $\{(x_{i1}, \ldots, x_{ip}, \widetilde{T}_i, \delta_i) : i = 1, \ldots, n\}$ for each censoring rate.

As a form of comparison, we considered existing sparse variable selection methods including the Bayesian Cox model introduced by Tang et al. (2017), denoted by (B), the Cox lasso method (Tibshirani, 1997), denoted by (C), and the sparse estimation method using approximated information criterion introduced by Su et al. (2016), denoted by (D), where our proposed method was denoted by (A). The two methods including (B) and (D) serve as two competing sparse Bayesian approaches, whereas we consider the lasso (C) as a competing sparse frequentist approach. We define $\widehat{\mathcal{M}} = \{1 \leq j \leq p \,|\widehat{\beta}_j \neq 0\}$ as an estimated set of indices associated with the nonzero covariates. In order to compare with the other competing methods, we examined three types of performance measures. Specifically, we considered the probability of obtaining the correct model, defined as $P_C = P(\mathcal{M} = \widehat{\mathcal{M}})$, the probability of obtaining an overfitted model, defined as $P_O = P(\mathcal{M} \subset \widehat{\mathcal{M}})$, and the expected value of incorrect nonzero covariates, defined as $E_{IN} = E(|\widehat{\mathcal{M}} \cap \mathcal{M}^c|)$, where $\mathcal{M}^c$ is the compliment of $\mathcal{M}$. For each scenario, 100 simulated data sets were generated, and three performance measures were empirically computed across all simulated data sets, *i.e.*, $\widehat{P}_C = \frac{1}{M}\sum_{m=1}^{M} I\{\mathcal{M} = \widehat{\mathcal{M}}_m\}$, $\widehat{P}_O = \frac{1}{M}\sum_{m=1}^{M} I\{\mathcal{M} \subset \widehat{\mathcal{M}}_m\}$, and $\widehat{E}_{IN} = \frac{1}{M}\sum_{m=1}^{M} |\widehat{\mathcal{M}}_m \cap \mathcal{M}^c|$, respectively, where $\widehat{\mathcal{M}}_m$ denotes an estimated set of nonzero

Table 4.1: Simulation results for the low-dimensional data ($p = 100$): We used the proposed method, denoted by (A), Bayesian Cox method, denoted by (B), Cox lasso method, denoted by (C), and Cox method with the approximated information criterion, denoted by (D). $\widehat{E}_{IN}$ is the expected number of the incorrect nonzero covariates, $\widehat{P}_C$ is the empirical probability of the correct model, $\widehat{P}_O$ is the empirical probability of the overfitted model, n is the sample size, p is the number of the covariates, and censor denotes the censoring rate. The variances are provided in parentheses.

n	p	censor	**(A)**			**(B)**			**(C)**			**(D)**		
			$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$	$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$	$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$	$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$
1000	100	20%	0.13	0.88	0.12	34.80	0.00	1.00	0.72	0.57	0.43	3.59	0.08	0.92
			(0.13)	(0.11)	(0.11)	(21.64)	(0.00)	(0.00)	(1.29)	(0.25)	(0.25)	(5.09)	(0.07)	(0.07)
		30%	0.06	0.94	0.06	32.07	0.00	1.00	0.73	0.62	0.38	2.86	0.12	0.88
			(0.06)	(0.06)	(0.06)	(16.83)	(0.00)	(0.00)	(1.86)	(0.24)	(0.24)	(4.22)	(0.11)	(0.11)
		40%	0.11	0.89	0.11	28.11	0.00	1.00	0.52	0.69	0.31	2.86	0.09	0.91
			(0.10)	(0.10)	(0.10)	(21.07)	(0.00)	(0.00)	(1.02)	(0.22)	(0.22)	(4.88)	(0.08)	(0.08)
3000	100	20%	0.05	0.95	0.05	58.11	0.00	1.00	0.12	0.92	0.08	2.81	0.05	0.95
			(0.05)	(0.05)	(0.05)	(19.59)	(0.00)	(0.00)	(0.21)	(0.07)	(0.07)	(2.80)	(0.05)	(0.05)
		30%	0.08	0.93	0.07	54.35	0.00	1.00	0.08	0.93	0.07	3.03	0.05	0.95
			(0.09)	(0.07)	(0.07)	(24.96)	(0.00)	(0.00)	(0.09)	(0.07)	(0.07)	(3.42)	(0.05)	(0.05)
		40%	0.11	0.90	0.10	53.19	0.00	1.00	0.14	0.92	0.08	3.32	0.05	0.95
			(0.12)	(0.09)	(0.09)	(19.85)	(0.00)	(0.00)	(0.38)	(0.07)	(0.07)	(3.92)	(0.05)	(0.05)

indices for the mth simulated data for $m = 1, \ldots, M$ and M denotes the total number of simulated data sets.

4.2 Simulation Results

Table 4.1 reports the performance measures for the low-dimensional scenarios. In order to transform the score and information functions, we chose the initial $\widehat{\beta}$ to be the Cox partial likelihood estimator. This $\widehat{\beta}$ estimator is identically utilized as the initial estimator of the Cox model based on the approximated information criterion. For lasso, we chose to work with the smallest λ value within one standard error of the minimum λ selected based on 10-fold cross validation. After transforming the score and information functions, 1000 iterations of the spike-and-slab Gibbs sampler were run with a total of 500 iterations of burn-in. For the existing Bayesian Cox model (B), we conducted 50 iterations of the EM algorithm to obtain the estimator. With $n = 1000$ and $p = 100$, we see that the proposed method (A) experiences the best

performances out of all four methods. Additionally, the proposed method has by far the least amount of incorrect covariates. Notably, lasso (C) performs the second best, but never classifies more than 70% of models correctly. For $n = 3000$ and $p = 100$, the proposed method (A) is still the best performing Bayesian method by a wide margin. We see that the lasso estimator (C) provides similar results to (A), now that n has been increased. Meanwhile, the existing Bayesian Cox method (B) and approach accompanying the approximated information criterion (D) experience decreased performances when the sample size is increased in the low-dimensional settings.

