# VARIABLE SELECTION FOR INTERVAL-CENSORED AND FUNCTIONAL SURVIVAL DATA

---

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

---

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

TIAN TIAN

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

VARIABLE SELECTION FOR INTERVAL-CENSORED
AND FUNCTIONAL SURVIVAL DATA

presented by Tian Tian,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

---

Dr. (Tony) Jianguo Sun

---

Dr. Guanyu Hu

---

Dr. Zhuoqiong He

---

Dr. Xinghe Wang

---

# ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my esteemed advisor Dr. (Tony) Jianguo Sun for his critical inspiration, unflinching encouragement and continuous support throughout the long journey of my doctorate study. It was my great honor to be one of his students and study under his guidance.

I extend my sincere gratitude to my advisory committee members: Drs. Guanyu Hu, Zhuoqiong He and Xinghe Wang for their insightful comments and suggestions on my work.

I also owe a debt of gratitude to all faculty in the Department of Statistics who have taught, encouraged, and assisted me during my graduate studies. I would like to thank our excellent staff Judy, Kathleen, Abbie, and Laura, for their generous help. I am so grateful to Dr. Larry Ries for helping me become a better instructor.

Moreover, I would like to extend many thanks to my cherished friends who helped me through difficult times, cheered me on, and made my life at Mizzou much fun and memorable.

Finally, my deepest gratitude goes to my husband, Dr. Qiwei Wu, my parents, Hui Zhu and Zongpeng Tian, my lovely and supportive parents-in-law, and other family members for their unconditional love, support, encouragement, and caring. Special thanks to my husband for bringing sunshine into my life in every dark moment, without whom I would not have been able to complete my Ph.D. study.

# TABLE OF CONTENTS

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# Variable Selection for Interval-censored and Functional Survival Data

Tian Tian

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

## ABSTRACT

Interval-censored data are a type of failure time data that is only known to belong to a time interval but cannot be observed precisely. Note that interval-censoring is often encountered in medical or health studies with periodic follow-ups nature and includes right- or left-censoring as a particular case. Moreover, the linear form of covariate effects in various survival models, such as the commonly used standard Cox model, may not always be a realistic assumption. To relax this, additive models, which assume nonlinear covariate effects, are useful alternatives to accommodate such nonlinearity. Recently, more and more attention has been drawn to the provision of variable selection for survival and functional data analysis when plenty of risk factors are available. But limited literature has investigated variable selection for the functional survival models with interval-censored data. In the dissertation, we shed light on variable selection for a series of additive survival models with high-dimensional interval-censored data via penalized estimation to address different statistical complexities.

In Chapter 2, we will focus on high-dimensional variable selection for the nonparametric additive Cox model with interval-censored failure time data to identify important risk factors with potential nonlinear covariate effects. For the problem,

we propose a penalized sieve maximum likelihood approach with the use of Bernstein polynomials approximation and group penalization. To implement the proposed method, an efficient group coordinate descent algorithm is developed and can be easily carried out for both low- and high-dimensional scenarios. Furthermore, a simulation study is performed to assess the performance of the presented approach and suggests that it works well in practice. The proposed method is applied to an Alzheimer's Disease Neuroimaging Initiative (ADNI) study to identify important and relevant genetic factors.

In Chapter 3, we will extend the method given in Chapter 2 to a broad class of nonparametric additive transformation models. A two-step regularization estimation procedure that combines Bernstein polynomials expansions with a Poisson-based data augmentation penalized expectation-maximization (EM) algorithm is developed for implementation. The proposed method is assessed through an extensive simulation study, and the results suggest that it works well in various scenarios. Finally, we apply the presented method to the ADNI data described in Chapter 2 with demographic and clinical factors for illustration.

In Chapter 4, we will perform the variable selection for a novel partially functional additive Cox model with both functional and scalar predictors under interval-censored failure time data and sparsely observed functional data. Specifically, we adopt Bernstein polynomials approximation to model the unspecified cumulative baseline hazard function and additive components and apply functional principal component analysis to extract functional features from trajectories of functional covariates. For implementation, a penalized sieve estimation approach with multiple group penalty functions is investigated, and a group coordinate descent algorithm is used. A simulation study

is conducted to demonstrate the finite-sample performance of the proposed method. The method is applied to the ADNI study introduced in the previous two chapters for illustration.

# Chapter 1

# Introduction

## 1.1 Variable Selection and Penalized Estimation

Variable selection plays an increasingly important role in the advancement of statistical techniques. A frequently encountered problem in regression analysis is which variables should be included in the model, especially when an enormous number of variables are available. Various traditional variable selection approaches such as stepwise, forward, forward-stage, or best subset regression have been investigated in the literature. However, these conventional methods exhibit instability and computational infeasibility, even in moderate dimensions.

To overcome this, other techniques such as penalization, also referred to as regularization or penalty-based methods, were well developed in literature, particularly in linear models. Most of them can simultaneously perform parameter estimation and variable selection, except ridge regression. Specifically, let $\boldsymbol{Y}$ be the response

vector, denote $\boldsymbol{X}$ as the covariate matrix, and $\boldsymbol{\beta}$ as the regression parameter vector, the penalized estimation for a linear model has the following expression:

$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}}{argmin} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \right\}, \tag{1.1}$$

where $P_\lambda(\boldsymbol{\beta})$ is a penalty function controlled by an unknown tuning parameter $\lambda$. The ridge regression is equivalent to using the penalty function with form $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2^2$ which is also referred to as $l_2$ penalty.

Tibshirani (1996) developed a least absolute shrinkage and selection operator (LASSO), also known as $l_1$ penalty, takes the form $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ in (1.1). The LASSO penalty is popular and easily implemented, but it lacks oracle property and may involve some noises and yield inconsistent variable selection results. The oracle property indicates that the final estimator performs as well as the one if the true model is known. Furthermore, other penalty functions were explored with enhanced selection accuracy. Among others, smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP), and adaptive LASSO (ALASSO) penalty functions were proposed by several authors, all of which possess the oracle property (Fan and Li, 2001; Zhang, 2010; Zou, 2006).

When group structures appear, several methods have been proposed for group selection that select or exclude a group of variables at the same time. Yuan and Lin (2006) investigated the group LASSO and employed the group coordinate descent for computation. The group LASSO is a group version of the LASSO penalty that tends to behave similarly to the LASSO, which may not achieve selection consistency and include some unimportant variables in the model with a high false positive rate. Huang *et al.* (2012) provided a systematic review of group selection and bi-level selec-

tion for linear models, and several concave group selection techniques were proposed, for example, the group MCP and SCAD penalization approaches.

## 1.2 Variable Selection for Survival Data

### 1.2.1 Regression Analysis of Failure Time Data

Variable selection for censored failure time data has attracted considerable attention in recent decades. The Cox's proportional hazards model (Cox, 1972) is likely the most commonly used one among all survival models, which takes the following form

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp\{\boldsymbol{X}\boldsymbol{\beta}\}$$

with $\lambda_0(t)$ being an unknown baseline hazard function. In addition to the Cox model, multiple survival models, such as proportional odds, additive hazards, and accelerated failure time models, are also popular and widely used in the problem of time-to-event data.

Notably, a broad class of semiparametric transformation models is another appealing approach to analyze survival data, which includes the Cox model and proportional odds model as special cases. The transformation model is flexible and has the following expression

$$\lambda(t|\boldsymbol{X}) = G\left[\lambda_0(t) \exp\{\boldsymbol{X}\boldsymbol{\beta}\}\right],$$

where $G(\cdot)$ is a pre-specified increasing transformation function and $\lambda_0(t)$ represents an unknown baseline hazard function (Dabrowska and Doksum, 1988). The right-

censored failure time data is frequently assumed in survival analysis, which means that the exact failure time is either observed or only known to be greater than a time point. The observed data under right-censoring is defined as $\{T_i^* = \min(T_i, C_i), \delta_i = \mathrm{I}(T_i \leq C_i)\}_{\{i=1,\ldots,n\}}$, where $T_i^*$, $T_i$, $C_i$ and $\delta_i$ are the observation time, the event time, the censoring time and the censoring indicator, respectively, for subject $i$.

Nevertheless, the interval-censored failure time data is often more appropriate than right-censoring in many cases where individuals have regular examination times. For the interval-censoring, we usually mean that the failure time of interest cannot be precisely observed but instead be observed or known to be within an interval (Sun, 2006). A typical example where such data occur is periodic follow-up studies such as clinical trials, in which study subjects are usually observed only from time to time or at discrete time points. It is apparent that interval-censored data include right-censored data as a special case.

Note that, in general, the analysis of interval-censored data is much more challenging than right-censored data due to the complicated data structure caused by interval censoring. One such situation is the analysis of right-censored data under the standard Cox model, for which a partial likelihood function that is independent of the baseline hazard function is available and thus makes the analysis as well as variable selection much more straightforward. In contrast, the same is not true for interval-censored data.

### 1.2.2 Variable Selection for Survival Models

A great deal of research for variable selection with right-censored failure time data has been studied. Tibshirani (1997) proposed penalized partial likelihood function for the

Cox model with the LASSO penalty. Fan and Li (2002) extended the SCAD penalty for the Cox model and Cox frailty model. Zhang and Lu (2007) carried out variable selection for the Cox model with the use of the ALASSO penalty. Liu and Zeng (2013) further extended the ALASSO to the semiparametric transformation models by using a penalized weighted negative partial log-likelihood function.

As mentioned earlier, analysis of interval-censored data is more challenging but sometimes more realistic and reasonable than right-censoring due to the unavailability of the partial likelihood function free of a baseline hazard function. Several methods have been proposed for variable selection based on interval-censored failure time data. Among others, Wu and Cook (2015) and Zhao *et al.* (2020) considered the selection problem under the Cox model and, moreover, Wu and Cook (2015) assumed that the baseline hazard function is piecewise constant. Furthermore, Zhao *et al.* (2020) proposed a broken adaptive ridge-based (BAR) penalized procedure and established the oracle property of the method. In addition, Scolas *et al.* (2016), Wu *et al.* (2020), Li *et al.* (2020) and Tian and Sun (2022) investigated the same problem but under a parametric mixture cure model, a partly linear Cox model with two sets of covariates, a class of semiparametric transformation models and the additive Cox model, respectively.

## 1.3 Variable Selection for Functional Data

Functional data analysis (FDA) is to analyze data that are in the form of functions. Functional data are infinite-dimensional data whose observations are functions defined on a continuum (e.g., time, space) but sampled at a finite number of points (Wang

*et al.*, 2016a). Functional data are possibly measured from dense or spare sampling schemes. Dense functional data are usually observed in a fine regular grid, such as spectral data and imaging data. Sparse functional data are often with observations taken irregularly and sparsely, for example, longitudinal trajectories of CD4 count and blood pressure.

Functional principal component analysis (FPCA) is one of the most prevalent techniques in the FDA that facilitates dimensionality reduction and extraction of functional features. Through FPCA, the intrinsically infinite dimensional functional data can be reduced to a finite-dimensional vector of random scores. The FPCA can capture the primary pattern of variations in the observations and decompose the stochastic process into multiple uncorrelated functional principal components (FPCs). As a result, these FPCs can be further treated as new covariates in the regression models.

Such a FPCA approach has been widely applied for functional regression models (Yao *et al.*, 2005a; Hall and Horowitz, 2007; Hall *et al.*, 2006; Morris, 2015; Yao *et al.*, 2005b). When functional data are sparely sampled, Yao *et al.* (2005b) proposed the principal analysis by conditional expectation (PACE) approach, and asymptotic properties were further established by Hall *et al.* (2006). Essentially, the PACE method utilizes the local polynomial regression method to estimate the covariance function and then achieves the estimates of FPCs by eigendecomposition of the estimated covariance function. The associated FPC score is estimated by conditional expectation with the assumption that the FPC scores follow a Gaussian distribution.

A rich literature has investigated variable selection for functional regression models. Among others, Kong *et al.* (2016) proposed penalized estimation for partially

functional linear regression with the use of the FPCA, and group smoothly clipped absolute deviation (SCAD) and SCAD penalty functions were applied for regularization of functional and scalar predictors, respectively. Matsui and Konishi (2011) developed the variable selection for functional regression models through the GSCAD penalty. Aneiros *et al.* (2015) investigated a nonconcave-penalized least squares estimation for high-dimensional partial linear regression with possibly functional variables. Aneiros *et al.* (2022) offered a comprehensive review of methodological advancements for variable selection in functional regression models.

## 1.4 Outline of the Dissertation

The remainder of this dissertation is organized as follows.

In Chapter 2, we will focus on variable selection for the nonparametric additive Cox model with interval-censored failure time data. For the problem, we propose a penalized sieve maximum likelihood approach with the use of Bernstein polynomials approximation and group penalization. To implement the proposed method, an efficient group coordinate descent algorithm is developed and can be easily carried out for both low- and high-dimensional scenarios. Furthermore, a simulation study is performed to assess the performance of the presented approach and suggests that it works well in practice. The proposed method is applied to an Alzheimer's disease study to identify important and relevant genetic factors.

In Chapter 3, we will extend the variable selection for interval-censored data to a broad class of nonparametric additive transformation models. A two-step regularization estimation procedure that combines Bernstein polynomials with a Poisson-based

data augmentation penalized expectation-maximization (EM) algorithm is developed for implementation. The proposed method is assessed through an extensive simulation study, and the results suggest that it works well in various scenarios. Finally, we apply the presented method to Alzheimer's Disease Neuroimaging Initiative (ADNI) data for illustration.

In Chapter 4, we will perform the variable selection for a novel partially functional additive Cox model with both functional and scalar predictors under interval-censored failure time data and sparsely and irregularly observed functional data. Specifically, we adopt Bernstein polynomials approximation to model unspecified cumulative baseline hazard functions and additive components and apply functional principal component analysis to exact functional features from trajectories of functional covariates. For implementation, a penalized sieve estimation approach with multiple group penalty functions is investigated, and a group coordinate descent algorithm is used. A simulation study is conducted to demonstrate the finite-sample performance of the proposed method. The method is applied to the ADNI study with genetic and clinical factors for illustration. Finally, several future research directions will be discussed in Chapter 5.

# Chapter 2

# Variable Selection for Nonparametric Additive Cox Model with Interval-censored Data

## 2.1 Introduction

As mentioned in Chapter 1, the standard Cox model is perhaps the most commonly used model for regression analysis of failure time data. One limitation is that it assumes that all covariate effects are linear. To relax this, the nonparametric additive Cox model, which allows for nonlinear covariate effects, has been proposed and investigated by some authors (Cai *et al.*, 2007; Cheng and Wang, 2011; Huang, 1999; Lu and McMahan, 2018). For the nonparametric additive Cox model, several authors have discussed variable selection problems for such model, but they only considered right-censored data (Du *et al.*, 2010; Lv *et al.*, 2018). However, the interval-censored mechanism is more realistic and appropriate in many real-world problems, especially

in medical and clinical trials with periodic examination times.

In this chapter, we consider the situation where the failure time of interest follows the nonparametric additive Cox model that allows nonlinear covariate effects and only interval-censored data are observed, for which it seems that there is no existence of an established variable selection procedure. For the problem, a penalized sieve maximum likelihood approach is proposed with the use of Bernstein polynomials and group penalty functions. The model considered here is much more general than and includes the models considered in Wu and Cook (2015) and Wu *et al.* (2020) as special cases, and in particular, unlike the existing methods, the proposed method can deal with any nonlinear covariate effects.

The remainder of this chapter is organized as follows. After introducing some notation and assumptions that will be used throughout the chapter, the proposed penalized sieve variable selection procedure will be presented in Section 2.2. In the method, Bernstein polynomials will be employed to approximate unknown functions and group penalty functions will be used for the selection of relevant nonlinear covariate effects. For the implementation of the proposed method, an efficient group coordinate descent algorithm will be developed in Section 2.3, which works for both low- and high-dimensional scenarios. Section 2.4 presents some results obtained from a simulation study to assess the finite sample performance of the proposed procedure, and they suggest that the approach works well for practical situations. In Section 2.5, the proposed method is applied to an Alzheimer's Disease study and Section 2.6 provides some concluding discussion and remarks.

## 2.2 Penalized, Sieve Variable Selection Procedure

Consider a failure time study that involves $n$ independent individuals. For subject $i$, let $T_i$ denote the associated failure time of interest and suppose that there exists a $p$-dimensional vector of covariates denoted by $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$. In the following, it will be assumed that $p$ could be either smaller or larger than the sample size $n$. To describe the covariate effects, we will assume that given $\boldsymbol{X}_i$, the cumulative hazard function of $T_i$ has the form

$$\Lambda(t|\boldsymbol{X}_i) = \Lambda_0(t)\, \exp\{\boldsymbol{\psi}(\boldsymbol{X}_i)\}\,, \tag{2.1}$$

where $\Lambda_0(t)$ denotes an unknown baseline cumulative hazard function and $\boldsymbol{\psi}(\boldsymbol{X}_i) = \sum_{j=1}^{p} \psi_j(X_{ij})$ with $\psi_j$ also being an unknown function. That is, $T_i$ follows the nonparametric additive Cox model with nonlinear covariate effects. For the identifiability of the model above, it will be assumed that $E\left[\psi_j(\cdot)\right] = 0$, $1 \leq j \leq p$ (Huang, 1999).

Suppose that the observed data on the $T_i$'s have the form $\mathcal{O} = \{(L_i, R_i], \boldsymbol{X}_i\}_{i=1}^{n}$, where $(L_i, R_i]$ indicates the interval that brackets $T_i$. That is, only interval-censored data are available. It is easy to see that $L_i = 0$ or $R_i = \infty$ corresponds to a left- or right-censored observation on $T_i$, respectively. In following, we will assume an independent or noninformative censoring mechanism, meaning that censoring intervals are independent of the failure times of interest (Sun, 2006). Then the observed likelihood function can be written as

$$L(\Lambda_0, \boldsymbol{\psi}) = \prod_{i=1}^{n} \left\{\exp\left[-\Lambda(L_i)\right] - \exp\left[-\Lambda(R_i)\right]\right\}$$

$$= \prod_{i=1}^{n} \left\{ \exp\left[ -\Lambda_0(L_i) e^{\sum_{j=1}^{p} \psi_j(X_{ij})} \right] - \exp\left[ -\Lambda_0(R_i) e^{\sum_{j=1}^{p} \psi_j(X_{ij})} \right] \right\} .$$

For variable selection, it would be natural to develop a penalized procedure based on the likelihood function above. On the other hand, it is apparent that this would not be easy since it involves unknown functions $\Lambda_0(t)$ and the $\psi_j(\cdot)$'s. To deal with this, by following Zhao *et al.* (2020) and others, we will first apply the sieve approach to approximate them by using Bernstein polynomials.

To illustrate the Bernstein polynomials approximation, consider defining the parameter space

$$\Theta = \{(\Lambda_0, \psi_1, ..., \psi_p) \in \mathcal{M}_0 \otimes \mathcal{M}_1 \otimes ... \otimes \mathcal{M}_p\}$$

for all unknown functions. Here $\mathcal{M}_0$ denotes the collection of all bounded and continuous nondecreasing, nonnegative functions over the interval $[c, u]$ with $c$ and $u$ usually taken to be $\min(L_i)$ and $\max(R_i)$ $(i = 1, ..., n)$, respectively, and $\mathcal{M}_j$ the collection of all bounded and continuous functions over the interval $[c_j, u_j]$ with $c_j$ and $u_j$ usually set to be $\min(X_j)$ and $\max(X_j)$, respectively, $j = 1, ..., p$. Also define the sieve space

$$\Theta_n = \{(\Lambda_{0n}, \psi_{1n}, ..., \psi_{pn}) \in \mathcal{M}_{0n} \otimes \mathcal{M}_{1n} \otimes ... \otimes \mathcal{M}_{pn}\} ,$$

where

$$\mathcal{M}_{0n} = \left\{ \Lambda_{0n}(t) = \sum_{k=0}^{m} \phi_k^* B_{0k}(t, m, c, u) : \sum_{0 \le k \le m} |\phi_k^*| \le M_{0n}, 0 \le \phi_0^* \le \phi_1^* \le ... \le \phi_m^* \right\} ,$$

and

$$\mathcal{M}_{jn} = \left\{ \psi_{jn}(X_j) = \sum_{k=0}^{m_j} \alpha_{jk} B_{jk}(X_j, m_j, c_j, u_j) : \sum_{0 \le k \le m_j} |\alpha_{jk}| \le M_{jn} \right\}, j = 1, ..., p.$$

In the above,

$$B_{0k}(t, m, c, u) = \binom{m}{k} \left( \frac{t-c}{u-c} \right)^k \left( 1 - \frac{t-c}{u-c} \right)^{m-k}, k = 0, 1, ..., m,$$

and

$$B_{jk}(X_j, m_j, c_j, u_j) = \binom{m_j}{k} \left( \frac{X_j - c_j}{u_j - c_j} \right)^k \left( 1 - \frac{X_j - c_j}{u_j - c_j} \right)^{m_j - k}, k = 0, 1, ..., m_j,$$

which are the Bernstein basis polynomials of $m$ and $m_j$ degrees, respectively. $M_{0n}$ and $M_{jn}$ can be used to control the size of sieve space (Zhou *et al.*, 2017). Note that in $\mathcal{M}_{0n}$, the constraint $0 \le \phi_0^* \le \phi_1^* \le ... \le \phi_m^*$ is needed to guarantee the monotonic properties of $\Lambda_{0n}(t)$ (Wang and Ghosh, 2012) and can be easily eliminated by the re-parametrization $\phi_0^* = e^{\phi_0}$, $\phi_l^* = \sum_{i=0}^{l} e^{\phi_i}$, $\forall\ 1 \le l \le m$.

It is worth noting that with the use of Bernstein polynomial approximations, one can reduce an infinite-dimensional estimation problem to a finite-dimensional estimation problem and thus it has the advantages of relatively easy implementation and computational efficiency (Zhou *et al.*, 2017). It is apparent that instead of Bernstein polynomials, one could apply other smooth functions such as spline functions or other polynomials. Some advantages of the use of Bernstein polynomials include that the needed monotonicity and nonnegativity constraints can be easily satisfied by some constraints on the parameters (Wang and Ghosh, 2012). In addition, among

all polynomials, the Bernstein polynomial has the optimal shape-preserving property (Carnicer and Pena, 1993) and it does not require the selection of interior knots.

Let $\boldsymbol{\phi} = (\phi_0, ..., \phi_m)'$ and $\boldsymbol{\alpha} = (\alpha_{10}, ..., \alpha_{1m_1}, ..., \alpha_{p0}, ..., \alpha_{pm_p})'$. Note that if the goal is to estimate $\Lambda_0$ and $\boldsymbol{\psi}$, it is natural to maximize the likelihood function $L(\Lambda_0, \boldsymbol{\psi})$ over the sieve space $\Theta_n$ or minimize the following negative log-likelihood function

$$\ell(\boldsymbol{\phi}, \boldsymbol{\alpha}) = -\frac{1}{n}\sum_{i=1}^{n} \log\left\{\exp\left(-\Lambda_{0n}(L_i)e^{\sum_{j=1}^{p}\psi_{jn}(X_j)}\right) - \exp\left(-\Lambda_{0n}(R_i)e^{\sum_{j=1}^{p}\psi_{jn}(X_j)}\right)\right\}$$

with respect to $\{\Lambda_{0n}, \psi_{1n}, ..., \psi_{pn}\}$. This suggests that for the variable selection on $\boldsymbol{X}$, we can minimize the penalized negative log-likelihood function

$$\ell_p(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \ell(\boldsymbol{\phi}, \boldsymbol{\alpha}) + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j)\,, \tag{2.2}$$

where $P_\lambda(\boldsymbol{\alpha}_j)$ denotes a penalty function of $\boldsymbol{\alpha}_j = (\alpha_{j0}, ..., \alpha_{jm_p})$ characterized by the tuning parameter $\lambda$. It is easy to see that $\boldsymbol{\alpha}_j = \mathbf{0}$ suggests that the corresponding covariate $X_j$ has no significant effect.

For the penalty function $P_\lambda(\boldsymbol{\alpha}_j)$, we will consider the following three group penalty functions. One is the group LASSO penalty function given by

$$P_\lambda(\boldsymbol{\alpha}_j) = \lambda_j \|\boldsymbol{\alpha}_j\|_2$$

with $\lambda_j = \lambda\sqrt{m_j + 1}$ and $\|\boldsymbol{\alpha}_j\|_2 = \sqrt{\sum_{l=0}^{m_j}\alpha_{jl}^2}$ (Yuan and Lin, 2006), and another is

the group SCAD penalty function

$$P_\lambda(\boldsymbol{\alpha}_j; a) = \begin{cases} \lambda_j \|\boldsymbol{\alpha}_j\|_2 & \text{if } \|\boldsymbol{\alpha}_j\|_2 \le \lambda_j\,, \\ \dfrac{2a\lambda_j\|\boldsymbol{\alpha}_j\|_2 - \|\boldsymbol{\alpha}_j\|_2^2 - \lambda_j^2}{2(a-1)} & \text{if } \lambda_j < \|\boldsymbol{\alpha}_j\|_2 \le a\lambda_j\,, \\ \dfrac{(a^2-1)\lambda_j^2}{2(a-1)} & \text{if } \|\boldsymbol{\alpha}_j\|_2 > a\lambda_j \end{cases}$$

with $\lambda_j = \lambda\sqrt{m_j + 1}$ and $a > 2$ (Fan and Li, 2001; Huang *et al.*, 2012). The third is the group MCP penalty function

$$P_\lambda(\boldsymbol{\alpha}_j; \gamma) = \begin{cases} \lambda_j\|\boldsymbol{\alpha}_j\|_2 - \dfrac{\|\boldsymbol{\alpha}_j\|_2^2}{2\gamma} & \text{if } \|\boldsymbol{\alpha}_j\|_2 \le \lambda_j\gamma\,, \\ \dfrac{\lambda^2\gamma}{2} & \text{if } \|\boldsymbol{\alpha}_j\|_2 > \lambda_j\gamma \end{cases}$$

with $\lambda_j = \lambda\sqrt{m_j + 1}$ and $\gamma > 1$ (Zhang, 2010; Huang *et al.*, 2012). Of course, one could apply other group penalty functions too.

## 2.3 An Efficient Group Coordinate Descent Algorithm

Let $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\alpha}}$ denote the estimators of $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ given by the minimization of the penalized negative log-likelihood function $\ell_p(\boldsymbol{\phi}, \boldsymbol{\alpha})$. In the following, for the determination of $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\alpha}}$, we will develop an optimization algorithm that estimates $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ alternately by following the idea of the group coordinate descent algorithm (Yuan and Lin, 2006; Yang and Zou, 2015; Breheny and Huang, 2015; Cao *et al.*, 2016; Lv *et al.*, 2018).

First we will consider the determination of $\hat{\boldsymbol{\alpha}}$ and for this, we will update each $\boldsymbol{\alpha}_j$ while keeping all other elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ fixed at their current estimates. More specifically, to update $\boldsymbol{\alpha}_j$ at the $(k+1)^{\text{th}}$ iteration, let $\hat{\boldsymbol{\phi}} = (\hat{\phi}_0^{(k+1)}, ..., \hat{\phi}_m^{(k+1)})$ and $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1^{(k+1)}, ..., \hat{\boldsymbol{\alpha}}_{j-1}^{(k+1)}, \hat{\boldsymbol{\alpha}}_j^{(k)}, ..., \hat{\boldsymbol{\alpha}}_p^{(k)})$ denote the current estimate of $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$, respectively. Also define $\ell_j'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) = \partial \ell(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}})/\partial \boldsymbol{\alpha}_j$, the $j$th sub-gradient vector, $H_j^{(k)} = \partial^2 \ell(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}})/\partial \boldsymbol{\alpha}_j^2$, the $j$th sub-Hessian matrix of $\ell(\boldsymbol{\phi}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}_j$ evaluated at $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\alpha}}$, and $h_j^{(k)}$ to be the largest eigenvalue of $H_j^{(k)}$. Note that both $\ell_j'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}})$ and $H_j^{(k)}$ depend on the current estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\alpha}}$. Let $\boldsymbol{\alpha}^* = (\hat{\boldsymbol{\alpha}}_1^{(k+1)}, ..., \hat{\boldsymbol{\alpha}}_{j-1}^{(k+1)}, \boldsymbol{\alpha}_j, \hat{\boldsymbol{\alpha}}_{j+1}^{(k)}, ..., \hat{\boldsymbol{\alpha}}_p^{(k)})$. The penalized negative log-likelihood function $\ell_p(\hat{\boldsymbol{\phi}}, \boldsymbol{\alpha}^*)$ can be approximated by

$$\begin{aligned}
\ell_p(\hat{\boldsymbol{\phi}}, \boldsymbol{\alpha}^*) &\approx \ell(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) + \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right]' \ell_j'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) \\
&\quad + \frac{1}{2}\left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right]' H_j^{(k)} \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right] + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j) \\
&\approx \ell(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) + \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right]' \ell_j'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) \\
&\quad + \frac{h_j^{(k)}}{2}\left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right]' \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right] + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j),
\end{aligned}$$

with the sub-Hessian matrix $H_j^{(k)}$ being replaced by its largest eigenvalue $h_j^{(k)}$. The updated estimate $\hat{\boldsymbol{\alpha}}_j^{(k+1)}$ of $\boldsymbol{\alpha}_j$ can be determined by

$$\hat{\boldsymbol{\alpha}}_j^{(k+1)} = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmin}} \left\{ \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\right]' \ell'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}}) + \frac{h_j^{(k)}}{2} \|\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(k)}\|_2^2 + P_\lambda(\boldsymbol{\alpha}_j) \right\}$$

and the closed-form solutions for various penalty functions will be shown subsequently. Here, one can employ group coordinate descent algorithm and determine each $\hat{\boldsymbol{\alpha}}_j^{(k+1)}$

$(j = 1, ..., p)$ individually while keeping all other parameters fixed at their current estimates. Note that Lv *et al.* (2018) suggested that $\hat{\boldsymbol{\alpha}}_j^{(k+1)}$ $(j = 1, ..., p)$ can be determined in parallel.

With the use of the group LASSO penalty function, for example, we have that

$$\hat{\boldsymbol{\alpha}}_j^{(k+1)} = S(\boldsymbol{c}_j; \lambda_j/h_j)\,,$$

where $\lambda_j = \lambda\sqrt{m_j + 1}$ and $S(\boldsymbol{c}_j; \lambda) = (1 - \lambda/\|\boldsymbol{c}_j\|)_+\boldsymbol{c}_j$ with $\boldsymbol{c}_j = \hat{\boldsymbol{\alpha}}_j^{(k)} - \ell_j'(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\alpha}})/h_j$. With the use of the group SCAD penalty function, it gives

$$\hat{\boldsymbol{\alpha}}_j^{(k+1)} = \begin{cases} S(\boldsymbol{c}_j; \lambda_j/h_j) & \text{if } \|\boldsymbol{c}_j\|_2 \le \lambda_j + \lambda_j/h_j\,, \\ \dfrac{\left[h_j(a-1) - \frac{a\lambda_j}{\|\boldsymbol{c}_j\|_2}\right]\boldsymbol{c}_j}{h_j a - h_j - 1} & \text{if } \lambda_j + \lambda_j/h_j < \|\boldsymbol{c}_j\|_2 \le a\lambda_j\,, \\ \boldsymbol{c}_j & \text{if } \|\boldsymbol{c}_j\|_2 > a\lambda_j. \end{cases}$$

If the group MCP penalty function is used, we have that

$$\hat{\boldsymbol{\alpha}}_j^{(k+1)} = \begin{cases} S\left(\dfrac{h_j\boldsymbol{c}_j}{h_j - 1/\gamma}; \dfrac{\lambda}{h_j - 1/\gamma}\right) & \text{if } \|\boldsymbol{c}_j\|_2 \le \lambda_j\gamma\,, \\ \boldsymbol{c}_j & \text{if } \|\boldsymbol{c}_j\|_2 > \lambda_j\gamma. \end{cases}$$

Note that as pointed out before, all $\psi_j(X_j)$'s need to be centered to avoid the model identification issue. For this, define

$$\hat{\psi}_{jn}^*(X_j) = \sum_{k=0}^{m_j} \hat{\alpha}_{jk}^* B_{jk}(X_j, m_j, c_j, u_j) \quad \text{and} \quad \bar{\psi}_{jn}^*(X_j) = \sum_{i=1}^{n} \hat{\psi}_{jn}^*(X_{ij})/n\,.$$

The final estimator of $\psi_j(X_j)$ will be defined as

$$\hat{\psi}_{jn}(X_j) = \hat{\psi}^*_{jn}(X_j) - \bar{\psi}^*_{jn}(X_j)\,, \qquad 1 \le j \le p\,.$$

For the determination of the estimate of $\boldsymbol{\phi}$ in the iteration, we suggest employing the Nelder-Mead simplex algorithm. The algorithm discussed above can be summarized as follows.

Step 1: Set $k = 0$ and give initial estimates $\hat{\boldsymbol{\phi}}^{(0)}$ and $\hat{\boldsymbol{\alpha}}^{(0)}$.

Step 2: At the $(k+1)^{\text{th}}$ iteration, determine $\hat{\boldsymbol{\phi}}^{(k+1)}$ by using the Nelder-Mead simplex algorithm with $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(k)}$.

Step 3: Also at the $(k+1)^{\text{th}}$ iteration, determine $\hat{\boldsymbol{\alpha}}_j^{(k+1)}$ by using the group coordinate descent algorithm by solving

$$\hat{\boldsymbol{\alpha}}_j^{(k+1)} = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmin}} \left\{ l(\hat{\boldsymbol{\phi}}^{(k+1)}, \boldsymbol{\alpha}^*) + P_\lambda(\boldsymbol{\alpha}_j) \right\}$$

for $j = 1, ..., p$. Then center $\hat{\psi}^*_{jn}(X_j)$ for each $j = 1, ..., p$.

Step 4: Repeat Steps 2 and 3 until convergence or $k$ exceeding a given large number.

To implement the algorithm above, it is apparent that one needs to choose a convergence criterion and the tuning parameter $\lambda$. For the stopping criterion, a commonly used one is to base the convergence on the absolute difference between two consecutive estimates or to stop the iteration if

$$\text{mean}\left[ \left| (\hat{\boldsymbol{\phi}}^{(k+1)}, \hat{\boldsymbol{\alpha}}^{(k+1)}) - (\hat{\boldsymbol{\phi}}^{(k)}, \hat{\boldsymbol{\alpha}}^{(k)}) \right| \right] < \epsilon$$

for a pre-selected constant $\epsilon$, which can be set to be $10^{-5}$ or $10^{-6}$. For the selection of $\lambda$, many methods can be used and in the numerical study below, we employ the extended Bayesian information criterion (EBIC) given by

$$EBIC(\lambda) = -2\,\ell(\boldsymbol{\phi}, \boldsymbol{\alpha}) + s(\lambda)\,\log(n) + 2\,\Gamma\,\log\binom{p}{s(\lambda)}$$

with $\Gamma = 1$ (Chen and Chen, 2008). In the above, $s(\lambda)$ denotes the total number of nonzero estimated covariate effects.

## 2.4 A Simulation Study

In this section, we present some results obtained from an extensive simulation study conducted to evaluate the finite-sample performance of the variable selection procedure proposed in the previous sections. In the study, the covariates were generated from the multivariate normal distribution with mean zero and the covariance matrix $\Sigma_X$ with the $(j,k)$th element being $0.5^{|j-k|}$. To generate the true failure times $T_i$'s, we assumed that they follow model (2.1) with $\Lambda_0(t) = (2t)^{1.25}$. For the covariate effects, we considered the following three scenarios with $s$ denoting the number of significant covariates or the covariates with nonzero effects.

- Scenario (a) with $s = 4$: $\psi_1(X_{1i}) = -6X_{1i} + 3$, $\psi_2(X_{2i}) = -5.4X_{2i} + 2.7$, $\psi_3(X_{3i}) = 4.6X_{3i} - 2.3$, $\psi_4(X_{4i}) = 3.5X_{4i} - 1.75$, and $\psi_j(X_{ji}) = 0$ for $j = 5, ..., p$.

- Scenario (b) with $s = 4$: $\psi_1(X_{1i}) = 2sin(\pi(X_{1i} - 0.5))$, $\psi_2(X_{2i}) = -6X_{2i} + 3$, $\psi_3(X_{3i}) = -1.8sin(1.5\pi X_{3i} + 0.5)$, $\psi_4(X_{4i}) = 2cos(\pi X_{4i})$, and $\psi_j(X_{ji}) = 0$ for $j = 5, ..., p$.

- Scenario (c) with $s = 6$: $\psi_1(X_{1i}) = 2sin(\pi(X_{1i} - 0.5))$, $\psi_2(X_{2i}) = -6X_{2i} + 3$, $\psi_3(X_{3i}) = -1.8sin(1.5\pi X_{3i} + 0.5)$, $\psi_4(X_{4i}) = 2cos(\pi X_{4i})$, $\psi_5(X_{5i}) = -5.4X_{5i} + 2.7$, $\psi_6(X_{6i}) = 4.6X_{6i} - 2.3$, and $\psi_j(X_{ji}) = 0$ for $j = 7, \ldots, p$.

Note that in scenario (a), we have four important covariates with all having linear effects, and in scenario (b), we also have four important covariates but three have nonlinear effects. In scenario (c), we have six important covariates and four of them are the same as in scenario (b) with the other two having linear effects.