It should be noted that it was also possible for us to test methods involving stepwise methods as proposed in Section 2.4 for the low-dimensional simulation settings. However, we must consider that if we had utilized backwards elimination to perform variable selection, the algorithm would have had to consider many models with many noisy variables prior to potentially arriving at the correct model with only four variables. Forward selection could be considered as an alternative, but this choice would rely on knowing that p is small, but the forward selection technique still fails to outperform the proposed method in these simulation settings.

Table 4.2 contains the performance measures for the high-dimensional scenarios. It should be noted that the approximated information criterion method (D) is incompatible with high-dimensional data, meaning high-dimensional results were compared solely between the three other methods. To choose $\widehat{\beta}$ for the transformation, we chose to work with the ridge estimator as the initial estimator in a high-dimensional setting. To use the ridge estimator, we first selected the tuning parameter based on the smallest λ within one standard error of the minimum of the λ value chosen via 10-fold cross validation. After choosing λ, $\widehat{\beta}$ was computed by minimizing the L_2 penalty function. To obtain the samples from the posterior distribution, 2600 iterations of

Table 4.2: Simulation results for the high-dimensional data ($p = 4000$): We used the proposed method, denoted by (A), Bayesian Cox method, denoted by (B), and Cox lasso method, denoted by (C). $\widehat{E}_{IN}$ is the expected number of the incorrect nonzero covariates, $\widehat{P}_C$ is the empirical probability of the correct model, $\widehat{P}_O$ is the empirical probability of the overfitted model, n is the sample size, p is the number of the covariates, and censor denotes the censoring rate. The variances of each member are provided below in parentheses.

n	p	censor	**(A)**			**(B)**			**(C)**		
			$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$	$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$	$\widehat{E}_{IN}$	$\widehat{P}_C$	$\widehat{P}_O$
1000	4000	20%	0.03	0.97	0.03	553.48	0.00	1.00	1.91	0.41	0.59
			(0.03)	(0.03)	(0.03)	(185.93)	(0.00)	(0.00)	(8.37)	(0.24)	(0.24)
		30%	0.04	0.96	0.04	492.63	0.00	1.00	1.77	0.50	0.50
			(0.04)	(0.04)	(0.04)	(183.71)	(0.00)	(0.00)	(11.55)	(0.25)	(0.25)
		40%	0.02	0.98	0.02	427.08	0.00	1.00	1.88	0.44	0.56
			(0.02)	(0.02)	(0.02)	(138.11)	(0.00)	(0.00)	(8.81)	(0.25)	(0.25)
3000	4000	20%	0.08	0.93	0.07	1695.04	0.00	1.00	0.68	0.67	0.33
			(0.09)	(0.07)	(0.07)	(406.50)	(0.00)	(0.00)	(1.57)	(0.22)	(0.22)
		30%	0.09	0.92	0.08	1562.78	0.00	1.00	0.56	0.76	0.24
			(0.10)	(0.07)	(0.07)	(497.41)	(0.00)	(0.00)	(2.31)	(0.18)	(0.18)
		40%	0.13	0.88	0.12	1415.71	0.00	1.00	0.40	0.83	0.17
			(0.13)	(0.11)	(0.11)	(382.23)	(0.00)	(0.00)	(3.17)	(0.14)	(0.14)

the spike-and-slab Gibbs sampler were run with a total of 100 iterations of burn-in. We set the iterations of the EM algorithm to be 50 for (B), and similarly chose the tuning parameter λ for (C) as mentioned above. For the setting with $n = 1000$ and $p = 4000$, we see that the proposed method (A) experiences the best performance out of all simulation settings. Meanwhile, we see that the spike-and-slab lasso via (B) and lasso (C) both have more difficulty detecting the sparsity structure within the simulated data. It should be noted that over 95% of the time, the proposed method (A) correctly selects the model while lasso (C) does not manage to do so more than 50% of the time. Similarly, less than 5 nonzero covariates are selected across 100 simulations for the proposed method. As for the simulations settings with $n = 3000$ and $p = 4000$, the proposed method (A) still maintains an excellent performance. Again, the existing Bayesian Cox model with spike-and-slab (B) experiences worsened performance as n increases. While lasso (C) experiences improved performance when n increases like the low-dimensional settings, the performance is still not comparable to the precision and performance of our method (A).

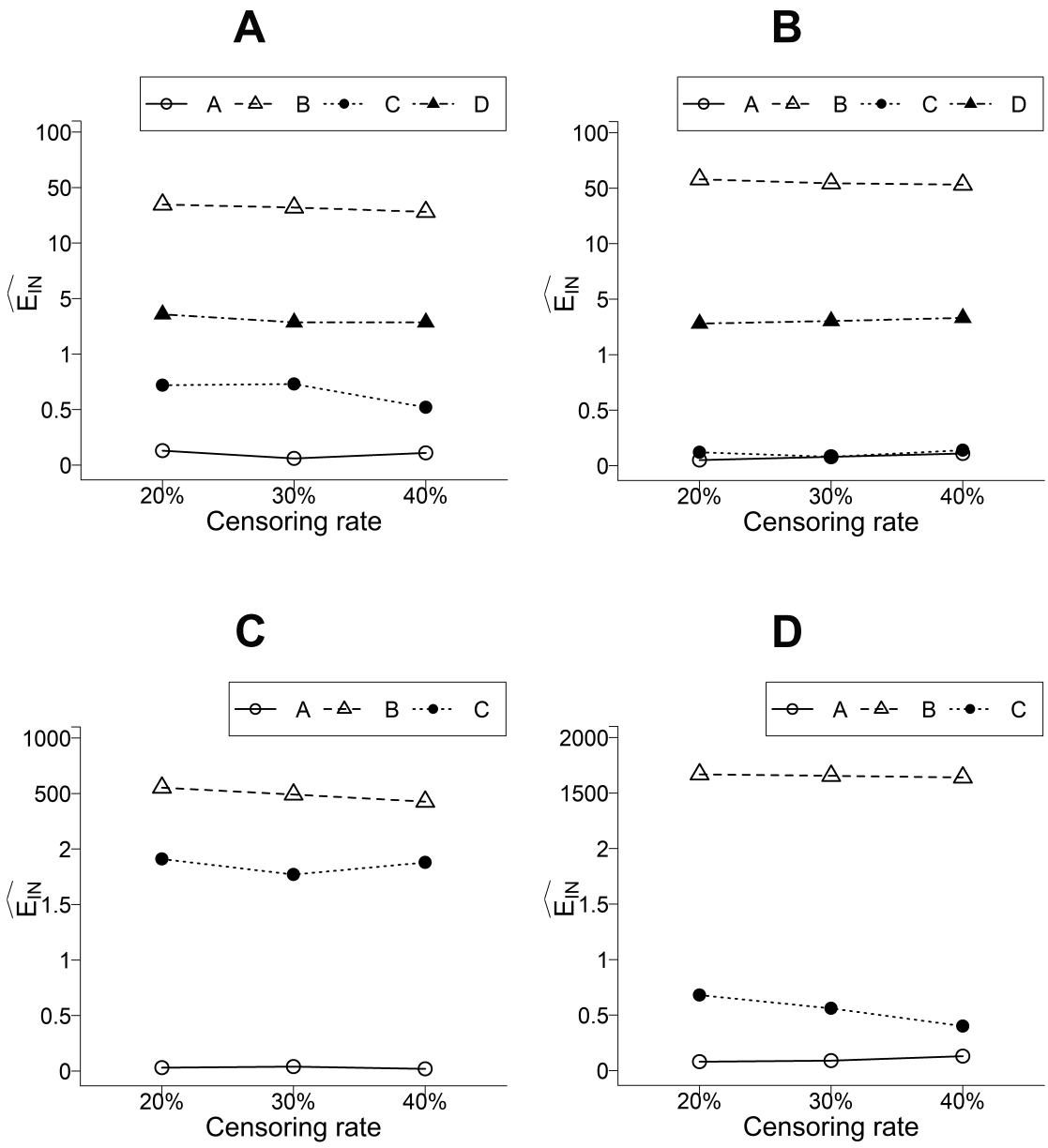


Figure 4.1: The empirical average of incorrect nonzero covariates across all scenarios: Panels A, B, C, and D depict the graphical representations of $\widehat{E}_{IN}$ for all scenarios including the two low dimensional settings ($n = 1000$, $p = 100$ and $n = 3000$, $p = 100$), and two high dimensional settings ($n = 1000$, $p = 4000$ and $n = 3000$, $p = 400$), respectively. In each panel, we included the proposed method (solid line with open circle), Bayesian Cox method (dashed line with open triangle), Cox lasso method (dotted line with filled circle), and the method with minimized approximated information criterion (dot-dashed line with filled triangle).

Figure 4.1 shows the graphical representations for the empirical average of incorrect nonzero covariates across all scenarios. Panels A, B, C, and D depict the graphical representations of $\widehat{E}_{IN}$ for all scenarios including the two low dimensional settings ($n = 1000$, $p = 100$ and $n = 3000$, $p = 100$), and two high dimensional settings ($n = 1000$, $p = 4000$ and $n = 3000$, $p = 400$), respectively.

In summary, it should be noted that while the performances of lasso (C), Bayesian Cox model (B), and Cox model with the approximated information criterion (D) are heavily influenced by sample sizes, the performances of the proposed method are relatively consistent across numerous sample sizes in this simulation. These results suggest that, in practice, the proposed method is a method that is capable of pinpointing sparsity structures within datasets regardless of sample size. Similarly, increasing the amount of noisy variables within the data can significantly influence the incorrect amount of incorrect covariates chosen by methods including (B) and (C), leading to overfitted models. On the other hand, our method (A) tends to maintain similar, if not improved, results when the total amount of covariates is increased.

Table 4.3 and Table 4.4 contain the performance of the prediction measures. We considered the areas under the receiver operating characteristic curve, denoted by AUC, the area under the precision-recall curve, denoted by PRC, and concordance index, denoted by CCI, as three predictive measures. It can be seen in the low-dimensional settings that ultimately all of the methods had a rather similar performance. However, in the high-dimensional setting, we note that (A) and (C) performed significantly better in the $n = 1000$ case than (B) which severely overfitted models. We see that for all three performance measures, (B) has a significantly worse performance than (A) and (C). In the $n = 3000$ case, our proposed method (A) performed slightly better than (C) when we considered the concordance index.