To generate interval-censored observations on the $T_i$'s, it was assumed that all subjects could be potentially observed at ten equally spaced time points between 0 and $\tau$ with $\tau$ selected to give the desired percentage of right-censored observations. For each subject and at each of these time points, the observation occurs with probability 0.5 independent of other observations. Then for subject $i$, $L_i$ was defined as the largest observation time point less than $T_i$ and $R_i$ as the smallest observation time point greater than $T_i$. For all Bernstein polynomials approximations, the degree was set to be $m_0 = m_j = 3$. The results given below are based on $n = 300$ or 500 and $p = 100, 500$ or 1000 with 100 replications.

Tables 2.1 - 2.3 present the results given by the proposed variable selection approach under the scenario (a) - (c), respectively, with the use of the group LASSO, MCP and SCAD penalty functions and the percentage of right-censored observations being around 30%. The results include the number of times that the $j^{th}$ additive component was selected ($\text{TP}_j$), the average number of correctly selected components whose true coefficients are nonzero (TP), and the average number of incorrectly selected components whose true coefficients are zero (FP), $j = 1, ..., s$. In addition, we also calculated and included in the tables the median of the mean squared errors

(MMSE) given by MSE $= s^{-1} \sum_{j=1}^{s} \int (\psi_j(x) - \hat{\psi}_j(x))^2 dx$ along with their standard deviation (SD) in 100 replications. In the analysis here, we set $a = 3.7$ for the group SCAD penalty function by following Fan and Li (2001) and Huang *et al.* (2012) and $\gamma = 1.1$ for the group MCP by following Zhang (2010) and Huang *et al.* (2012). These authors have shown that the results are relatively robust with respect to the values of these parameters.

One can see from the tables that the proposed variable selection procedure with all of three penalty functions seems to perform well for the situations considered. In particular, the results suggest that all linear and nonlinear effects could be correctly detected with high percentages. On the other hand, the results indicate that compared to the group MCP and group SCAD penalty functions, the method with the group LASSO penalty function tends to identify more irrelevant variables and yield relatively larger models, which is especially the case in the high-dimensional situations. One possible reason for this is that the group selection consistency of the group LASSO relies on the assumption that the design matrix satisfies the irrepresentable condition, which is difficult to satisfy in $p \gg n$ scenarios as pointed out in the literature (Zhang, 2010; Huang *et al.*, 2012).

The results above also indicate that as expected, the FP increased when $p$ increased and decreased when $n$ increased, while the TP and MMSE were relatively stable when $p$ and $n$ changed. To see the possible effect of the right-censoring percentage on the proposed approach, we repeated the study giving Table 3 except setting the percentage of right-censored observations being around 45% and give the results in Table 4. In addition, we also repeated the study giving Table 3 except setting $m_0 = m_j = 5$ to see the possible effects of the values of the $m_j$'s on the proposed

method and present the results in Table 5. It is apparent that all results are similar and they suggest that the proposed approach is relatively robust with respect to both the censoring percentage and the values of the $m_j$'s.

## 2.5 Analysis of an Alzheimer's Disease Study

Now we apply the variable selection approach proposed in the previous sections to the data arising from the Alzheimer's Disease Neuroimaging Initiative (ADNI), an ongoing longitudinal multi-site study. Among others, one of the main objectives of the study is to identify significant clinical, imaging, genetic or biochemical biomarkers related to the progression or status of Alzheimer's Disease (AD) such as the AD conversion time from the initial visit. The patients in the study are classified into three groups based on their cognitive status: cognitively normal (CN), mild cognitive impairment (MCI), and AD. By following others, we will focus on the participants in the MCI group (Li *et al.*, 2017, 2018; Hu *et al.*, 2018; Wu *et al.*, 2020; Wang *et al.*, 2012). Due to the periodic follow-up nature, only interval-censored data are available on the AD conversion time.

For the analysis, following Li *et al.* (2017) and Wu *et al.* (2020), we will consider the 280 individuals who have complete information on the 327,354 single nucleotide polymorphisms (SNPs) and four demographic and clinical factors in the MCI group with the focus on identifying significant SNPs with either linear or nonlinear effects on the risk of AD conversion given four demographic and clinical factors. The four factors are participants' Alzheimer's Disease Assessment Scale Score of 13 items (ADAS13), Rey auditory verbal learning test score of immediate recall (RAVLT.i), functional

assessment questionnaire score (FAQ), and MRI volume of middle temporal gyrus (MidTemp). The SNPs were read by the Illumina method and then translated into continuous covariates with values 0, 1, and 2, indicating a SNP without allele T, heterozygous with one allele T, and homozygous with two alleles T, respectively. The additive Cox model can be used to investigate the possibly nonlinear effects of the SNPs on the risk of developing AD.

Before the application of the proposed approach, as Wu *et al.* (2020), we first used the mid-point imputation method to convert the interval-censored data to right-censored data and then applied the sure independent screening procedure given in Fan *et al.* (2010) to identify the top 500 SNPs. Table 6 presents the SNPs selected among these 500 SNPs by the proposed method with a degree of 3 for all Bernstein polynomials approximations and the use of the three group penalty functions considered in the simulation study. In total, 40 SNPs were selected by the three group penalty functions with 31, 13, or 14 SNPs selected by the group LASSO, group MCP or group SCAD penalty function, respectively. In particular, five SNPs, rs1455888, rs1467025, rs1619465, rs170519 and rs1953851, were selected by all three penalty functions. Note that as seen and discussed in the previous section, the group LASSO is much more conservative and selected more SNPs in high-dimensional scenarios. Among the 40 selected SNPs, 16 were also selected by Wu *et al.* (2020), and only rs1467025 among the five commonly selected SNPs was selected by Wu *et al.* (2020), which assumed that all SNPs only had linear effects.

To further investigate the selected SNPs, Figure 1 shows the estimated effects of the SNPs selected by all three group penalty functions. Figures 2 - 4 display the estimated effects of the selected SNPs by each of the three group penalty functions,

respectively. Note that all of the covariates were re-scaled from 0 to 1 and the covariate effects were calculated only at the three points. One can see that many of the selected SNPs seem to have nonlinear covariates effects and the estimated effects are similar or consistent with respect to the three penalty function. For instance, the estimated curve of rs1455888 implies that the risk of AD conversion is higher among those with heterozygous rs1455888 with one allele T than those with homozygous rs1455888 with two alleles T and those without allele T. In other words, the risk of AD conversion does not increase linearly with increase in the number of allele T in SNP rs1455888.

## 2.6 Discussion and Concluding Remarks

This chapter discussed the variable selection for the nonparametric additive Cox model based on interval-censored failure time data, and for the problem, a penalized, sieve variable selection procedure with the use of Bernstein polynomials approximations was developed. The proposed method can be seen as a generalization of that given by Wu *et al.* (2020), which only considered the selection of the covariates with linear effects. A simulation study was performed to assess the proposed method and indicated that it works well for practical situations. Furthermore, the approach was applied to the ADNI genetic data, which motivated this study, and identified some significant SNPs that had nonlinear effects on the risk of AD conversion.

Note that in the proposed variable selection procedure, Bernstein polynomials were employed to approximate unknown functions. Similar methods could be developed with the use of other smooth function approximation such as splines. In the preceding sections, we only considered the group LASSO, MCP and SCAD penalty functions

and it is apparent that the procedure method would still be valid if some other group penalty functions were used. However, the implementation could be different under different group penalty functions. Also in the preceding sections, we have focused on the nonparametric additive Cox model and it is straightforward to apply the idea discussed above to other models.

The proposed method has some limitations or there exist several directions for future research related to the proposed approach. One is that the focus here has been on the low-dimensional ($p < n$) and high-dimensional ($p > n$) situations or the proposed method cannot be applied to ultra-high dimensional ($p \gg n$) cases. In other words, for the ultra-high dimensional situation, a new sure independent screening procedure similar to that given in Fan *et al.* (2010) for right-censored data needs to be developed. Another assumption imposed above is that the censoring mechanism is independent or the mechanism behind censoring intervals is independent of the failure time of interest. It is clear that this could be violated sometimes (Sun, 2006) and thus it would be useful to generalize the method proposed above to the dependent interval censoring situation. Of course, it would be desirable or helpful to provide some theoretical justification to or establish the asymptotic properties of the proposed method, which will not be straightforward.

Table 2.1: Simulation results under scenario (a) with the right censoring percentage being around 30%

| $n$ | $p$ | **Penalty** | **TP1** | **TP2** | **TP3** | **TP4** | **TP** | **FP** | **MMSE(SD)** |
|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | GLASSO | 100 | 100 | 100 | 100 | 4 | 1.63 | 0.543 (0.232) |
| | | GMCP | 100 | 100 | 100 | 99 | 3.99 | 0.5 | 0.146 (0.200) |
| | | GSCAD | 100 | 100 | 100 | 99 | 3.99 | 1.37 | 0.156 (0.246) |
| | 500 | GLASSO | 100 | 99 | 98 | 100 | 3.97 | 1.9 | 0.720 (0.258) |
| | | GMCP | 100 | 99 | 96 | 98 | 3.93 | 0.94 | 0.110 (0.279) |
| | | GSCAD | 100 | 98 | 96 | 97 | 3.91 | 1.17 | 0.132 (0.322) |
| | 1000 | GLASSO | 100 | 99 | 98 | 100 | 3.97 | 3.05 | 0.785 (0.258) |
| | | GMCP | 100 | 98 | 92 | 99 | 3.89 | 1.15 | 0.126 (0.300) |
| | | GSCAD | 100 | 98 | 94 | 99 | 3.91 | 1.94 | 0.130 (0.334) |
| 500 | 100 | GLASSO | 100 | 100 | 100 | 100 | 4 | 0.71 | 0.420 (0.167) |
| | | GMCP | 100 | 99 | 99 | 99 | 3.97 | 0.08 | 0.083 (0.182) |
| | | GSCAD | 100 | 100 | 100 | 99 | 3.99 | 0.11 | 0.088 (0.130) |
| | 500 | GLASSO | 100 | 100 | 98 | 100 | 3.98 | 0.43 | 0.578 (0.205) |
| | | GMCP | 100 | 100 | 98 | 100 | 3.98 | 0.24 | 0.084 (0.143) |
| | | GSCAD | 100 | 100 | 98 | 99 | 3.97 | 0.07 | 0.079 (0.153) |
| | 1000 | GLASSO | 100 | 100 | 100 | 100 | 4 | 0.44 | 0.626 (0.202) |
| | | GMCP | 100 | 100 | 100 | 99 | 3.99 | 0.39 | 0.096 (0.138) |
| | | GSCAD | 100 | 100 | 99 | 98 | 3.97 | 0.11 | 0.080 (0.142) |

Table 2.2: Simulation results under scenario (b) with the right censoring percentage being around 30%

| $n$ | $p$ | **Penalty** | **TP1** | **TP2** | **TP3** | **TP4** | **TP** | **FP** | **MMSE(SD)** |
|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | GLASSO | 100 | 99 | 99 | 100 | 3.98 | 3.64 | 0.808 (0.254) |
| | | GMCP | 100 | 100 | 100 | 100 | 4.00 | 0.92 | 0.174 (0.327) |
| | | GSCAD | 100 | 100 | 100 | 100 | 4.00 | 1.35 | 0.177 (0.371) |
| | 500 | GLASSO | 100 | 97 | 98 | 100 | 3.95 | 9.68 | 1.038 (0.324) |
| | | GMCP | 100 | 100 | 100 | 100 | 4.00 | 1.97 | 0.156 (0.226) |
| | | GSCAD | 100 | 100 | 100 | 100 | 4.00 | 5.24 | 0.216 (0.429) |
| | 1000 | GLASSO | 99 | 96 | 98 | 100 | 3.93 | 17.06 | 1.113 (0.348) |
| | | GMCP | 99 | 99 | 100 | 99 | 3.97 | 2.89 | 0.177 (0.296) |
| | | GSCAD | 99 | 98 | 99 | 100 | 3.96 | 9.02 | 0.311 (0.788) |
| 500 | 100 | GLASSO | 100 | 100 | 99 | 100 | 3.99 | 0.58 | 0.827 (0.232) |
| | | GMCP | 100 | 100 | 100 | 100 | 4 | 0.21 | 0.138 (0.210) |
| | | GSCAD | 100 | 100 | 100 | 100 | 4 | 0.24 | 0.139 (0.213) |
| | 500 | GLASSO | 100 | 100 | 98 | 100 | 3.98 | 2.26 | 0.882 (0.248) |
| | | GMCP | 100 | 100 | 100 | 100 | 4 | 0.4 | 0.139 (0.168) |
| | | GSCAD | 100 | 100 | 100 | 100 | 4 | 0.33 | 0.138 (0.168) |
| | 1000 | GLASSO | 100 | 98 | 98 | 100 | 3.96 | 4.01 | 0.940 (0.282) |
| | | GMCP | 100 | 100 | 100 | 100 | 4 | 0.46 | 0.138 (0.199) |
| | | GSCAD | 100 | 100 | 100 | 100 | 4 | 0.69 | 0.134 (0.200) |

Table 2.3: Simulation results under scenario (c) with the right censoring percentage being around 30%

| $n$ | $p$ | **Penalty** | **TP1** | **TP2** | **TP3** | **TP4** | **TP5** | **TP6** | **TP** | **FP** | **MMSE(SD)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | GLASSO | 98 | 99 | 92 | 100 | 100 | 97 | 5.86 | 2.65 | 1.006 (0.317) |
| | | GMCP | 100 | 99 | 96 | 100 | 100 | 99 | 5.94 | 0.79 | 0.166 (0.269) |
| | | GSCAD | 100 | 99 | 99 | 100 | 100 | 100 | 5.98 | 1.43 | 0.198 (0.316) |
| | 500 | GLASSO | 100 | 92 | 92 | 100 | 99 | 99 | 5.82 | 11.10 | 1.119 (0.325) |
| | | GMCP | 100 | 100 | 98 | 100 | 100 | 100 | 5.98 | 2.44 | 0.209 (0.349) |
| | | GSCAD | 100 | 99 | 98 | 99 | 100 | 100 | 5.96 | 1.98 | 0.185 (0.310) |
| | 1000 | GLASSO | 98 | 93 | 85 | 100 | 99 | 96 | 5.71 | 20.36 | 1.210 (0.351) |
| | | GMCP | 98 | 95 | 95 | 99 | 100 | 96 | 5.83 | 4.78 | 0.231 (0.555) |
| | | GSCAD | 100 | 100 | 99 | 99 | 100 | 98 | 5.96 | 3.85 | 0.190 (0.487) |
| 500 | 100 | GLASSO | 100 | 99 | 96 | 100 | 100 | 99 | 5.94 | 0.62 | 0.992 (0.250) |
| | | GMCP | 100 | 100 | 99 | 100 | 100 | 100 | 5.99 | 0.14 | 0.133 (0.199) |
| | | GSCAD | 100 | 100 | 99 | 100 | 100 | 100 | 5.99 | 0.02 | 0.133 (0.201) |
| | 500 | GLASSO | 100 | 94 | 97 | 100 | 100 | 99 | 5.90 | 1.96 | 1.069 (0.275) |
| | | GMCP | 100 | 99 | 99 | 100 | 99 | 100 | 5.97 | 0.28 | 0.128 (0.226) |
| | | GSCAD | 100 | 99 | 99 | 100 | 99 | 100 | 5.97 | 0.16 | 0.129 (0.227) |
| | 1000 | GLASSO | 99 | 94 | 94 | 100 | 100 | 98 | 5.85 | 3.47 | 1.154 (0.294) |
| | | GMCP | 100 | 100 | 100 | 100 | 100 | 100 | 6.00 | 0.50 | 0.143 (0.183) |
| | | GSCAD | 100 | 100 | 100 | 100 | 100 | 100 | 6.00 | 0.41 | 0.140 (0.175) |

Table 2.4: Simulation results under scenario (c) with the right censoring percentage being around 45%

| $n$ | $p$ | **Penalty** | **TP1** | **TP2** | **TP3** | **TP4** | **TP5** | **TP6** | **TP** | **FP** | **MMSE(SD)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | GLASSO | 98 | 92 | 80 | 100 | 100 | 94 | 5.64 | 2.19 | 1.111 (0.363) |
| | | GMCP | 99 | 99 | 98 | 100 | 100 | 99 | 5.95 | 1.50 | 0.202 (0.424) |
| | | GSCAD | 100 | 99 | 98 | 100 | 100 | 100 | 5.97 | 1.11 | 0.198 (0.358) |
| | 500 | GLASSO | 99 | 84 | 81 | 100 | 97 | 94 | 5.55 | 7.79 | 1.249 (0.379) |
| | | GMCP | 100 | 99 | 99 | 100 | 98 | 99 | 5.95 | 5.53 | 0.238 (0.774) |
| | | GSCAD | 100 | 99 | 98 | 99 | 98 | 99 | 5.93 | 4.25 | 0.216 (0.617) |
| | 1000 | GLASSO | 94 | 84 | 76 | 99 | 99 | 89 | 5.41 | 16.14 | 1.287 (0.379) |
| | | GMCP | 97 | 90 | 88 | 99 | 100 | 91 | 5.65 | 9.09 | 0.317 (1.098) |
| | | GSCAD | 97 | 93 | 93 | 100 | 100 | 94 | 5.77 | 7.37 | 0.245 (0.971) |
| 500 | 100 | GLASSO | 99 | 94 | 90 | 100 | 99 | 95 | 5.77 | 0.46 | 1.117 (0.318) |
| | | GMCP | 100 | 100 | 99 | 100 | 100 | 100 | 5.99 | 0.39 | 0.134 (0.209) |
| | | GSCAD | 100 | 100 | 99 | 100 | 100 | 100 | 5.99 | 0.11 | 0.136 (0.208) |
| | 500 | GLASSO | 99 | 87 | 85 | 100 | 100 | 91 | 5.62 | 1.32 | 1.162 (0.334) |
| | | GMCP | 100 | 99 | 100 | 100 | 100 | 100 | 5.99 | 0.71 | 0.138 (0.204) |
| | | GSCAD | 100 | 100 | 100 | 100 | 100 | 100 | 6.00 | 0.38 | 0.141 (0.179) |
| | 1000 | GLASSO | 95 | 85 | 69 | 100 | 99 | 84 | 5.32 | 2.05 | 1.332 (0.361) |
| | | GMCP | 100 | 99 | 100 | 100 | 100 | 99 | 5.98 | 0.97 | 0.157 (0.252) |
| | | GSCAD | 100 | 100 | 100 | 100 | 100 | 99 | 5.99 | 0.58 | 0.146 (0.218) |