Table 4.3: Predictive results for the low-dimensional data ($p = 100$): We used the proposed method, denoted by (A), Bayesian Cox method, denoted by (B), Cox lasso method, denoted by (C), and Cox method with the approximated information criterion, denoted by (D). Performance was assessed via three measures abbreviated as AUC (area under the receiver operating characteristic curve), PRC (area under the precision-recall curve), and CCI (concordance index). n is the sample size, p is the number of the covariates, and censor denotes the censoring rate, where the standard deviations are provided in parentheses

n	p	censor	**(A)**			**(B)**			**(C)**			**(D)**		
			AUC	PRC	CCI	AUC	PRC	CCI	AUC	PRC	CCI	AUC	PRC	CCI
1000	100	20%	0.834	0.831	0.821	0.832	0.829	0.821	0.834	0.831	0.822	0.831	0.828	0.820
			(0.012)	(0.015)	(0.009)	(0.012)	(0.015)	(0.009)	(0.012)	(0.015)	(0.009)	(0.012)	(0.015)	(0.009)
		30%	0.837	0.833	0.825	0.836	0.831	0.824	0.837	0.832	0.825	0.835	0.831	0.823
			(0.013)	(0.016)	(0.010)	(0.013)	(0.016)	(0.010)	(0.011)	(0.015)	(0.009)	(0.013)	(0.016)	(0.010)
		40%	0.837	0.833	0.825	0.836	0.832	0.824	0.837	0.833	0.825	0.835	0.831	0.823
			(0.013)	(0.016)	(0.010)	(0.013)	(0.016)	(0.010)	(0.013)	(0.016)	(0.010)	(0.013)	(0.016)	(0.010)
3000	100	20%	0.836	0.834	0.824	0.835	0.832	0.824	0.836	0.834	0.824	0.835	0.833	0.824
			(0.006)	(0.008)	(0.005)	(0.006)	(0.008)	(0.005)	(0.006)	(0.008)	(0.005)	(0.006)	(0.008)	(0.005)
		30%	0.837	0.835	0.825	0.837	0.834	0.825	0.836	0.834	0.825	0.837	0.834	0.824
			(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.008)	(0.004)
		40%	0.837	0.834	0.825	0.836	0.834	0.825	0.836	0.834	0.825	0.836	0.833	0.824
			(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.004)

Table 4.4: Predictive results for the high-dimensional data ($p = 4000$): We used the proposed method, denoted by (A), Bayesian Cox method, denoted by (B), and Cox lasso method, denoted by (C). Performance was assessed via three measures abbreviated as AUC (area under the receiver operating characteristic curve), PRC (area under the precision-recall curve), and CCI (concordance index). n is the sample size, p is the number of the covariates, and censor denotes the censoring rate, where the standard deviations are provided in parentheses

n	p	censor	**(A)**			**(B)**			**(C)**		
			AUC	PRC	CCI	AUC	PRC	CCI	AUC	PRC	CCI
1000	4000	20%	0.838	0.831	0.827	0.811	0.802	0.811	0.837	0.833	0.825
			(0.012)	(0.016)	(0.009)	(0.014)	(0.019)	(0.010)	(0.011)	(0.015)	(0.008)
		30%	0.837	0.829	0.828	0.817	0.807	0.815	0.837	0.832	0.824
			(0.011)	(0.016)	(0.008)	(0.012)	(0.017)	(0.008)	(0.012)	(0.016)	(0.008)
		40%	0.838	0.831	0.828	0.824	0.815	0.817	0.837	0.831	0.825
			(0.011)	(0.014)	(0.008)	(0.012)	(0.016)	(0.010)	(0.013)	(0.016)	(0.009)
3000	4000	20%	0.838	0.833	0.829	0.803	0.790	0.832	0.837	0.834	0.825
			(0.006)	(0.008)	(0.004)	(0.006)	(0.009)	(0.005)	(0.006)	(0.008)	(0.005)
		30%	0.838	0.832	0.828	0.817	0.805	0.832	0.835	0.832	0.824
			(0.006)	(0.007)	(0.004)	(0.007)	(0.009)	(0.004)	(0.006)	(0.008)	(0.004)
		40%	0.838	0.833	0.828	0.827	0.817	0.832	0.837	0.833	0.825
			(0.006)	(0.008)	(0.004)	(0.006)	(0.010)	(0.004)	(0.005)	(0.008)	(0.004)

Additionally, we see that (B) performs significantly worse than (A) and (C) when we considered the performance measures of AUC and PRC.

Chapter 5

Real Data Application

5.1 Primary Biliary Cirrhosis Data

We demonstrate our method by first performing an analysis on the primary biliary cirrhosis (PBC) data collected by the Mayo Clinic between 1974 and 1984. This dataset, which is publicly available in the *survival* package in R, consists of 418 observations and 17 covariates. Prior to applying our method on the PBC dataset, we first performed some filtering measures. All observations that contained missing covariate values were removed, leaving us with 276 observations. The PBC dataset, unlike most survival datasets, consists of three censoring indicators. The first indicator, which corresponds to 0, represents patients who survived and did not need a liver transplant. The indicator corresponding to 1 corresponds to patients who survived, but needed a liver transplant, while the indicator corresponding to 2 corresponds to patients who died from PBC. To ensure censoring was binary, we assigned patients who survived the censoring indicator of 0 and patients who died the censoring indicator of 1. After performing these adjustments to the PBC data, we were left with 111 censored observations, creating a censoring rate of 40.22%.

We first standardized all of the covariates. This was due to the varying scales

Table 5.1: Results for the PBC data set: We used the proposed method (A), Bayesian Cox method (B), Cox lasso method (C), and Cox method with the approximated information criterion (D).