Table 2.5: Simulation results under scenario (c) with the right censoring percentage being around 30% and $m_j = 5$

| $n$ | $p$ | **Penalty** | **TP1** | **TP2** | **TP3** | **TP4** | **TP5** | **TP6** | **TP** | **FP** | **MMSE(SD)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | GLASSO | 98 | 98 | 93 | 100 | 99 | 95 | 5.83 | 2.53 | 1.139 (0.310) |
| | | GMCP | 100 | 99 | 96 | 100 | 100 | 99 | 5.94 | 2.94 | 0.278 (0.417) |
| | | GSCAD | 100 | 100 | 98 | 100 | 98 | 97 | 5.93 | 0.79 | 0.247 (0.378) |
| | 500 | GLASSO | 100 | 91 | 88 | 100 | 99 | 97 | 5.75 | 8.12 | 1.302 (0.304) |
| | | GMCP | 100 | 100 | 98 | 100 | 98 | 100 | 5.96 | 7.22 | 0.385 (0.699) |
| | | GSCAD | 100 | 99 | 97 | 100 | 98 | 97 | 5.91 | 4 | 0.349 (0.632) |
| | 1000 | GLASSO | 98 | 91 | 91 | 100 | 99 | 96 | 5.75 | 17.81 | 1.332 (0.326) |
| | | GMCP | 99 | 96 | 97 | 99 | 100 | 97 | 5.88 | 11.39 | 0.465 (1.183) |
| | | GSCAD | 100 | 98 | 97 | 100 | 100 | 97 | 5.92 | 8.02 | 0.507 (0.766) |
| 500 | 100 | GLASSO | 100 | 99 | 98 | 100 | 100 | 97 | 5.94 | 0.54 | 1.146 (0.239) |
| | | GMCP | 100 | 100 | 99 | 100 | 100 | 100 | 5.99 | 1.75 | 0.171 (0.222) |
| | | GSCAD | 100 | 99 | 97 | 100 | 100 | 100 | 5.96 | 0.02 | 0.172 (0.254) |
| | 500 | GLASSO | 100 | 94 | 97 | 100 | 100 | 97 | 5.88 | 1.62 | 1.206 (0.258) |
| | | GMCP | 100 | 99 | 99 | 100 | 99 | 100 | 5.97 | 1.98 | 0.174 (0.242) |
| | | GSCAD | 100 | 100 | 100 | 100 | 99 | 100 | 5.99 | 0.14 | 0.172 (0.202) |
| | 1000 | GLASSO | 99 | 94 | 95 | 100 | 100 | 97 | 5.85 | 2.52 | 1.296 (0.271) |
| | | GMCP | 100 | 99 | 99 | 100 | 100 | 100 | 5.98 | 2.21 | 0.177 (0.241) |
| | | GSCAD | 100 | 100 | 100 | 100 | 100 | 99 | 5.99 | 0.44 | 0.164 (0.204) |

Table 2.6: List of the selected SNPs for the ADNI data

| SNP Name | GLASSO | GMCP | GSCAD |
|---|---|---|---|
| rs10089267 | + | | |
| rs10178349 | + | | |
| rs10512849 | + | | |
| rs10771958 | + | | |
| rs1158464 | + | | |
| rs1161397 | + | | |
| rs11704226 | + | | + |
| rs12042017 | + | | |
| rs12334868 | + | | + |
| rs12454238 | + | | + |
| rs12555515 | + | | |
| rs12589973 | + | | |
| rs13037957 | + | | |
| rs1330312 | + | | |
| rs1336703 | + | | |
| rs138957 | + | | |
| rs1397228 | + | | |
| rs1399417 | + | + | |
| rs1455888 | + | + | + |
| rs1467025 | + | + | + |
| rs1475950 | + | | |
| rs1499966 | + | | |
| rs1619465 | + | + | + |
| rs1638438 | + | | |
| rs170519 | + | + | + |
| rs1953851 | + | + | + |
| rs2050635 | + | | |
| rs2255417 | + | | |
| rs2295894 | + | | |
| rs2428754 | + | | |
| rs2514897 | + | | |
| rs1063666 | | + | + |
| rs10799802 | | + | + |
| rs11128570 | | + | + |
| rs12694110 | | + | + |
| rs10879354 | | + | |
| rs17170619 | | + | |
| rs1868750 | | + | |
| rs1470002 | | | + |
| rs2444907 | | | + |

Figure 2.1: Estimated effects of the five SNPs selected by all of the three penalty functions: row 1 - GLASSO, row 2 - GMCP, row 3 - GSCAD

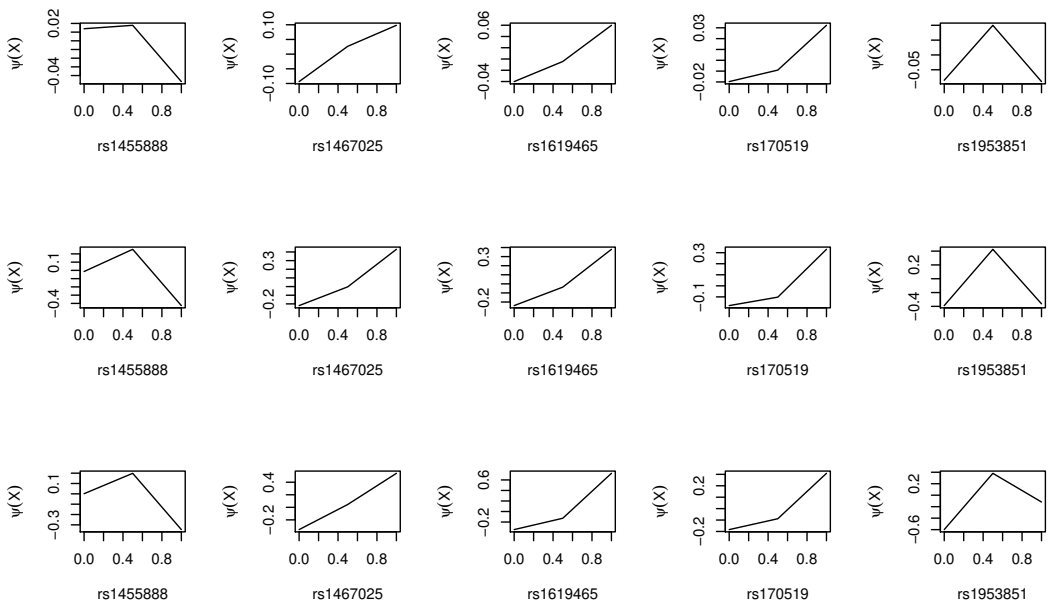


Figure 2.2: Estimated effects of the 31 SNPs selected by the GLASSO penalty

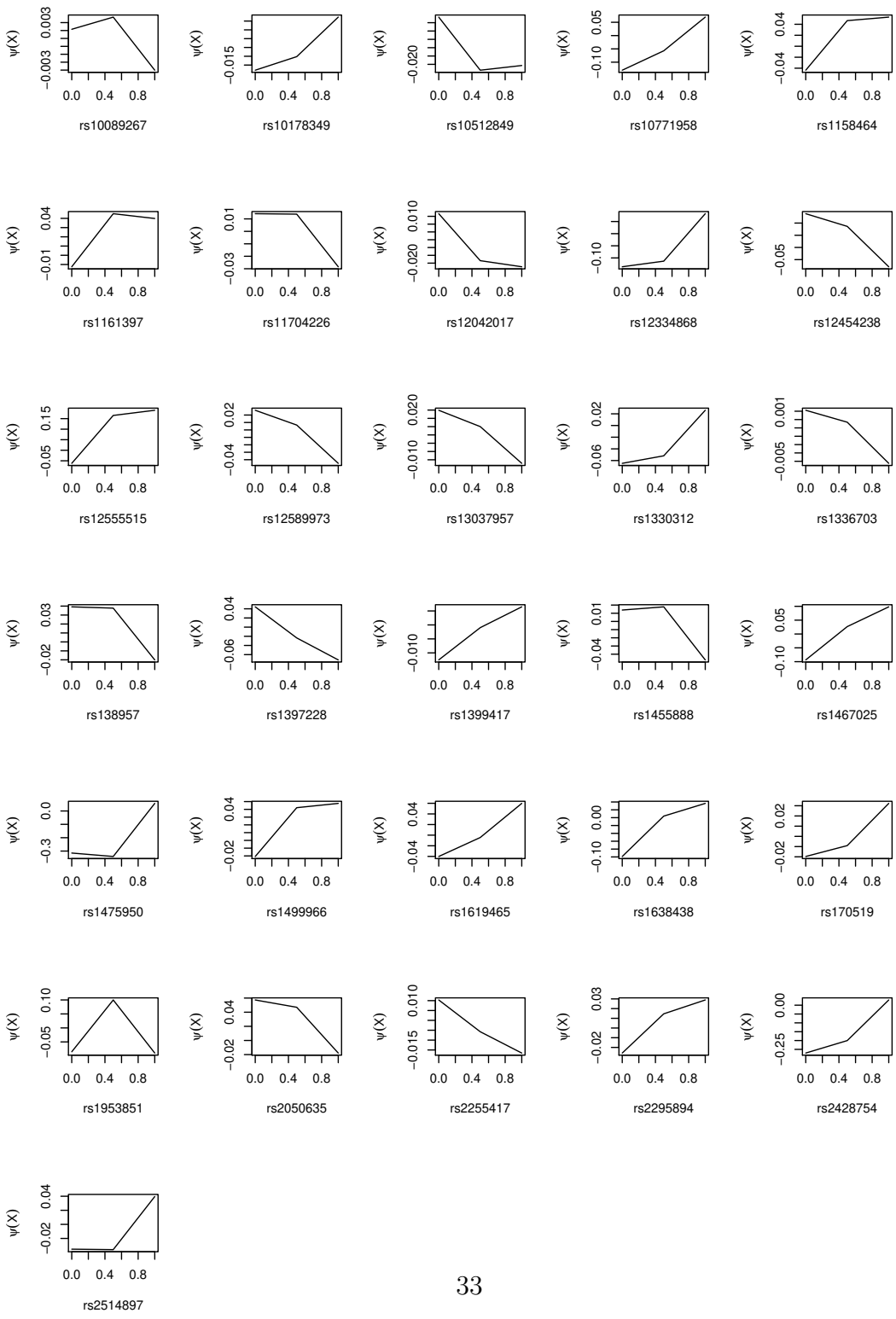


Figure 2.3: Estimated effects of the 13 SNPs selected by the GMCP penalty

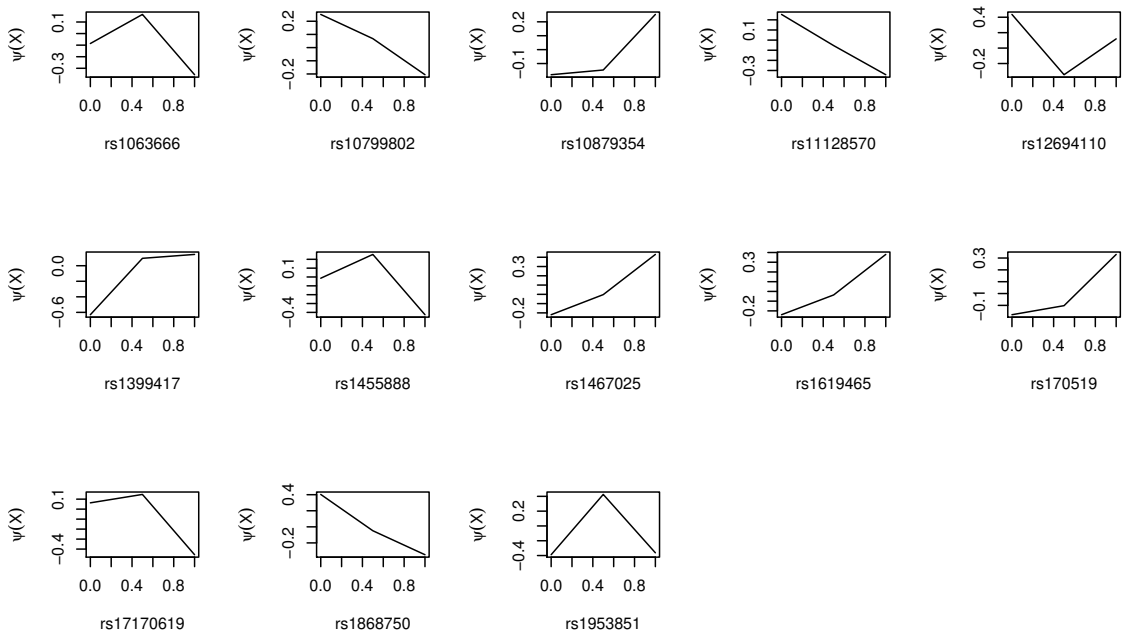
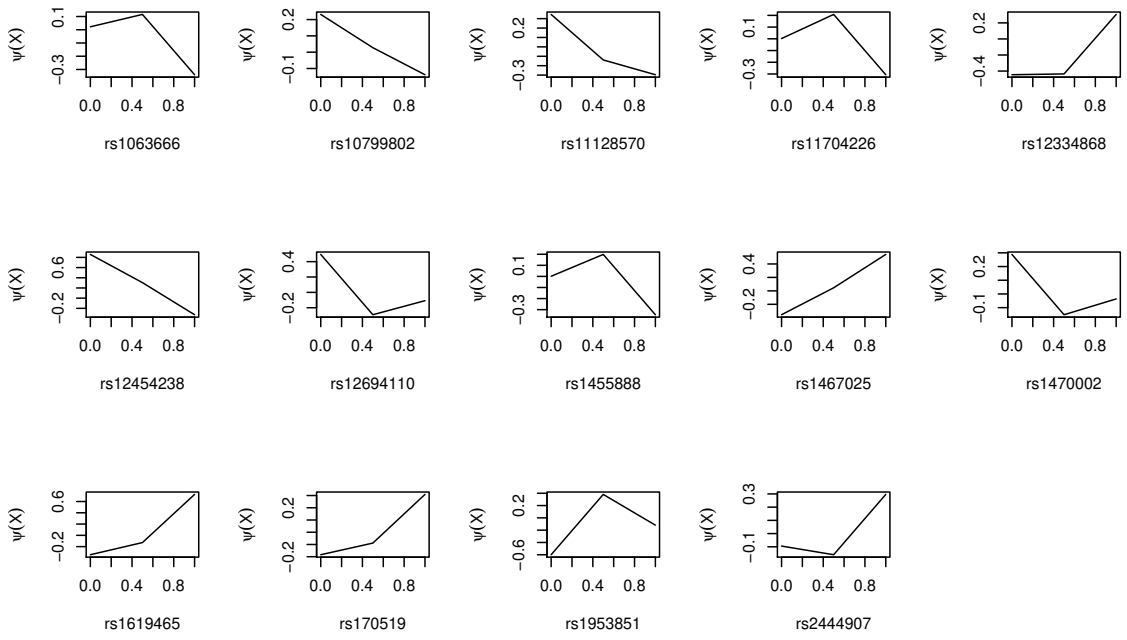


Figure 2.4: Estimated effects of the 14 SNPs selected by the GSCAD penalty



# Chapter 3

# Variable Selection for Nonlinear Covariate Effects with Interval-censored Failure Time Data

## 3.1 Introduction

This chapter discusses variable selection when one faces general, high-dimensional interval-censored failure time data, which commonly occur in many areas such as epidemiological, medical and public health studies (Sun, 2006). By interval-censored data, we usually mean that the failure time of interest is known or observed only to belong to an interval instead of being observed exactly. It is apparent that right-censored data, the type of failure time data discussed most in the literature, can be seen as a special case of interval-censored data. In addition to high dimensionality and interval censoring, it can also often be the case that covariates may have non-

linear effects on the failure time of interest, which makes the variable selection more challenging.

Several methods have been proposed for variable selection based on interval-censored failure time data (Wu and Cook, 2015; Zhao *et al.*, 2020; Wu *et al.*, 2020; Li *et al.*, 2020). However, most of the existing methods have assumed that or only considered the situation where covariate effects are linear, and it is well-known that sometimes, this may not be true in practice. In other words, some covariates may have nonlinear effects on the failure time of interest (Cai *et al.*, 2007; Cheng and Wang, 2011; Huang, 1999; Lu and McMahan, 2018). To address this, in the following, we will propose a penalized sieve maximum likelihood variable selection procedure under a general class of semiparametric additive transformation models.

The semiparametric additive transformation model provides a flexible and comprehensive way to model the failure time variable and a key feature is that it allows for arbitrary forms or patterns of the effects of covariates on the failure time of interest. In the proposed method, we will adopt the sieve approach with the use of Bernstein polynomials to approximate unknown functions involved, which can significantly reduce the computational burden among other advantages. Also in the proposed procedure, group penalty functions will be used. In particular, we will focus on the group least absolute shrinkage and selection operator (GLASSO) penalty function (Yuan and Lin, 2006), the group smoothly clipped absolute deviation (GSCAD) penalty function (Fan and Li, 2001), and the group minimum concave penalty (GMCP) penalty function (Zhang, 2010) although other group penalty functions could be applied too.

The remainder of the chapter is organized as follows. In Section 3.2, we will first introduce some notation and the semiparametric additive transformation model and

then describe the other needed background. Section 3.3 will present the proposed, penalized sieve maximum likelihood variable selection procedure, and in particular, a novel EM algorithm will be developed for the implementation of the proposed approach. In Section 3.4, some results from a simulation study conducted to assess the finite-sample performance of the proposed method are provided and indicate that it works well for the situations considered. We will apply the procedure to a set of real data arising from an Alzheimer's Disease study in Section 3.5 and Section 3.6 concludes with some discussion and remarks.

## 3.2 Notation, Models and Preparations

Consider a failure time study with $n$ independent individuals and let $T_i$ and $\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})'$ denote the failure time of interest and a $p$-dimensional vector of covariates associated with the $i$th subject, respectively. For $T_i$, suppose that only an interval $(L_i, R_i]$ such that $L_I < T_i \leq R_i$ is observed, $i = 1, ..., n$. That is, only interval-censored data are observed and the observed data have the form $\{(\boldsymbol{X}_i, L_i < T_i \leq R_i), i = 1, ..., n\}$.

For the covariate effect, we will assume that given $\boldsymbol{X}_i$, $T_i$ follows the semiparametric additive transformation model with the cumulative hazards function having the form

$$\Lambda(t|\boldsymbol{X}_i) = G\left[\Lambda_0(t)\, \exp\left\{\boldsymbol{f}(\boldsymbol{X})\right\}\right] = G\left[\Lambda_0(t)\, \exp\left\{\sum_{j=1}^{p} f_j(X_j)\right\}\right]. \tag{3.1}$$

In the above, $G(\cdot)$ denotes a pre-specified increasing transformation function, $\Lambda_0(t)$ is an unknown baseline cumulative hazard function, and $f_1, ..., f_p$ are unknown func-

tions. It is easy to see that the class of models above is quite flexible and includes many commonly used models as special cases. For instance, the choices of $G(x) = x$ and $G(x) = log(1+x)$ give the additive Cox model and the additive proportional odds model, respectively. For the identification of the model above, we will assume that $E[f_j(\cdot)] = 0$ for all $1 \leq j \leq p$ and the censoring is independent (Huang, 1999; Sun, 2006). Also we will adopt the sparsity assumption although $p$ can be small ($p < n$) or large ($p > n$) (Lv *et al.*, 2018).