Covariate	**(A)**	**(B)**	**(C)**	**(D)**
trt	No	No	No	No
age	Yes	Yes	Yes	Yes
sex	No	No	Yes	No
ascites	Yes	No	Yes	No
hepato	Yes	No	No	No
spiders	Yes	No	Yes	No
edema	Yes	Yes	Yes	Yes
bili	Yes	Yes	Yes	Yes
chol	Yes	No	No	No
albumin	Yes	Yes	Yes	Yes
copper	Yes	Yes	Yes	Yes
alk.phos	No	No	No	No
ast	Yes	No	Yes	Yes
trig	No	No	No	No
platelet	No	No	No	No
protime	Yes	No	Yes	Yes
stage	Yes	Yes	Yes	Yes

within the PBC data, specifically with regards to variables such as alkaline phosphatase and serum bilirubin. The initial estimator $\widehat{\beta}$ was chosen as the maximum likelihood estimator provided by the partial likelihood function. After transforming the score and information functions, 1000 iterations of the spike-and-slab Gibbs sampler were run with a total of 500 iterations of burn-in. Again, we compared our results on the standardized survival data set with the existing Bayesian Cox model (B), Cox lasso method (C), and Cox method using approximated information criterion (D). As this data set was already explored in Tibshirani (1997) and Nabi & Su (2017), we reported their original results for (C) and (D).

Table 5.1 reported the results for the PBC data set. Six variables were selected by all methods, which consist of age, the presence of edema, the amount of serum bilirubin in mg/dl, albumin in gm/dl, urine copper in ug/day, and histologic stage

of disease. Two additional variables were selected by all of the methods except the Bayesian lasso (B), which consisted of the AST enzyme in U/liter, and the prothrombin time in seconds. Notice that sex was selected by the lasso (C) but not via any other methods. In the PBC data, 242 out of 276 subjects are female, creating a large imbalance for sex which may have contributed to the result where only the lasso estimator detected sex as significant. Four additional covariates were selected via our method (A) and were not selected by the method with the approximated information criterion (D). These covariates are the presence of ascites, the presence of hepatomegaly, the presence of spiders, and the level of serum cholesterol in mg/dl.

Figure 5.1 contains Kaplan-Meier plots for these four covariates. For the cholesterol covariate, the groups were split based on the median value. The lasso estimator (C) selected the presence of ascites and the presence of spiders to be significant, but did not select the presence of hepatomegaly or the level of serum cholesterol to be significant. However, two studies after the initial lasso publication support the significance of these variables (Janičko et al., 2013; Uddenfeldt & Danielsson, 2000). In all four plots, there is clear visual evidence to suggest there are significant differences between the survival curves of the groups. Additionally, the p-values from the Kaplan-Meier log-rank tests, which are calculated and displayed within each Kaplan-Meier plot, suggest that there are statistically significant differences between the two groups for each variable.

Within the context of clinical trials, under-fitted models can lead to severe consequences, as they might overlook important and significant covariates which are associated with patient survival times. The results of this analysis suggest that the proposed method can detect important and significant covariates that may not necessarily be reported by other existing methods within low-dimensional settings.

5.2 Lung Adenocarcinoma Data

Next, we demonstrate the performance of our method in a high-dimensional setting with the lung adenocarcinoma data presented by Beer et al. (2002). The data set consists of 86 observations with 7129 genes serving as covariates. Of the 86 observations, 62 are censored, yielding a censoring rate of 72.09%. As the data set was high-dimensional, the initial $\widehat{\beta}$ for the transformation was chosen utilizing an initial ridge estimator. The tuning parameter was chosen as the smallest λ within one standard error of the minimum λ chosen by leave-one-out cross-validation (LOOCV). After choosing λ, we used the L_2 penalty function to calculate $\widehat{\beta}$. After transforming the score and information functions, 5500 iterations of the spike-and-slab process were run with 500 iterations of burn-in.

Ultimately, our method detected 28 genes as being significant to predicting the survival of patients with lung adenocarcinoma. To elaborate on the results, we discuss four of the significant genes reported by our proposed method, which correspond to genes which have been previously reported to have strong associations with lung cancer. These genes are referred to as PRKACB, GAPDH, KLF6, and STX1A. In previous studies, there has been strong evidence to suggest that downregulation in the PRKACB genes can increase risk for those with lung adenocarcinoma (Chen et al., 2013). Similarly, overexpression in GAPDH, KLF6, and STX1A also leads to the increased risk (Puzone et al., 2013; DiFeo et al., 2008; Beer et al., 2002).

To further examine these genes, we showcase Kaplan-Meier plots in Figure 5.2 that correspond to the four previously mentioned genes. Although the genes have continuous values, observations were split into two groups based on the median values for each gene. In the Kaplan-Meier plots, we observe a clear difference between survival times within the groups plotted. For PRKACB, GAPDH, and STX1A, the p-values from the Kaplan-Meier log-rank test (Bland & Altman, 2004) also suggest

significant differences between the subgroups of each gene. These results are consistent with other studies, and suggest that our proposed method is correctly detecting signals within a high-dimensional setting with a high censoring rate.

The Kaplan-Meier log-rank test for the KLF6 gene yields a p-value of 0.06, which normally suggests that there is no significant difference between the survival curves. However, the loss of power within the Kaplan-Meier log-rank test can be attributed to the early crossings between the two survival curves and relatively small sample size (Li et al., 2015). Visually, it is clear after the time point of 25 that there is a clear difference in survival probabilities between those who experience overexpression in the KLF6 gene and those who do not. This gene presents an important example of our method being able to detect a significant gene that may have been overlooked by other popular methods such as the log-rank test. The results of this analysis suggest that even in high-dimensional settings with high censoring rates, our method is able to detect significant variables and sparsity structures.