Under the assumptions above, one can easily derive the observed likelihood function as

$$L(\Lambda_0, \boldsymbol{f}) = \prod_{i=1}^{n} \left\{ \exp\left[ -G\left\{ \Lambda_0(L_i) e^{\sum_{j=1}^{p} f_j(X_{ij})} \right\} \right] - \exp\left[ G\left\{ -\Lambda_0(R_i) e^{\sum_{j=1}^{p} f_j(X_{ij})} \right\} \right] \right\}, \tag{3.2}$$

where $\boldsymbol{f} = (f_1, ..., f_p)$. It is apparent that the direct use of the likelihood function above would not be easy because of the involvement of unknown functions $H_0$ and $\boldsymbol{f}$. To deal with this, following others (Zhao *et al.*, 2020; Li *et al.*, 2020), we suggest to employ the sieve approach first to approximate $\boldsymbol{f}$ using Bernstein polynomials. Specifically, define the parameter space

$$\Theta = \{(\Lambda_0, f_1, ..., f_p) \in (\mathcal{H}_0 \otimes \mathcal{M}_1 \otimes ... \otimes \mathcal{M}_p)\}$$

and the sieve parameter space

$$\Theta_n = \{(\Lambda_0, f_{1n}, ..., f_{pn}) \in (\mathcal{H}_0 \otimes \mathcal{M}_{1n} \otimes ... \otimes \mathcal{M}_{pn})\} .$$

In the above,

$$\mathcal{M}_{jn} = \left\{ f_{jn}(X_j) = \sum_{k=0}^{m_j} \alpha_{jk} B_{jk}(X_j, m_j, c_j, u_j) : \sum_{0 \le k \le m_j} |\alpha_{jk}| \le M_{jn} \right\}, \; j = 1, ..., p \,,$$

with

$$B_{jk}(X_j, m_j, c_j, u_j) = \binom{m_j}{k} \left( \frac{X_j - c_j}{u_j - c_j} \right)^k \left( 1 - \frac{X_j - c_j}{u_j - c_j} \right)^{m_j - k}, k = 0, 1, ..., m_j \,,$$

which are the Bernstein polynomials with the degree of freedoms $m_j = o(n^a)$ for some $a \in (0, 1)$.

Define $\boldsymbol{\alpha} = (\alpha_{10}, ..., \alpha_{1m_1}, ..., \alpha_{p0}, ..., \alpha_{pm_p})'$. Then the likelihood function in (3.2) can be rewritten as

$$L(\Lambda_0, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \left\{ \exp\left[ -G\left\{ \Lambda_0(L_i) e^{\sum_{j=1}^{p} f_{jn}(X_{ij})} \right\} \right] - \exp\left[ G\left\{ -\Lambda_0(R_i) e^{\sum_{j=1}^{p} f_{jn}(X_{ij})} \right\} \right] \right\}. \tag{3.3}$$

Note that instead of Bernstein polynomials, one could use other smoothing functions such as B-splines. The advantages of Bernstein polynomials include their simplicity and no requirement of specifying interior knots (Wang and Ghosh, 2012; Zhou *et al.*, 2017; Carnicer and Pena, 1993).

## 3.3 Penalized Sieve Maximum Likelihood Variable Selection

### 3.3.1 Penalized Procedure

Now we discuss the variable or covariate selection or how to identify relevant and important covariates. For this, motivated by the literature, we propose to minimize the penalized negative log likelihood function

$$\ell_p(\boldsymbol{\alpha}, \Lambda_0) = -\frac{1}{n} logL(\Lambda_0, \boldsymbol{\alpha}) + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j) = \ell_1(\Lambda_0, \boldsymbol{\alpha}) + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j)\,, \tag{3.4}$$

where $P_\lambda(\boldsymbol{\alpha}_j)$ denotes a group penalty function with respect to $\boldsymbol{\alpha}_j = (\alpha_{j0}, ..., \alpha_{jm_p})$ and $\lambda$ is a tuning parameter. It is apparent that $\boldsymbol{\alpha}_j = \mathbf{0}$ will mean that the covariate $X_j$ has zero or non-significant effect on the failure time of interest.

For the penalty function $P_\lambda(\boldsymbol{\alpha}_j)$, although one can use any, we will focus on GLASSO, GMCP, and GSCAD penalty functions in the following. The GLASSO penalty function, proposed by Yuan and Lin (2006), is given by $P_\lambda(\boldsymbol{\alpha}_j) = \lambda_j \|\boldsymbol{\alpha}_j\|_2$ with $\lambda_j = \lambda\sqrt{m_j + 1}$ and $\|\boldsymbol{\alpha}_j\|_2 = \sqrt{\sum_{l=0}^{m_j} \alpha_{jl}^2}$, while the GMCP penalty function has the form

$$P_\lambda(\boldsymbol{\alpha}_j; \eta) = \begin{cases} \lambda_j \|\boldsymbol{\alpha}_j\|_2 - \dfrac{\|\boldsymbol{\alpha}_j\|_2^2}{2\eta} & \text{if } \|\boldsymbol{\alpha}_j\|_2 \leq \lambda_j \eta\,, \\ \dfrac{\lambda_j^2 \eta}{2} & \text{if } \|\boldsymbol{\alpha}_j\|_2 > \lambda_j \eta \end{cases}$$

with $\lambda_j = \lambda\sqrt{m_j + 1}$ and $\eta > 1$ (Zhang, 2010). Finally the GSCAD penalty function

takes the form

$$P_\lambda(\boldsymbol{\alpha}_j;\omega) = \begin{cases} \lambda_j \|\boldsymbol{\alpha}_j\|_2 & \text{if } \|\boldsymbol{\alpha}_j\|_2 \le \lambda_j\,, \\ \dfrac{2\omega\lambda_j\|\boldsymbol{\alpha}_j\|_2 - \|\boldsymbol{\alpha}_j\|_2^2 - \lambda_j^2}{2(\omega-1)} & \text{if } \lambda_j < \|\boldsymbol{\alpha}_j\|_2 \le \omega\lambda_j\,, \\ \dfrac{(\omega^2-1)\lambda_j^2}{2(\omega-1)} & \text{if } \|\boldsymbol{\alpha}_j\|_2 > \omega\lambda_j \end{cases}$$

with $\lambda_j = \lambda\sqrt{m_j+1}$ and $\omega > 2$ (Fan and Li, 2001). Fan and Li (2001) suggested to set $\omega = 3.7$ in the above. For the selection of the tuning parameter $\lambda$, many procedures can be used, and in the following, we will adopt the extended Bayesian information criterion (EBIC) by following Chen and Chen (2008).

It is easy to see that the direct minimization of the penalized negative log-likelihood function $\ell_p(\boldsymbol{\alpha}, \Lambda_0)$ given in (3.4) is difficult or not feasible. Corresponding to this, we will develop a penalized EM algorithm in the next subsection by introducing subject-specific Poisson random variables in the data augmentation. For the baseline cumulative hazard function $\Lambda_0$, we will take the nonparametric approach and it will be seen that one advantage of this is that the closed-form solution of $\Lambda_0$ can be obtained.

### 3.3.2 The Penalized EM Algorithm

To present the developed penalized EM algorithm, first not that for the transformation function $G$ in model (3.1), one can adopt the Laplace transformation and write it as a frailty-induced transformation function (Wang *et al.*, 2016b; Zeng *et al.*, 2016; Li *et al.*, 2020).

$$G(x) = -log \int_0^\infty \exp(-x\,\mu)\,\rho(\mu|r)\,d\mu$$

In the above, $\rho(t)$ denotes the density function of a frailty variable $\mu$ with the support $[0, \infty)$. It is easy to show that if $\rho(\mu|r)$ is set to be the density function of the gamma distribution with mean one and variance $r$, it yields $G(x) = \log(1 + rx)$, the class of logarithmic transformations (Kosorok *et al.*, 2004). It follows that the likelihood function given in (3.3) can be rewritten as

$$L(\Lambda_0, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \int_{\mu_i} \left\{ \exp\left[-\Lambda_0(L_i)\mu_i e^{\sum_{j=1}^{p} f_{jn}(X_{ij})}\right] - \exp\left[-\Lambda_0(R_i)\mu_i e^{\sum_{j=1}^{p} f_{jn}(X_{ij})}\right] \right\} \rho(\mu_i|r) d\mu_i \,. \tag{3.5}$$

In other words, the transformation above transfers model (3.1) to the additive Cox frailty model.

Let $t_1 < t_2 < \cdots < t_{K_n}$ denote the ordered unique time points of $\{ L_i > 0, R_i < \infty; i = 1, ..., n \}$ and $h_k$ be the jump size of $\Lambda_0$ at $t_k$. Define $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{h}')$ with $\boldsymbol{h} = (h_1, ..., h_{K_n})'$. Then the likelihood function can be rewritten as

$$\begin{aligned} L(\boldsymbol{\theta}) = & \prod_{i=1}^{n} \int_{f_i} \exp\left[ - \sum_{t_k \le L_i} h_k \exp\left( \sum_{j=1}^{p} f_{jn}(X_{ij}) \right) \mu_i \right] \\ & \times \left\{ 1 - \exp\left[ - \sum_{L_i < t_k \le R_i} h_k \exp\left( \sum_{j=1}^{p} f_{jn}(X_{ij}) \right) \mu_i \right] \right\}^{I(R<\infty)} \\ & \times \rho\left(\mu_i \mid r\right) \mathrm{d}\mu_j \,. \end{aligned} \tag{3.6}$$

As mentioned above, for the baseline cumulative hazard function $\Lambda_0$, we take the nonparametric approach. To augment the observed data, by following Li *et al.* (2020), let $\{Y_{ik}, k = 1, ..., K_n\}$ be a set of independent Poisson random variables with the mean $h_k \exp\{\sum_{j=1}^{p} f_{jn}(X_{ij})\}$ for each $i$. Then the likelihood function above can be reexpressed as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{\mu_i} \Pr\left( \sum_{t_k \le L_i} Y_{ik} = 0 \right) \Pr\left( \sum_{L_i < t_k \le R_i} Y_{ik} > 0 \right)^{I(R_i<\infty)} \rho\left(\mu_i \mid r\right) \mathrm{d}\mu_i \,. \tag{3.7}$$

This suggests that one can treat the $\mu_i$'s and $Y_{ik}$'s as the pseudo complete data and the corresponding likelihood function has the form

$$L_C(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K_n} p\left(Y_{ik} \mid h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right) \mu_i\right) \rho\left(\mu_i \mid r\right) \tag{3.8}$$

with $\sum_{t_k \le L_i} Y_{ik} = 0$ and $\sum_{L_i < t_k \le R_i} Y_{ik} > 0$ if $R_i < \infty$ and $\sum_{t_i \le L_i} Y_{ik} = 0$ if $R_i = \infty$.

In the E-step of the EM algorithm, we need to calculate the conditional expectations of $Y_{ik}$ and $\mu_i$ given the observed data and the current estimates of all unknowns. For simplicity, we will denote them by $E(Y_{ik})$ and $E(\mu_i)$ by omitting the conditional arguments. More specifically, we have that

$$\mathrm{E}\left(Y_{ik}\right) = \left\{ \frac{h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right)}{\exp\left\{-G\left(V_i\right)\right\} - \exp\left\{-G\left(W_i\right)\right\}} \int_{\mu_i} \frac{\mu_i \left(\exp\left\{-\mu_i V_i\right\} - \exp\left\{-\mu_i W_i\right\}\right)}{1 - \exp\left\{-\mu_i \left(W_i - V_i\right)\right\}} \right.$$

$$\left. \times \rho\left(\mu_i \mid r\right) \mathrm{d}\mu_i I\left(L_i < t_k \le R_i\right) + h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right) \mathrm{E}\left(\mu_j\right) I\left(t_k > R_i\right) \right\} I\left(R_i < \infty\right)$$

$$+ h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right) \mathrm{E}\left(\mu_i\right) I\left(t_k > L_i\right) I\left(R_i = \infty\right) ,$$

and

$$\mathrm{E}\left(\mu_i\right) = \frac{\exp\left\{-G\left(V_i\right)\right\} G'\left(V_i\right) - \exp\left\{-G\left(W_i\right)\right\} G'\left(W_i\right)}{\exp\left\{-G\left(V_i\right)\right\} - \exp\left\{-G\left(W_i\right)\right\}} I\left(R_i < \infty\right) + G'\left(V_i\right) I\left(R_i = \infty\right)$$

where $V_i = \sum_{t_k \le L_i} h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right)$, and $W_i = \sum_{t_k \le R_i} h_k \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right)$. Note that by setting $\rho\left(\mu_i | r\right)$ to be the gamma density function as mentioned above,

we have that

$$\int_{\mu_i} \mu_i^2 \exp\left(-\mu_i x\right) \rho\left(\mu_i \mid r\right) \mathrm{d}\mu_i = (1+r)(rx+1)^{-r^{-1}-2} .$$

Also one can use the Gauss quadrature technique to calculate

$$\int_{\mu_i} \frac{\mu_i \left(\exp\left\{-\mu_i V_i\right\} - \exp\left\{-\mu_i W_i\right\}\right)}{1 - \exp\left\{-\mu_i \left(W_i - V_i\right)\right\}} \rho\left(\mu_i \mid r\right) \mathrm{d}\mu_i .$$

In the M-step of the EM algorithm, one can first update $h_k$ with the following closed-form expression

$$h_k = \frac{\sum_{i=1}^{n} \mathrm{E}\left(Y_{ik}\right)}{\sum_{i=1}^{n} \mathrm{E}\left(\mu_i\right) \exp\left(\sum_{j=1}^{p} f_{jn}(X_{ij})\right)}, \quad k = 1, \ldots, K_n . \tag{3.9}$$

It is easy to see that the existence of the closed-form solution above can considerably reduce the computational burden of the optimization process, an appealing feature with the use of the Poisson-based data augmentation. For the estimation of $\boldsymbol{\alpha}$, at the $l$th iteration, we need to minimize

$$Q_p(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)}) = Q_2(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)}) + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j) ,$$

where $\boldsymbol{\theta}^{(l)}$ denotes the estimate of $\theta$ at the $l$th iteration and

$$Q_2(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K_n} \left\{ E(Y_{ik}) \sum_{j=1}^{p} f_{jn}(X_{ij}) + E(Y_{ik}) log \left[ \sum_{i=1}^{n} E(\mu_i) exp(\sum_{j=1}^{p} f_{jn}(X_{ij})) \right] \right\} .$$

For this, we suggest to employ the group coordinate descent algorithm to update the estimates of $\boldsymbol{\alpha}$ (Yuan and Lin, 2006; Yang and Zou, 2015; Breheny and Huang, 2015; Cao

*et al.*, 2016; Lv *et al.*, 2018).

Specifically, for estimation of $\boldsymbol{\alpha}$, we will update each $\boldsymbol{\alpha}_j$ while keeping all other elements of $\boldsymbol{\alpha}$ fixed at their current estimates sequentially. To obtain the updated estimate $\hat{\boldsymbol{\alpha}}_j^{(l+1)}$ at the $(l+1)$th iteration, let $\hat{\boldsymbol{a}}_j^{(l+1)} = (\hat{\boldsymbol{\alpha}}_1^{(l+1)}, ..., \hat{\boldsymbol{\alpha}}_{j-1}^{(l+1)}, \hat{\boldsymbol{\alpha}}_j^{(l)}, ..., \hat{\boldsymbol{\alpha}}_p^{(l)})$, the current estimate of $\boldsymbol{\alpha}$ for which the first $(j-1)$th elements have been updated at the $(l+1)$th iteration, and define $Q_j'(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)}) = \partial Q_2(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)})/\partial \boldsymbol{\alpha}_j$ and $\boldsymbol{H}_j^{(l)} = \partial^2 Q_2(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(l)})/\partial \boldsymbol{\alpha}_j^2$, both evaluated at $\hat{\boldsymbol{a}}_j^{(l+1)}$. Also let $h_j^{(l)}$ denote the largest eigenvalue of $\boldsymbol{H}_j^{(l)}$ and $\hat{\boldsymbol{a}}_j^{(l+1)*} = (\hat{\boldsymbol{\alpha}}_1^{(l+1)}, ..., \hat{\boldsymbol{\alpha}}_{j-1}^{(l+1)}, \boldsymbol{\alpha}_j, \hat{\boldsymbol{\alpha}}_{j+1}^{(l)}, ..., \hat{\boldsymbol{\alpha}}_p^{(l)})$. Note that $Q_j'(\hat{\boldsymbol{a}}_j^{(l+1)}; \boldsymbol{\theta}^{(l)})$ depends on the current estimates $\hat{\boldsymbol{a}}_j^{(l+1)}$, and the objective function $Q_p(\hat{\boldsymbol{a}}_j^{(l+1)*}; \boldsymbol{\theta}^{(l)})$ can be approximated by

$$Q_p(\boldsymbol{\alpha}^*; \boldsymbol{\theta}^{(l)}) \approx Q_2(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)}) + \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right]' Q_j'(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)}) + \frac{1}{2}\left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right]' \boldsymbol{H}_j^{(l)} \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right] + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j)$$

$$\approx Q_2(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)}) + \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right]' Q_j'(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)}) + \frac{h_j^{(l)}}{2}\left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right]' \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right] + \sum_{j=1}^{p} P_\lambda(\boldsymbol{\alpha}_j)\,.$$

In the above, the sub-Hessian matrix $\boldsymbol{H}_j^{(l)}$ is replaced by its largest eigenvalue $h_j^{(l)}$. It follows that the updated estimate $\hat{\boldsymbol{\alpha}}_j^{(l+1)}$ of $\boldsymbol{\alpha}_j$ can be determined by

$$\hat{\boldsymbol{\alpha}}_j^{(l+1)} = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmin}} \left\{ \left[\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\right]' Q_j'(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)}) + \frac{h_j^{(l)}}{2} \|\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j^{(l)}\|_2^2 + P_\lambda(\boldsymbol{\alpha}_j) \right\},$$

which can yield different closed-form expressions for different penalty functions as shown below.

Withe the use of the GLASSO penalty function, the formula above gives

$$\hat{\boldsymbol{\alpha}}_j^{(l+1)} = S(\boldsymbol{d}_j; \lambda_j/h_j)\,,$$

where $\lambda_j = \lambda\sqrt{m_j + 1}$ and $S(\boldsymbol{d}_j; \lambda) = (1 - \lambda/\|\boldsymbol{d}_j\|)_+ \boldsymbol{d}_j$ with $\boldsymbol{d}_j = \hat{\boldsymbol{\alpha}}_j^{(l)} - Q_j'(\hat{\boldsymbol{\alpha}}; \boldsymbol{\theta}^{(l)})/h_j$. By

using the GMCP penalty function, one can obtain

$$\hat{\boldsymbol{\alpha}}_j^{(l+1)} = \begin{cases} S\left(\dfrac{h_j \boldsymbol{d}_j}{h_j - 1/\eta}; \dfrac{\lambda_j}{h_j - 1/\eta}\right) & \text{if } \|\boldsymbol{d}_j\|_2 \le \lambda_j \eta\,, \\ \boldsymbol{d}_j & \text{if } \|\boldsymbol{d}_j\|_2 > \lambda_j \eta. \end{cases}$$

If the group SCAD penalty function is used, we have that

$$\hat{\boldsymbol{\alpha}}_j^{(l+1)} = \begin{cases} S(\boldsymbol{d}_j; \lambda_j/h_j) & \text{if } \|\boldsymbol{d}_j\|_2 \le \lambda_j + \lambda_j/h_j\,, \\ \dfrac{\left[h_j(\omega-1) - \frac{\omega\lambda_j}{\|\boldsymbol{d}_j\|_2}\right]\boldsymbol{d}_j}{h_j\omega - h_j - 1} & \text{if } \lambda_j + \lambda_j/h_j < \|\boldsymbol{d}_j\|_2 \le \omega\lambda_j\,, \\ \boldsymbol{d}_j & \text{if } \|\boldsymbol{d}_j\|_2 > \omega\lambda_j. \end{cases}$$

Note that as mentioned above, all of the $f_j(X_j)$'s need to be centered to avoid the model identification issue. For this, define

$$\hat{f}_{jn}^*(X_j) = \sum_{k=0}^{m_j} \hat{\alpha}_{jk}^* B_{jk}(X_j, m_j, c_j, u_j) \quad \text{and} \quad \bar{f}_{jn}^*(X_j) = \sum_{i=1}^{n} \hat{f}_{jn}^*(X_{ij})/n\,.$$

Then by following Huang (1999), one can estimate $f_j(X_j)$ by

$$\hat{f}_{jn}(X_j) = \hat{f}_{jn}^*(X_j) - \bar{f}_{jn}^*(X_j)\,, \qquad 1 \le j \le p\,.$$

For given $\lambda$, the EM algorithm discussed above can be summarized as follows.

Step (1): Set initial values of $\boldsymbol{\theta}^{(0)}$.

Step (2): At the $(l+1)$th iteration, calculate $E(Y_{ik}), E(\mu_i)$.

Step (3): Obtain each element in $\hat{\boldsymbol{h}}^{(l+1)}$ at the $(l+1)$th iteration by equation (3.9).

Step (4): Determine $\hat{\boldsymbol{\alpha}}_j^{(l+1)}$ at the $(l+1)$th iteration by the group decent algorithm

$$\hat{\boldsymbol{\alpha}}_j^{(l+1)} = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmin}} \left\{ l(\hat{\boldsymbol{h}}^{(l+1)}, \hat{\boldsymbol{\alpha}}^{(l+1)*}) + P_\lambda(\boldsymbol{\alpha}_j) \right\}$$

for $j = 1, ..., p$. Then center $\hat{f}^*_{jn}(X_j)$ for each $j = 1, ..., p$.

Step (5): Repeat Steps (2) and (4) until convergence or $l$ exceeding a given large number.

For the selection of the tuning parameter $\lambda$, as mentioned above, various criteria can be used. In the numerical study below, we will employ the EBIC defined as

$$EBIC(\lambda) = -2\, log\ell(\hat{\boldsymbol{\theta}}) + S\, \log(n) + 2\,\xi\, \log \binom{p}{S}.$$

Here $\hat{\boldsymbol{\theta}}$ denotes the estimate of $\boldsymbol{\theta}$, $\ell(\cdot)$ the logarithm of the observed likelihood function given in (3.2), $S$ the total number of the estimated nonzero covariate effects, and $\xi \in [0, 1]$. It is easy to see that if $\xi = 0$, the EBIC becomes the BIC.

## 3.4 A Simulation Study

In this section, we present some results obtained from a simulation study conducted to evaluate the empirical performance of the variable selection approach proposed in the previous sections. In the study, we first generated the covariates from the multivariate normal distribution with mean zero and the correlation $corr(X_j, X_k) = 0.5^{|j-k|}$ between $X_j$ and $X_k$. Given the covariates $X_{ij}$'s, the true failure times $T_i$'s were generated based on the additive transformation model

$$\Lambda(t|\boldsymbol{X}_i) = G\left[\Lambda_0(t)\, \exp\left\{\sum_{j=1}^{p} f_j(X_{ij})\right\}\right]$$

with $\Lambda_0(t) = \log(1 + 0.5t)$ and $G(x) = \log(1 + rx)/r$ $(r \gg 0)$.

For the covariate effects, we considered the following three scenarios.

- Scenario (a): $f_1(x) = 8x - 4$, $f_2(x) = -8x + 4$, $f_3(x) = 8x - 4$, $f_4(x) = -8x + 4$, and $f_j(x) = 0\,(j = 4, ..., p)$;

- Scenario (b): $f_1(x) = 8x - 4$, $f_2(x) = 4cos(\pi x)$, $f_3(x) = -4cos(\pi x)$, $f_4(x) = 6sin(\pi x) - 3$, and $f_j(x) = 0\,(j = 4, ..., p)$;

- Scenario (c): $f_1(x) = 5cos(\pi x)$, $f_2(x) = 10sin(\pi x) - 5$, $f_3(x) = -5sin(1.5\pi(x - 0.5))$, $f_4(x) = -5cos(2\pi x)$, $f_5(x) = 5sin(1.5\pi(x - 0.5))$, $f_6(x) = -5cos(-\pi x)$, and $f_j(x) = 0\,(j = 6, ..., p)$.

Let $S$ denote the number of the covariates that have significant effects or nonzero coefficients. Then we have $S = 4$, 4 and 6 for the three situations above, respectively. Also scenario (a) has four significant covariates all having linear effects, scenario (b) includes one covariate with linear effect and three with nonlinear effects, and scenario (c) involves six significant covariates all having nonlinear effects.

For the generation of the observed data, for each subject, we first generated a sequence of observation times from the homogeneous Poisson process with the study length of 4 and the inter-examination times following the exponential distribution with the mean of 0.2. Then for subject $i$, $L_i$ and $R_i$ were set to be the largest observation time that was smaller than $T_i$ and the smallest observation time that was larger than $T_i$. For the two parameters $\eta$ and $\omega$ in the GMCP and GSCAD penalty functions, they were chosen to be 2.7 and 3.7, respectively, following Fan and Li (2001) and Zhang (2010). Furthermore, we set $\xi = 1$ in the EBIC, $m_0 = m_j = 3$ for Bernstein polynomial approximations, and $r = 0$, 0.5 and 1 for the link function $G$. As mentioned above, $r = 0$ gives the additive Cox model, while $r = 1$ corresponds to the additive proportional odds model. The results given below are based on the sample size $n = 300$ or 500 and the total number of covariates $p = 100$ or 500 with 100 replications.

Table 3.1 presents the results of the variable or covariate selection given by the proposed procedure with the use of the GLASSO, GMCP and GSCAD penalty functions under scenario (a). In the table, we calculated the average of the number of times, $TP_j$, when a

covariate was selected for each of the covariates with significant effects, the average of the number of the selected covariates, $TP$, whose true effects are significant, the average of the number of the selected covariates, $FP$, whose true effects are not significant. In addition, we determined the median of the mean squared errors, denoted by MMSE, given by $S^{-1}\sum_{j=1}^{S}\int\{f_j(x)-\hat{f}_j(x)\}^2\,dx$ along with their standard deviation denoted by SD.

The results given by the proposed approach under scenarios (b) and (c) are given in Tables 3.2 and 3.3, respectively. One can see from the three tables that the proposed variable selection procedure seems to give similar results and perform well under the three scenarios and with respect to all of three penalty functions in general. It appears that all important or significant covariates, no matter with linear or nonlinear effects, can be identified with high percentages or probabilities although the probabilities slightly decreased when the number of covariates increased. As expected, the method with the use of the GLASSO penalty function tends to yield slightly larger FP or the models with more noise and prediction errors than the other two penalty functions. Also as expected, the method gave better performance when the sample size increased.

## 3.5 An Application

Now we apply the sieve penalized variable selection procedure proposed in the previous sections to a study on Alzheimer's Disease (AD), the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is an ongoing, longitudinal and multi-site study. One primary objective of the ADNI is to investigate various factors, including genetic, demographic and clinical factors, and identify these that can be used for early diagnosis, tracking, and prevention of AD. The study consists of three groups of patients classified based on their cognitive conditions, cognitive normal, mild cognitive impairment (MCI), and AD. Among others, one event or variable of interest is the time from the initial visit date to the AD conversion

and on it, only interval-censored data are available because of the periodic examination and monitoring nature for all study participants.

For the analysis here, following Li *et al.* (2020), we will focus on the 310 patients in the MCI group with complete information on 24 factors described below. Among them, 5 factors are demographic factors, 12 are clinical factors, and the rest are associated to the patient's MRI volumetric data. More specifically, the five demographic factors are the participant's baseline age, gender (1 for male and 0 for female), apolipoprotein E $\epsilon 4$ (APOE$\epsilon 4$) genotype, years of receiving education (PTEDUCAT), and marital status (MaritalStatus, 1 for married and 0 otherwise). The clinical factors are the participant's Alzheimer's Disease Assessment Scale scores of 11 and 13 items (ADAS11 and ADAS13), delayed word recall score in ADAS (ADASQ4), clinical dementia rating scale-sum of boxes score (CDRSB), mini-mental state examination score (MMSE), Rey auditory verbal learning test score of immediate recall (RAVLT.i), learning ability (RAVLT.l), the total number of words that were forgotten in the RAVLT delayed memory test (RAVLT.f), and the percentage of words that were forgotten in the RAVLT delayed memory test (RAVLT.perc.f) as well as the participant's digit symbol substitution test score (DIGITSCOR), trails B score (TRABSCOR), and functional assessment questionnaire score (FAQ). The participant's MRI volumetric data consist of ventricles, hippocampus, whole brain (WholeBrain), entorhinal, fusiform gyrus (Fusiform), middle temporal gyrus (MidTemp), and intracerebral volume (ICV).

To apply the proposed method, as in the simulation, we chose $G(x) = \log(1 + r\,x)/r$ in model (3.1) and considered several values of $r$ between 0 and 2 with an increment of 0.1. For each value of $r$, the EBIC with $\xi = 1$ was used to select the optimal tuning parameter $\lambda$ and it turned out that the smallest EBIC value was given by $r = 0$, corresponding to the additive Cox model. Also as in the simulation, we set $m_0 = m_j = 3$ for the Bernstein polynomial approximation and scaled all covariates from 0 to 1. Table 3.4 gives the factors selected by the proposed procedure with the use of the GLASSO, GMCP, and GSCAD

penalty functions with $r = 0$. One can see from Table 3.4 that three factors, ADAS13, RAVLT.i and MidTemp, were selected by all three penalty functions and also they are the only factors selected by the GMCP and GSCAD penalty functions. In contrast, the GLASSO selected eight more factors.

Figures 3.1 - 3.3 display the estimated effects of the factors selected by using the GLASSO, GMCP, and GSCAD penalty functions, respectively. It seems that all three factors, ADAS13, RAVLT.i and MidTemp had significant and nonlinear effects on the AD conversion. More specifically, ADAS13 appears to be positively related to the risk of the AD conversion, while the other two were negatively related. All other selected factors did not seem to have significant effects. As discussed above, Li *et al.* (2020) considered the same problem discussed here under the class of semiparametric transformation models but assumed linear effects for all covariates. They identified the same three factors as above plus FAQ by using various penalty functions. Their method also suggested that RAVLT.i and MidTemp were negatively related to the risk of the AD conversion, and the ADAS13 was positively related.

## 3.6 Discussion and Concluding Remarks

In this chapter, we discussed variable selection for a class of semiparametric additive transformation models based on high-dimensional interval-censored failure time data, and a key feature of the model is that it allows for the existence and estimation of nonlinear covariate effects. For the problem, a sieve penalized maximum likelihood procedure was proposed and the proposed method can be seen as a generalization of the method given in Li *et al.* (2020). Although only the GLASSO, GMCP, and GSCAD penalty functions were considered, the proposed approach allows for the use of other group penalty functions, and for the implementation of the method, an efficient and reliable EM algorithm was developed

with the use of Poisson variables for the data augmentation. The numerical study indicated that the proposed procedure performed well for both low-dimensional and high-dimensional practical situations.

It is worth noting that the proposed method has some limitations or there exist several directions for future research. One is that in the preceding sections, we have assumed that interval censoring is non-informative or independent of the failure time of interest (Sun, 2006). As pointed out in the literature (Du *et al.*, 2022), this may not be true sometimes and in the presence of informative censoring, the use of the methods that ignore it could yield biased results. In other words, it would be useful to develop a method similar to that proposed above but allowing for informative interval censoring. Another assumption used above is that all covariates effects are constant and it is apparent that in some situations, the effects may be time-dependent (Lv *et al.*, 2018). In other words, some new methods need to be developed when there exist time-varying factors under the interval-censored data context.

Many authors have discussed the analysis of multivariate interval-censored data, and in particular, Sun *et al.* (2021) investigated the variable selection for such situations. However, their method applies only to the situation where covariate effects are linear and thus new procedures have to be developed when one faces nonlinear covariate effects. One choice would be to generalize the method proposed above to multivariate interval-censored data. As another direction for future research, it is apparent that it would be helpful to develop some methods for estimating the standard errors of the estimated covariate effects. Also it would useful to provide some theoretical justifications to or establish the asymptotic properties of the proposed method such as the oracle property and thus derive confidence bands for the estimated effect curves.

Table 3.1: Simulation results under scenario (a)

| n | p | r | Penalty | TP1 | TP2 | TP3 | TP4 | TP | FP | MMSE(SD) |
|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | 0 | glasso | 98 | 95 | 97 | 99 | 3.89 | 2.31 | 2.147 (0.581) |
| | | | gmcp | 98 | 99 | 100 | 98 | 3.95 | 0.13 | 0.375 (0.161) |
| | | | gscad | 100 | 100 | 100 | 99 | 3.99 | 0.09 | 0.355 (0.158) |
| | | 0.5 | glasso | 99 | 88 | 92 | 99 | 3.78 | 1.35 | 2.635 (0.633) |
| | | | gmcp | 99 | 95 | 92 | 100 | 3.86 | 1.37 | 0.250 (0.633) |
| | | | gscad | 100 | 92 | 91 | 99 | 3.82 | 1.33 | 0.250 (0.739) |
| | | 1 | glasso | 99 | 82 | 86 | 99 | 3.66 | 1.56 | 2.975 (0.693) |
| | | | gmcp | 100 | 98 | 95 | 100 | 3.93 | 0.41 | 0.457 (0.404) |
| | | | gscad | 99 | 100 | 97 | 98 | 3.94 | 0.53 | 0.464 (0.319) |
| | 500 | 0 | glasso | 99 | 91 | 97 | 100 | 3.87 | 4.83 | 2.574 (0.506) |
| | | | gmcp | 99 | 98 | 100 | 99 | 3.96 | 0.19 | 0.324 (0.217) |
| | | | gscad | 99 | 99 | 100 | 100 | 3.98 | 0.22 | 0.335 (0.209) |
| | | 0.5 | glasso | 98 | 77 | 85 | 98 | 3.58 | 1.64 | 3.410 (0.588) |
| | | | gmcp | 100 | 58 | 69 | 98 | 3.25 | 1.57 | 0.500 (1.316) |
| | | | gscad | 100 | 57 | 68 | 98 | 3.23 | 1.44 | 1.307 (1.311) |
| | | 1 | glasso | 98 | 70 | 75 | 99 | 3.42 | 2.52 | 3.813 (0.448) |
| | | | gmcp | 100 | 87 | 87 | 100 | 3.74 | 0.51 | 0.446 (0.748) |
| | | | gscad | 97 | 89 | 89 | 99 | 3.74 | 0.82 | 0.438 (0.717) |
| 500 | 100 | 0 | glasso | 100 | 99 | 99 | 98 | 3.96 | 1.21 | 1.722 (0.34) |
| | | | gmcp | 99 | 99 | 100 | 100 | 3.98 | 0.04 | 0.277 (0.106) |
| | | | gscad | 98 | 100 | 100 | 100 | 3.98 | 0.07 | 0.280 (0.106) |
| | | 0.5 | glasso | 100 | 100 | 99 | 100 | 3.99 | 1.07 | 2.093 (0.336) |
| | | | gmcp | 100 | 99 | 99 | 100 | 3.98 | 0.58 | 0.187 (0.254) |
| | | | gscad | 100 | 98 | 97 | 100 | 3.95 | 0.42 | 0.189 (0.333) |
| | | 1 | glasso | 100 | 98 | 97 | 100 | 3.95 | 0.77 | 2.468 (0.435) |
| | | | gmcp | 100 | 100 | 100 | 99 | 3.99 | 0.11 | 0.359 (0.132) |
| | | | gscad | 99 | 100 | 100 | 97 | 3.96 | 0.18 | 0.340 (0.131) |
| | 500 | 0 | glasso | 100 | 95 | 99 | 99 | 3.93 | 1.91 | 2.111 (0.593) |
| | | | gmcp | 99 | 100 | 99 | 97 | 3.95 | 0.1 | 0.296 (0.123)) |
| | | | gscad | 99 | 100 | 100 | 98 | 3.97 | 0.05 | 0.296 (0.124) |
| | | 0.5 | glasso | 100 | 92 | 91 | 98 | 3.81 | 1.4 | 2.478 (0.647) |
| | | | gmcp | 100 | 94 | 95 | 100 | 3.89 | 0.65 | 0.195 (0.621) |
| | | | gscad | 100 | 94 | 95 | 100 | 3.89 | 0.82 | 0.192 (0.625) |
| | | 1 | glasso | 100 | 81 | 81 | 97 | 3.59 | 1.05 | 2.974 (0.711) |
| | | | gmcp | 99 | 98 | 97 | 100 | 3.94 | 0.2 | 0.358 (0.397) |
| | | | gscad | 100 | 96 | 96 | 100 | 3.92 | 0.28 | 0.362 (0.398) |