Figure 5.1: Kaplan-Meier plots for the PBC data set: Panels A, B, C, and D contain the survival probabilities for the presence of ascites, presence of spiders, presence of hepatomegaly, and amount of serum cholesterol, respectively. For A, B, and C, the green lines indicate the presence of the covariate, whereas the red lines indicate the absence of the covariate. For D, the green line represents serum cholesterol levels higher than the median level, whereas the red line represents subjects who had serum cholesterol levels lower than the median level. In each panel, p-values for the log-rank test are provided.

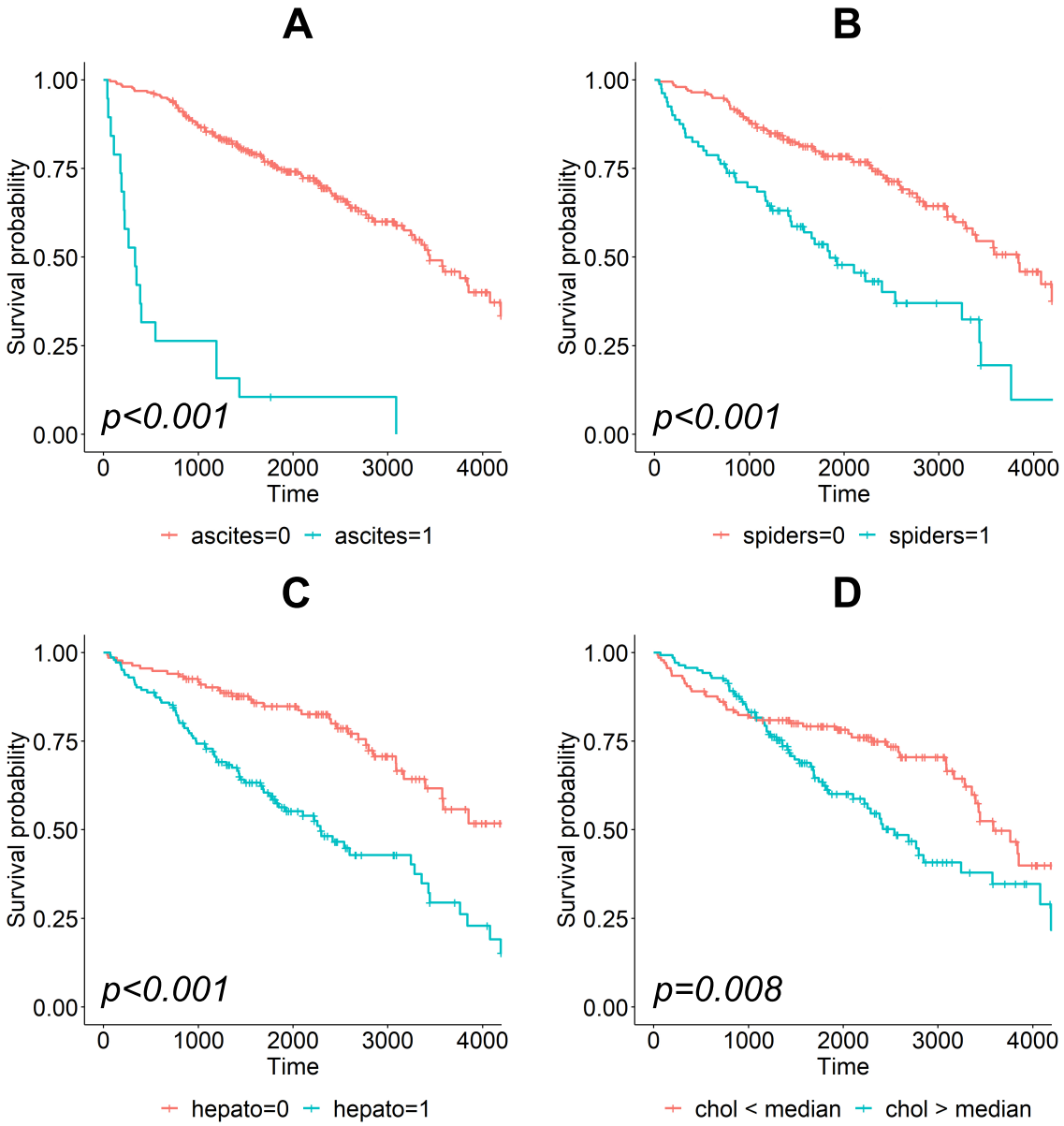
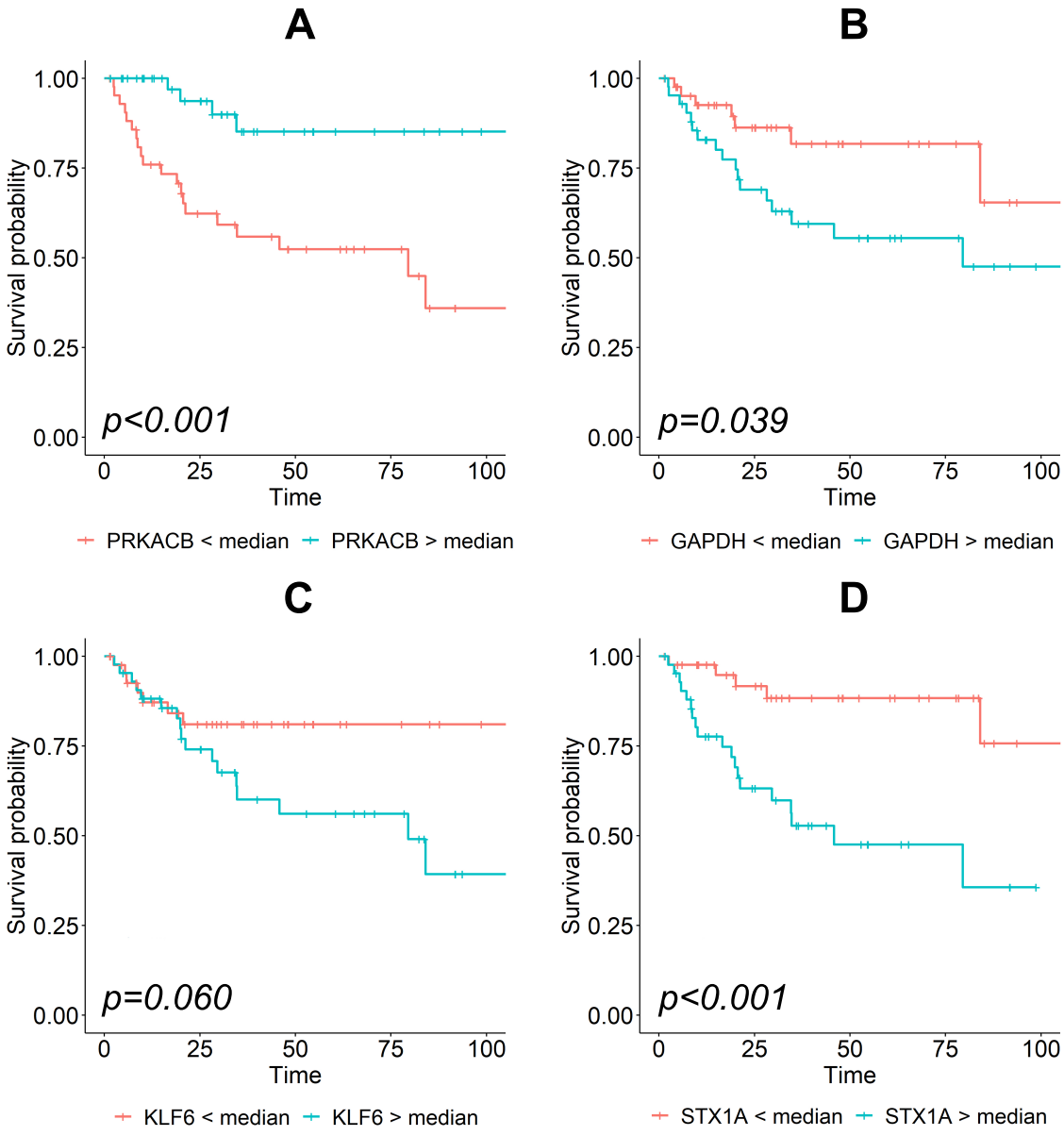


Figure 5.2: Kaplan-Meier plots for the lung adenocarcinoma dataset: Panels A, B, C, and D contain the survival probability for the PRKACB, GAPDH, KLF6, and STX1A genes, respectively. The green lines indicate levels of the gene higher than the median level, whereas the red lines indicate levels of the gene lower than the median level. In each panel, p-values for the log-rank test are provided.



Chapter 6

Discussion

In this thesis, we have proposed a sparse variable selection method that transforms the nonlinear estimating equation for the partial likelihood function into a linear estimating equation framework prior to performing a method of sparse variable selection. Although our primary focus was motivated by improving sparse variable selection within the context of clinical trials and cancer gene-expression profiling, the method can easily be extended to other low- and high-dimensional contexts within the Cox model.

In our simulation section, we found that the proposed method is capable of providing an excellent performance in both low-dimensional and high-dimensional settings, while being relatively unaffected by censoring rates of the simulated data sets. Additionally, we saw that all of the competing methods had performances that were heavily influenced by the sample size, whereas the proposed method was relatively consistent. Lastly, we also saw that increasing the number of covariates led to worse performance in competing methods, whereas our method tended to maintain similar, if not improved, results when the total number of covariates was increased.

In our analysis on the PBC data, we found that the proposed method is capable of detecting several significant covariates that were not deemed significant by any of the

competing methods. We were able to confirm that our method's choices were sensible based on existing literature and examination of the Kaplan-Meier curves, suggesting that our method was capable of detecting significant covariates in scenarios, where other approaches may provide seemingly reasonable, but under-fitted models.

In our real data analysis on lung adenocarcinoma data, we found that the proposed method is capable of detecting signal and joint effects on the survival outcome among many noisy covariates. The proposed method was able to efficiently select 28 significant genes associated with decreased survival in patients with lung adenocarcinoma, whereas competing sparse variable selection methods were unable to identify any signal amongst the genes. The results suggest that in future genetic studies, our method can be utilized as a reliable way to efficiently detect joint effects of genes associated with survival times, serving as an improvement over the popular forms of univariate analysis that have been commonly used within the field.

It should be noted that the ridge estimator chosen for the initial transformation can be sensitive and time-consuming to obtain in high-dimensional settings. In our simulations, we chose to utilize 10-fold cross validation to select our regularization parameter λ for better computational speed. However, in our real data application, LOOCV was utilized to select a more precise λ at the cost of computational time.

Lastly, in future work, we may look into alternatives to the SVS Gibbs sampler for the variable selection step. Although the SVS Gibbs sampler has demonstrated exceptional performance, improvements may be possible by devising a variable selection method specific to the proposed transformed linear estimating equation.