Table 3.2: Simulation results under scenario (b)

| n | p | r | Penalty | TP1 | TP2 | TP3 | TP4 | TP | FP | MMSE(SD) |
|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | 0 | glasso | 99 | 98 | 99 | 99 | 3.95 | 2.13 | 3.022 (0.492) |
| | | | gmcp | 99 | 99 | 99 | 100 | 3.97 | 0.16 | 1.383 (0.487) |
| | | | gscad | 100 | 100 | 99 | 100 | 3.99 | 0.18 | 1.385 (0.242) |
| | | 0.5 | glasso | 100 | 100 | 100 | 96 | 3.96 | 2.58 | 3.006 (0.521) |
| | | | gmcp | 99 | 99 | 100 | 100 | 3.98 | 0.51 | 1.408 (0.314) |
| | | | gscad | 100 | 100 | 100 | 99 | 3.99 | 0.46 | 1.404 (0.280) |
| | | 1 | glasso | 98 | 94 | 99 | 89 | 3.8 | 1.47 | 3.045 (0.696) |
| | | | gmcp | 99 | 98 | 100 | 100 | 3.97 | 0.59 | 1.390 (0.498) |
| | | | gscad | 99 | 97 | 99 | 100 | 3.95 | 0.52 | 1.417 (0.645) |
| | 500 | 0 | glasso | 98 | 99 | 98 | 99 | 3.94 | 2.43 | 3.918 (0.596) |
| | | | gmcp | 99 | 94 | 99 | 100 | 3.92 | 0.22 | 1.388 (0.781) |
| | | | gscad | 99 | 92 | 96 | 100 | 3.87 | 0.3 | 1.424 (0.838) |
| | | 0.5 | glasso | 99 | 99 | 100 | 97 | 3.95 | 5.77 | 3.575 (0.591) |
| | | | gmcp | 96 | 88 | 99 | 97 | 3.8 | 1.01 | 1.462 (0.962) |
| | | | gscad | 90 | 82 | 95 | 93 | 3.6 | 0.5 | 1.485 (1.339) |
| | | 1 | glasso | 96 | 85 | 97 | 82 | 3.6 | 2.21 | 4.234 (0.652) |
| | | | gmcp | 83 | 72 | 95 | 90 | 3.4 | 0.85 | 1.634 (1.482) |
| | | | gscad | 89 | 77 | 99 | 94 | 3.59 | 0.99 | 1.592 (1.264) |
| 500 | 100 | 0 | glasso | 100 | 100 | 100 | 100 | 4 | 0.12 | 3.278 (0.355) |
| | | | gmcp | 99 | 100 | 100 | 100 | 3.99 | 0.06 | 1.343 (0.178) |
| | | | gscad | 100 | 100 | 100 | 100 | 4 | 0.05 | 1.331 (0.180) |
| | | 0.5 | glasso | 100 | 100 | 100 | 97 | 3.97 | 0.13 | 3.602 (0.304) |
| | | | gmcp | 100 | 100 | 100 | 100 | 4 | 0.08 | 1.370 (0.187) |
| | | | gscad | 100 | 100 | 100 | 100 | 4 | 0.09 | 1.370 (0.187) |
| | | 1 | glasso | 100 | 100 | 100 | 96 | 3.96 | 0.99 | 2.565 (0.335) |
| | | | gmcp | 99 | 100 | 100 | 99 | 3.98 | 0.14 | 1.314 (0.211) |
| | | | gscad | 100 | 100 | 100 | 100 | 4 | 0.09 | 1.308 (0.205) |
| | 500 | 0 | glasso | 100 | 100 | 99 | 99 | 3.98 | 1.82 | 3.103 (0.465) |
| | | | gmcp | 100 | 100 | 99 | 100 | 3.99 | 0.01 | 1.364 (0.178) |
| | | | gscad | 100 | 100 | 100 | 100 | 4 | 0.06 | 1.368 (0.184) |
| | | 0.5 | glasso | 99 | 99 | 100 | 98 | 3.96 | 1.75 | 3.099 (0.468) |
| | | | gmcp | 98 | 100 | 99 | 100 | 3.97 | 0.23 | 1.379 (0.294) |
| | | | gscad | 99 | 97 | 99 | 100 | 3.95 | 0.11 | 1.376 (0.250) |
| | | 1 | glasso | 97 | 98 | 100 | 91 | 3.86 | 1.35 | 3.446 (0.591) |
| | | | gmcp | 99 | 99 | 99 | 99 | 3.96 | 0.55 | 1.397 (0.383) |
| | | | gscad | 99 | 99 | 99 | 98 | 3.95 | 0.43 | 1.414 (0.382) |

Table 3.3: Simulation results under scenario (c)

| n | p | r | Penalty | TP1 | TP2 | TP3 | TP4 | TP5 | TP6 | TP | FP | MMSE(SD) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 100 | 0 | glasso | 99 | 98 | 100 | 99 | 99 | 100 | 5.95 | 2.06 | 9.734 (0.666) |
| | | | gmcp | 96 | 97 | 98 | 99 | 99 | 99 | 5.88 | 0.49 | 6.121 (0.744) |
| | | | gscad | 98 | 97 | 97 | 99 | 99 | 98 | 5.88 | 0.53 | 6.148 (0.747) |
| | | 0.5 | glasso | 99 | 95 | 99 | 98 | 99 | 99 | 5.89 | 1.64 | 9.424 (0.756) |
| | | | gmcp | 97 | 95 | 92 | 95 | 99 | 99 | 5.77 | 0.55 | 6.697 (0.925) |
| | | | gscad | 96 | 94 | 92 | 94 | 99 | 98 | 5.73 | 0.67 | 6.795 (0.958) |
| | | 1 | glasso | 99 | 92 | 99 | 96 | 99 | 99 | 5.84 | 0.99 | 8.885 (0.743) |
| | | | gmcp | 93 | 91 | 84 | 92 | 98 | 98 | 5.56 | 0.34 | 7.746 (0.938) |
| | | | gscad | 92 | 90 | 85 | 86 | 98 | 97 | 5.48 | 0.32 | 7.796 (1.040) |
| | 500 | 0 | glasso | 100 | 95 | 100 | 98 | 99 | 100 | 5.92 | 3.04 | 10.538 (0.431) |
| | | | gmcp | 99 | 93 | 92 | 93 | 98 | 98 | 5.73 | 0.84 | 6.416 (0.930) |
| | | | gscad | 96 | 93 | 91 | 91 | 97 | 97 | 5.65 | 0.94 | 7.160 (0.843) |
| | | 0.5 | glasso | 98 | 82 | 99 | 88 | 100 | 99 | 5.66 | 1.09 | 10.329 (0.560) |
| | | | gmcp | 95 | 89 | 91 | 78 | 98 | 98 | 5.49 | 0.58 | 7.297 (1.040) |
| | | | gscad | 92 | 87 | 89 | 69 | 95 | 97 | 5.29 | 0.69 | 7.402 (1.167) |
| | | 1 | glasso | 99 | 85 | 98 | 89 | 100 | 100 | 5.71 | 1.7 | 10.064 (0.602) |
| | | | gmcp | 94 | 92 | 90 | 82 | 96 | 97 | 5.51 | 0.53 | 7.066 (0.941) |
| | | | gscad | 91 | 87 | 86 | 78 | 93 | 96 | 5.31 | 0.39 | 7.137 (1.064) |
| 500 | 100 | 0 | glasso | 100 | 100 | 100 | 100 | 100 | 100 | 6 | 1.13 | 9.273 (0.519) |
| | | | gmcp | 100 | 100 | 100 | 100 | 100 | 100 | 6 | 0.13 | 5.872 (0.470) |
| | | | gscad | 99 | 98 | 100 | 98 | 100 | 99 | 5.94 | 0.06 | 5.873 (0.667) |
| | | 0.5 | glasso | 100 | 100 | 100 | 100 | 100 | 100 | 6 | 1.29 | 8.673 (0.577) |
| | | | gmcp | 100 | 97 | 100 | 99 | 100 | 100 | 5.96 | 0.21 | 6.393 (0.608) |
| | | | gscad | 100 | 96 | 100 | 99 | 99 | 100 | 5.94 | 0.26 | 6.393 (0.624) |
| | | 1 | glasso | 100 | 99 | 100 | 98 | 100 | 100 | 5.97 | 0.81 | 8.190 (0.646) |
| | | | gmcp | 100 | 96 | 99 | 97 | 99 | 99 | 5.9 | 0.25 | 6.347 (0.793) |
| | | | gscad | 100 | 95 | 100 | 100 | 99 | 100 | 5.94 | 0.3 | 6.331 (0.582) |
| | 500 | 0 | glasso | 100 | 95 | 100 | 98 | 100 | 100 | 5.93 | 1.76 | 9.538 (0.626) |
| | | | gmcp | 100 | 99 | 99 | 98 | 100 | 100 | 5.96 | 0.29 | 5.939 (0.562) |
| | | | gscad | 100 | 99 | 100 | 97 | 100 | 99 | 5.95 | 0.24 | 5.937 (0.609) |
| | | 0.5 | glasso | 100 | 93 | 100 | 96 | 99 | 100 | 5.88 | 1.5 | 9.442 (0.728) |
| | | | gmcp | 100 | 90 | 99 | 94 | 100 | 100 | 5.83 | 0.25 | 6.590 (0.656) |
| | | | gscad | 100 | 92 | 99 | 96 | 100 | 100 | 5.87 | 0.42 | 6.558 (0.652) |
| | | 1 | glasso | 100 | 90 | 100 | 95 | 100 | 100 | 5.85 | 1.04 | 8.646 (0.669) |
| | | | gmcp | 98 | 88 | 98 | 96 | 100 | 100 | 5.8 | 0.31 | 6.428 (0.654) |
| | | | gscad | 99 | 89 | 99 | 96 | 100 | 100 | 5.83 | 0.38 | 6.409 (0.614) |

Table 3.4: Variable selection results with $r = 0$ of ADNI data

| **Covariate** | **GLASSO** | **GMCP** | **GSCAD** |
|---|---|---|---|
| Gender | | | |
| MaritalStatus | | | |
| AGE | | | |
| PTEDUCAT | | | |
| APOE$\epsilon$4 | + | | |
| ADAS11 | + | | |
| ADAS13 | + | + | + |
| ADASQ4 | + | | |
| CDRSB | | | |
| MMSE | + | | |
| RAVLT.i | + | + | + |
| RAVLT.l | | | |
| RAVLT.f | | | |
| RAVLT.perc.f | + | | |
| DIGITSCOR | | | |
| TRABSCOR | + | | |
| FAQ | + | | |
| Ventricles | | | |
| Hippocampus | | | |
| WholeBrain | | | |
| Entorhinal | + | | |
| Fusiform | | | |
| MidTemp | + | + | + |
| ICV | | | |

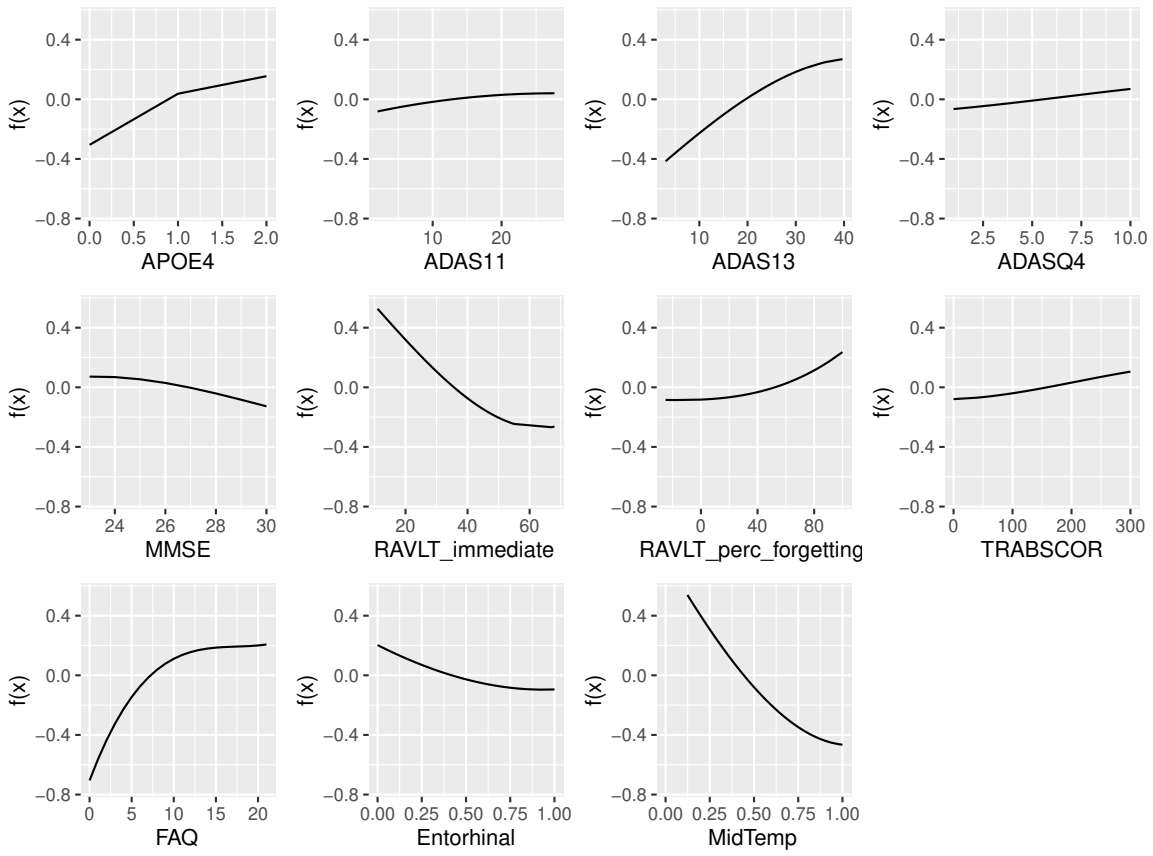



Figure 3.1: Estimated effects of covariates selected by GLASSO penalty

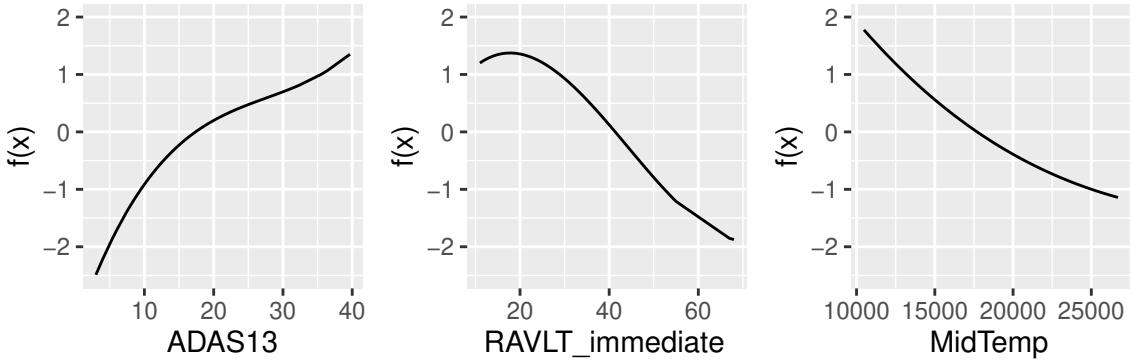



Figure 3.2: Estimated effects of covariates selected by GMCP penalty

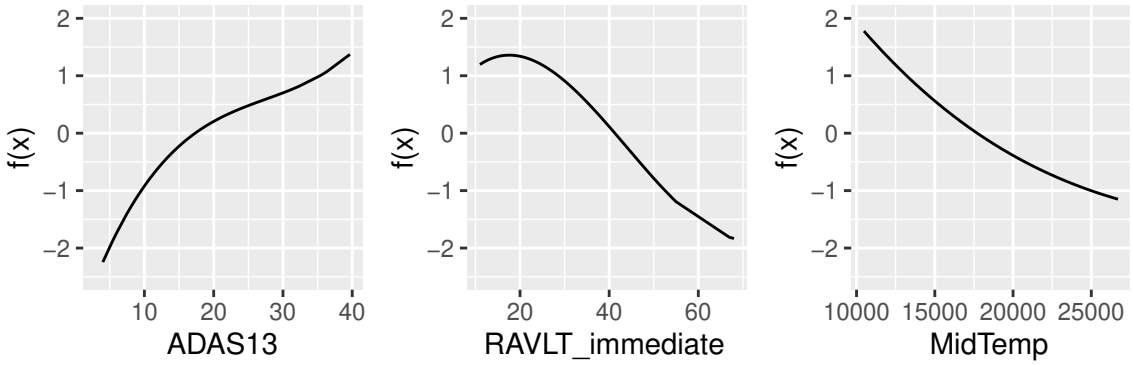



Figure 3.3: Estimated effects of covariates selected by GSCAD penalty

# Chapter 4

# Variable Selection for Partially Functional Additive Cox Model with Interval-censored Data

## 4.1 Introduction

In regression analysis of time-to-event data, Cox's proportional hazards model, with the merits of straightforward interpretation and easy implementation, has become the most popular and widely used model that assumes covariates take a linear form (Cox, 1972). However, the applications of the standard Cox model are sometimes too restricted in practice owing to the assumption of linear covariate effects. It is appealing to adopt additive Cox models that assume nonlinear covariates effects to provide flexible modeling for time-to-event data to incorporate nonlinearity. In addition, by treating longitudinal predictors as functional data, more information from longitudinal trajectory of multiple predictors can be obtained, which can effectively improve estimation and variable selection performance. Motivated by the time of AD progression along with a large number of genetic factors

and sparsely observed longitudinal neurocognitive assessments in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, we propose to perform the variable selection by using a novel partially functional additive Cox model (PFACM) with a set of functional predictors and a set of scalar predictors with possibly nonlinear covariate effects. In particular, the longitudinal trajectories of neurocognitive assessments can be modeled by functional data analysis techniques, namely, functional principal component analysis (FPCA).

The proposed partially functional additive Cox model (PFACM) is an extension of the standard Cox model that combines a functional linear Cox regression model (Kong *et al.*, 2018; Gellar *et al.*, 2015; Qu *et al.*, 2016; Lee *et al.*, 2015; Hao *et al.*, 2021) and a nonparametric additive Cox model (Huang, 1999; Du *et al.*, 2010; Lv *et al.*, 2018; Wu *et al.*, 2020). And it can capture both underlying trajectories of longitudinal covariates and nonlinear patterns of scalar covariates. Most existing functional survival models tend to depend on the assumption of proportional hazards structure and right-censored data. For instance, Kong *et al.* (2018) proposed the functional linear Cox regression model (FLCRM) to discover functional characteristics from surface data by applying functional principal component analysis (FPCA). Lee *et al.* (2015) extended the estimation of the FLCRM to the Bayesian framework. Gellar *et al.* (2015) and Qu *et al.* (2016) investigated penalized partial likelihood functions. These methods with functional survival models are only designed for right-censored failure time data. However, interval-censoring is sometimes more realistic and appropriate and often observed in medical and health studies with regular follow-ups.

The analysis of interval-censored data intensifies considerable technical difficulty in computation and implementation because there is no partial likelihood function feasible. One needs to cope with both unknown baseline hazard functions and coefficients. For functional regression models with interval-censored data, Ye *et al.* (2015) proposed a joint modeling method based on functional data analysis for longitudinal data and interval-censored time. Shi *et al.* (2022) studied a functional proportional hazard cure rate model with functional

and scalar predictors for sparsely and irregularly sampled longitudinal data and interval-censored failure time data. To deal with such sparsely and irregularly sampled longitudinal data, the authors adopted the principal analysis by conditional expectation (PACE) approach proposed by Yao *et al.* (2005a) for functional disease trajectory feature extraction.