References

Ahn, M., Zhang, H. H., & Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, *22*(4), 1539-1562. doi: 10.5705/ss.2011.054

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., ... others (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, *8*, 816-824. doi: 10.1038/nm733

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability*, *29*(4), 885–895. doi: 10.2307/3214721

Bland, J. M., & Altman, D. G. (2004). The logrank test. *BMJ*, *328*(7447), 1073. doi: 10.1136/bmj.328.7447.1073

Broyden, C. G. (1967). Quasi-Newton methods and their application to function minimisation. *Mathematics of Computation*, *21*(99), 368–381. doi: 10.2307/2003239

Chen, Y., Gao, Y., Tian, Y., & Tian, D.-L. (2013). PRKACB is downregulated in non-small cell lung cancer and exogenous PRKACB inhibits proliferation and invasion of LTEP-A2 cells. *Oncology Letters*, *5*, 1803-1808. doi: 10.3892/ol.2013.1294

Cox, D. R. (1972). Regressioin models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-220. doi: 10.1111/j.2517-6161.1972.tb00899.x

DiFeo, A., Feld, L., Rodriguez, E., Wang, C., Beer, D. G., Martignetti, J. A., & Narla, G. (2008). A functional role for KLF6-SV1 in lung adenocarcinoma prognosis and chemotherapy response. *Cancer Research*, *68*(4), 965-970. doi: 10.1158/0008-5472.CAN-07-2604

Fan, J., & Jiang, J. (2009). Non- and semi-parametric modeling in survival analysis. In *New developments in biostatistics and bioinformatics* (p. 3-33). New Jersey: World Scientific & Higher Education.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101-148.

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis*. Hoboken, New Jersey: John Wiley & Sons.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, *13*(3), 317-322. doi: 10.1093/comjnl/13.3.317

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339-373.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, *24*(109), 23–26. doi: 10.2307/2004873

Greenwood, M. (1926). The natural duration of cancer. *Reports On Public Health and Medical Subjects*, *33*, 1-26.

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, *53*(4), 423-442. doi: 10.1111/j.1467-842X.2011.00641.x

Güler, E. N. (2017). Gene expression profiling in breast cancer and its effect on therapy selection in early-stage breast cancer. *European Journal of Breast Health*, *13*(4), 168-174. doi: 10.5152/ejbh.2017.3636

Hastie, T., & Tibshirani, R. (1990). *Generalized additive model*. New York: Chapman and Hall.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82. doi: 10.1080/00401706.1970.10488635

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer-Verlag.

Ishwaran, H., Kogalur, U. B., & Rao, J. (2010). spikeslab: Prediction and variable selection using spike and slab regression. *R Journal*, *2*(2), 68-73. doi: 10.32614/RJ-2010-018

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730-773. doi: 10.1214/009053604000001147

Janičko, M., Veselíny, E., Leško, D., & Jarčuška, P. (2013). Serum cholesterol is a significant and independent mortality predictor in liver cirrhosis patients. *Annals of Hepatology*, *12*(4), 413-419. doi: 10.1016/S1665-2681(19)31342-0

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457-481. doi: 10.2307/2281868

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, *66*, 221-244. doi: 10.1007/s10463-013-0429-6

Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, *10*(1), e0116774. doi: 10.1371/journal.pone.0116774

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535-1546. doi: 10.1080/01621459.1994.10476894

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023-1032. doi: 10.2307/2290129

Myers, D. J., & Wallen, J. M. (2019). *Cancer, lung adenocarcinoma*. Treasure Island, FL: StatPearls Publishing.

Nabi, R., & Su, X. (2017). coxphMIC: An R package for sparse estimation of Cox proportional hazards models via approximated information criteria. *R Journal*, *9*(1), 229-238. doi: 10.32614/RJ-2017-018

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681-686. doi: 10.1198/016214508000000337

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 185–207. doi: 10.2307/2344317

Puzone, R., Savarino, G., Salvi, S., Dal Bello, M. G., Barletta, G., Genova, C., ... Pfeffer, U. (2013). Glyceraldehyde-3-phosphate dehydrogenase gene over expression correlates with poor prognosis in non-small cell lung cancer patients. *Molecular Cancer*, *12*, 97.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461 – 464. doi: 10.1214/aos/1176344136

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*(111), 647–656. doi: 10.2307/2004840

Singh, K., Gupta, N., et al. (2012). Palateless custom bar supported overdenture: A treatment modality to treat patient with severe gag reflex. *Indian Journal of Dental Research*, *23*(2), 145-148. doi: 10.4103/0970-9290.100416

Su, X., Wijayasinghe, C. S., Fan, J., & Zhang, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics*, *72*(3), 751–759. doi: 10.1111/biom.12484

Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*, *33*(18), 2799-2807. doi: 10.1093/bioinformatics/btx300

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

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

Uddenfeldt, P., & Danielsson, Å. (2000). Primary biliary cirrhosis: survival of a cohort followed for 10 years. *Journal of Internal Medicine*, *248*(4), 292-298. doi: 10.1046/j.1365-2796.2000.00733.x

Volinsky, C. T., & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, *56*(1), 256-262. doi: 10.1111/j.0006-341X.2000.00256.x

Xu, J. (2012). High-dimensional Cox regression analysis in genetic studies with censored survival outcomes. *Journal of Probability and Statistics*, *2012*, 478680. doi: 10.1155/2012/478680

Yang, H., Baladandayuthapani, V., Rao, A. U., & Morris, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *Journal of the American Statistical Association*, *115*(529), 90-106. doi: 10.1080/01621459.2019.1609969

Yang, H., Zhu, H., & Ibrahim, J. G. (2018). MILFM: Multiple index latent factor model based on high-dimensional features. *Biometrics*, *74*(3), 834-844. doi: 10.1111/biom.12866

Zhang, H. H., & Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, *94*(3), 691–703. doi: 10.1093/biomet/asm037