To our knowledge, only limited research has been focusing on variable selection for functional survival models, even for right-censored data (Fang *et al.*, 2016; Shi *et al.*, 2021). To fill the research gap, we propose a flexible partially functional additive Cox model (PFACM) to distinguish critical predictors with underlying functional and nonlinear structures under the interval-censored failure time data framework. Specifically, a unified penalized sieve estimation approach is developed where the PACE method and Bernstein polynomials are used for modeling the functional predictors and scalar predictors with nonlinear covariate effects, respectively. Moreover, three group penalty functions of the group least absolute shrinkage and selection operator (GLASSO) (Yuan and Lin, 2006), group smoothly clipped absolute deviation (GSCAD) (Fan and Li, 2001; Huang *et al.*, 2012), and group minimum concave penalty (GMCP) (Zhang, 2010; Huang *et al.*, 2012) are considered for regularization of predictors, and a group coordinate descent algorithm is adopted for implementation.

The main contributions of the proposed method are three-fold. Firstly, our method is capable of handling the estimation of covariates with potential nonlinear effects as well as functional covariates. Both of them are infinite-dimensional. Secondly, the proposed penalized estimation method can regularize estimates of all functional covariates and scalar covariates simultaneously while identifying significant covariates with possibly nonlinear effects and sparsely measured longitudinal covariates under the Cox model framework. Thirdly, the proposed method is designed for the interval-censored failure time data, which is more reasonable but complicated for some real-world problems, such as the ADNI data with regular examination times. Note that interval-censoring includes right-censoring as a particular case.

The remainder of this chapter is organized as follows. Section 4.2 introduces the proposed partially functional additive Cox regression model and constructs the observed data likelihood function. The Bernstein polynomials and principal analysis by conditional expectation (PACE) methods are adopted to approximate unknown baseline hazard functions and additive components as well as functional covariates. In Section 4.3, we employ a penalized estimation procedure using the GLASSO, GMCP, and GSCAD penalty functions. The group coordinate descent algorithm is developed for the implementation in Section 4.4. In Section 4.5, a simulation study is performed to evaluate the empirical performance of the variable selection method. The ADNI study described in the previous chapters is applied to the proposed method in Section 4.6, and concluding remarks are presented in Section 4.7.

## 4.2 Partially Functional Additive Cox Model

Consider a failure time study that involves two sets of predictors, which are a set of $P$-dimensional time-independent scalar predictors $Z$ with potential nonlinear covariate effects and a set of $J$-dimensional functional predictors $X(s)$, $s \in S$, where $S$ is the observation time window of the trajectories. Suppose that the study recruits $n$ subjects ($i = 1, ..., n$) and $T_i$'s are the actual failure time of interest. We aim to identify a small subset of $Z$ and $X(s)$ that are relevant to the failure time $T$.

In many real-world problems, the failure time $T$ cannot be observed exactly and may suffer from interval-censoring, which means that the exact failure time $T$ is only known within a time interval that covers $T \in (L_i, R_i]_{i=1,...,n}$. For the $j$th functional predictor with $i$th subject at $k$th visit, one can only observe covariates $Y_{ij}(s_{ik})$ with measurement errors instead of the true covariates $X_{ij}(s_{ik})$, where $Y_{ij}(s_{ik}) = X_{ij}(s_{ik}) + \epsilon_{ijk}$ and $\epsilon_{ijk}$'s are independent and identically distributed random errors with mean zero and variance $\sigma^2_{\epsilon_j}$ ($i =$

$1, ..., n; j = 1, ...J; k = 1, ..., K_i)$. To simplify the notation, we let $Y_{ijk} = Y_{ij}(s_{ik})$ and $Y_{ijk}$ is the $j$th longitudinal outcomes for subject $i$ at the time $s_{ik}(s_{ik} \leq L_i)$. Therefore, the observed data is $(\{Y_{1jk}, 1 \leq k \leq K_1\}, Z_1, (L_1, R_1]), \ldots, (\{Y_{njk}, 1 \leq \ k \leq K_n\}, Z_n, (L_n, R_n])$, where $(L_i, R_i]$ indicates the interval that brackets $T_i$. That is, only interval-censored data are observed. It can be seen that $L_i = 0$ or $R_i = \infty$ represents a left- or right-censored observation on $T_i$, respectively.

Under the partially functional additive Cox model (PFACM), the cumulative hazard function of the $i$th subject is

$$H_i(t) = H_0(t) \exp\left\{\sum_{p=1}^{P} f_p(Z_i) + \sum_{j=1}^{J} \int_{\mathcal{S}} X_{ij}(s)\beta_j(s)ds\right\}, \tag{4.1}$$

where $Z_i = (Z_{i1}, \ldots, Z_{iP})^{\mathrm{T}}$, $\beta_j(\cdot)$ is the $j$th coefficient function, and $T \in [0, \tau]$ for some finite $\tau > 0$. The model (4.1) is determined by the unknown functions $\beta_j(\cdot)$ and $f_p(\cdot)$, and the cumulative baseline hazard function $H_0(\cdot)$. For the identifiability of the model above, it will be assumed that $E[f_p(\cdot)] = 0$, $1 \leq p \leq P$ (Huang, 1999).

In the following, we will assume an independent or noninformative interval-censoring mechanism (Sun, 2006). Then the observed likelihood function can be written as

$$\begin{aligned} L(H_0, f_p, \beta_j) &= \prod_{i=1}^{n} \{\exp[-H(L_i)] - \exp[-H(R_i)]\} \\ &= \prod_{i=1}^{n} \left\{\exp\left[-H_0(L_i)e^{\sum_{p=1}^{P} f_p(Z_{ip}) + \sum_{j=1}^{J} \int_{\mathcal{S}} X_{ij}(s)\beta_j(s)ds}\right]\right. \\ &\quad \left. - \exp\left[-H_0(R_i)e^{\sum_{p=1}^{P} f_p(Z_{ip}) + \sum_{j=1}^{J} \int_{\mathcal{S}} X_{ij}(s)\beta_j(s)ds}\right]\right\}. \end{aligned} \tag{4.2}$$

### 4.2.1 Bernstein Polynomials Approximation

Following Zhou *et al.* (2017), we adopt the sieve approach with the use of the Bernstein polynomials expansion to approximate $H_0(t)$ and $f_p(\cdot)$. The Bernstein polynomials expansion for modeling the monotone and nonnegative function $H_0(t)$ is given by

$$H_{0n}(t) = \sum_{l=0}^{m} \psi_l^* B_{0l}(t, m, c, u),\ 0 \leq \psi_0^* \leq \psi_1^* \leq ... \leq \psi_m^*\,,$$

where $B_{0l}(t, m, c, u)$ is the Bernstein basis polynomials of degree $m = o(n^{\omega})$ for some $\omega \in (0, 1)$ and has the form

$$B_{0l}(t, m, c, u) = \binom{m}{l} \left(\frac{t-c}{u-c}\right)^l \left(1 - \frac{t-c}{u-c}\right)^{m-l},\ l = 0, 1, ..., m.$$

Note that the constraint $0 \leq \psi_0^* \leq \psi_1^* \leq ... \leq \psi_m^*$ is required to guarantee the monotonic properties of $\Lambda_{0n}(t)$ and can be easily addressed by the re-parameterization $\psi_0^* = e^{\psi_0}$, $\psi_d^* = \sum_{i=0}^{l} e^{\psi_i}$, $\forall\ 1 \leq l \leq m$ (Zhou *et al.*, 2017; Wang and Ghosh, 2012).

The Bernstein polynomials expansion for $f_p(\cdot)$ is

$$f_{pn}(Z_p) = \sum_{l=0}^{m_p} \alpha_{pl} B_{pl}(Z_p, m_p, c_p, u_p),\ p = 1, ..., P\,,$$

where

$$B_{pl}(Z_p, m_p, c_p, u_p) = \binom{m_p}{l} \left(\frac{Z_p - c_p}{u_p - c_p}\right)^l \left(1 - \frac{Z_p - c_p}{u_p - c_p}\right)^{m_p - l},\ l = 0, 1, ..., m_p\,.$$

$B_{pl}(Z_p, m_p, c_p, u_p)$ is the Bernstein basis polynomials of $m_p$ degrees. It is apparent that there is no constraint on $f_p(\cdot)$, and the re-parameterization is not needed. Spline and other polynomial functions are also applicable to replace the Bernstein polynomials. Zhou *et al.* (2017) mentioned that the Bernstein polynomials expansion has the optimal shape preserv-

ing property that can naturally model $H_0$ with the nonnegative and monotone properties, and no specification of interior knots is required in the approximation. In the following subsection, we will employ the functional principal component analysis (FPCA) to characterize the pattern of sparsely and irregularly observed trajectories of functional predictors.

### 4.2.2 Functional Principal Component Analysis

In the following, we propose to estimate the trajectories of covariates $X(s)$ and the coefficients functions $\beta(\cdot)$ by the functional principal component analysis (FPCA) (Yao *et al.*, 2005a; Shi *et al.*, 2022; Li and Luo, 2019). To extract information from the sparsely-measured longitudinal data, following Yao *et al.* (2005a), we consider the principal analysis conditional expectation (PACE) approach, which can accommodate sparsely and irregularly sampled longitudinal data.

For the PACE approach, let $\mu_j(s) = E(X_j(s))$ be the mean function and $G_j(s, s') = \text{Cov}(X_j(s) - \mu_j(s), X_j(s') - \mu_j(s'))$ be the covariance function of the functional predictors $\{X_j(s), s \in S\}$. It is natural to consider the Karhunen-Lovève decomposition in the FPCA approach, and then $X_{ij}(s)$ can be expanded as

$$X_{ij}(s) = \mu_j(s) + \sum_{r=1}^{\infty} \xi_{ijr}\phi_{jr}(s),$$

where $\phi_{jr}(s)$ is the $r$th functional principal component (FPC) with constraint that $\int_{\mathcal{S}} \phi_{jr}(s)\phi_{jr'}(s) = \delta_{rr'}$, and $\delta_{rr'} = 1$ if $r = r'$ and 0 otherwise. $\xi_{ijr}$ is the corresponding FPC score for the $i$th subject and $j$th predictor.

Define the FPC score $\xi_{ijr}$ as

$$\xi_{ijr} = \int_{\mathcal{S}} \left(X_{ij}(s) - \mu_j(s)\right) \phi_{jr}(s) ds$$

with $E(\xi_{ijr}) = 0$ and $\text{Var}(\xi_{ijr}) = \lambda_{jr}$, where $\lambda_{j1} \geq \lambda_{j2} \geq ... \geq 0$. By Mercer's theorem, the covariance function $G_j(s, s')$ can be written as

$$G_j(s, s') = \sum_{r=1}^{\infty} \lambda_{jr} \phi_{jr}(s) \phi_{jr}(s').$$

Next, we will consider estimating the FPCs and first develop smoothed estimates of the mean and covariance functions. Let $\hat{\mu}_j(s)$ be the estimates of the mean function and denote $\hat{G}_j(s, s')$ as the estimates of the covariance function. One can obtain the estimated mean function and covariance function $G_j(s, s')$ by local linear smoother and a two-dimensional kernel smoother, respectively. Therefore, the estimates for $\phi_{jr}(s)$ and $\lambda_{jr}$ can be solved via the following eigenequation

$$\int_{\mathcal{S}} \hat{G}_j(s, s') \phi_{jr}(s) ds = \lambda_{jr} \phi_{jr}(s')$$

with the constraints $\|\phi_{jr}\|^2 = 1$ and $\langle \phi_{jr}, \phi_{jr'} \rangle = 1$ if $r = r'$, and 0 otherwise. By using spectral decomposition to the discretized covariance function $\hat{G}_j(s, s')$, the estimated eigenequation $\hat{\phi}_r(s)$ can be solved.

Following Yao *et al.* (2005a), we will take advantage of the PACE method and calculate the FPC scores of the $i$th subject, the $j$th predictors, and the $r$th FPC as follows.

$$\hat{\xi}_{ijr} = \hat{E}(\xi_{ijr} \mid Y_{ij}) = \lambda_{jr} \hat{\phi}_{ijr}^T \hat{\Sigma}_{Y_{ij}}^{-1} (Y_{ij} - \hat{\mu}_{ij}),$$

where $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijK_i})^T$, $\hat{\Sigma}_{Y_{ij}} = \widetilde{G}_{ij} + \hat{\sigma}^2 I_{K_i}$. Here, $\hat{\phi}_{ijr}$ and $\hat{\mu}_{ij}$ represent the vectors of values of $\hat{\phi}_{jr}(s)$ and $\hat{\mu}_j(s)$ evaluated at time points $s_{ik}$. And $\widetilde{G}_{ij}$ denotes the matrix of values of $\hat{G}_j(s, s')$ evaluated at the two-dimensional grid consisting of $s_{ik}$.

Based on the basis of FPCs $\{\phi_{jr}(s), r = 1, \ldots, \infty\}$, $\beta_j(s)$ can be expanded as

$$\beta_j(s) = \sum_{r=1}^{\infty} \beta_{jr}\phi_{jr}(s),$$

where $\beta_{jr} = \int_{\mathcal{S}} \beta_j(s)\phi_{jr}(s)ds$ are the basis coefficients. Since the number of FPCs and FPCs scores are infinite, one can approximate $X_{ij}(s)$ and $\beta_j(s)$ by remaining the first $R_n$ number of FPCs and associated FPC scores in practice. Denote $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jR_n})$. Consequently, the integral $\int_{\mathcal{S}} X_{ij}(s)\beta_j(s)ds$ in (1) can be computed and the model (4.1) can be re-expressed as

$$H_i(t) = H_{0n}^*(t) \exp\left\{\sum_{p=1}^{P} f_{pn}(Z_{ip}) + \sum_{j=1}^{J}\sum_{r=1}^{R_n} \xi_{ijr}\beta_{jr}\right\}, i = 1, ..., n, \tag{4.3}$$

where $H_{0n}^*(t) = H_{0n}(t)exp\{\sum_{j=1}^{J} \int_{\mathcal{S}} \mu_j(s)\beta_j(s)ds\}$. It is worth noting that the Bernstein polynomials expansion and the FPCA approach convert the estimation problem with infinite-dimensional parameters to a simpler one with finite-dimensional parameters.

## 4.3 Penalized Estimation and Variable Selection Procedure

In this section, we will establish penalized estimation procedure for both $Z$ and $X(\cdot)$. Let $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, \boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ contain all unknown parameters, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T = (\beta_{11}, ..., \beta_{1R_n}, ..., \beta_{J1}, ..., \beta_{JR_n})^T$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_P)^T = (\alpha_{10}, ..., \alpha_{1m_1}, ..., \alpha_{P0}, ..., \alpha_{Pm_P})^T$ and $\boldsymbol{\psi} = (\psi_0, ..., \psi_m)^T$. To carry out variable selection and obtain the sparse estimator of $\boldsymbol{\theta}$, we first let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_{P+J})^T = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_P, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T$, it is natural to employ

the following penalized negative log-likelihood

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \sum_{d=1}^{P+J} P_\lambda(\boldsymbol{\eta}_d)\,, \tag{4.4}$$

where $P_\lambda(\boldsymbol{\eta}_d)$ denotes a group penalty function of $\boldsymbol{\eta}_d$ characterized by the tuning parameter $\lambda$. And $\ell(\boldsymbol{\theta})$ is the negative log-likelihood function with respect to $\boldsymbol{\theta}$ derived from the observed data likelihood function (4.2) and takes the following form

$$\ell(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log\Big\{\exp\Big(-H_{0n}^*(L_i)e^{\sum_{p=1}^{P} f_{pn}(Z_{ip})+\sum_{j=1}^{J}\sum_{r=1}^{r_n}\xi_{ijr}\beta_{jr}}\Big) \\ - \exp\Big(-H_{0n}^*(R_i)e^{\sum_{p=1}^{P} f_{pn}(Z_{ip})+\sum_{j=1}^{J}\sum_{r=1}^{r_n}\xi_{ijr}\beta_{jr}}\Big)\Big\}\,,$$

where $H_{0n}^*(t) = H_{0n}(t)\exp\{\sum_{j=1}^{J}\int_{\mathcal{S}}\mu_j(s)\beta_j(s)ds\}$. It is easy to see that $\boldsymbol{\eta}_d = \mathbf{0}$ suggests that the corresponding covariate $Z_p$ or $X_j$ has no significant effect. Then sparse estimator for $\boldsymbol{\eta}_d$ can be obtained by minimizing the penalized negative log-likelihood function (4.4).

For the penalty function $P_\lambda(\boldsymbol{\eta}_d)$, we will consider the following three group penalty functions. One is the group LASSO penalty function given by

$$P_\lambda(\eta_d) = \lambda_d\|\boldsymbol{\eta}_d\|_2$$

with $\lambda_d = \lambda\sqrt{m_d+1}$ and $\|\boldsymbol{\eta}_d\|_2 = \sqrt{\sum_{l=0}^{m_d}\eta_{dl}^2}$ (Yuan and Lin, 2006), and the group MCP penalty function is also considered

$$P_\lambda(\boldsymbol{\eta}_d;\gamma) = \begin{cases} \lambda_d\|\boldsymbol{\eta}_d\|_2 - \dfrac{\|\boldsymbol{\eta}_d\|_2^2}{2\gamma} & \text{if } \|\boldsymbol{\eta}_d\|_2 \le \lambda_d\gamma\,, \\ \dfrac{\lambda_d^2\gamma}{2} & \text{if } \|\boldsymbol{\eta}_d\|_2 > \lambda_d\gamma \end{cases}$$

with $\lambda_d = \lambda\sqrt{m_d+1}$ and $\gamma > 1$ (Zhang, 2010). The third one is the group SCAD penalty

function

$$P_\lambda(\boldsymbol{\eta}_d;\kappa)=\begin{cases}\lambda_d\,\|\boldsymbol{\eta}_d\|_2 & \text{if } \|\boldsymbol{\eta}_d\|_2\le\lambda_d\,,\\ \dfrac{2\kappa\lambda_d\|\boldsymbol{\eta}_d\|_2-\|\boldsymbol{\eta}_d\|_2^2-\lambda_d^2}{2(\kappa-1)} & \text{if } \lambda_d<\|\boldsymbol{\eta}_d\|_2\le\kappa\lambda_d\,,\\ \dfrac{(\kappa^2-1)\lambda_d^2}{2(\kappa-1)} & \text{if } \|\boldsymbol{\eta}_d\|_2>\kappa\lambda_d\end{cases}$$

with $\lambda_d=\lambda\sqrt{m_d+1}$ and $\kappa>2$ (Fan and Li, 2001; Huang *et al.*, 2012). It is possible to investigate other group penalty functions, and one needs to develop an associated estimation procedure for different penalty functions.

## 4.4 Group Coordinate Descent Algorithm

We employ the group coordinate descent algorithm for implementation. The group coordinate descent algorithm is stable and efficient and can be applied for low- and high-dimensional cases. Each unknown function is approximated by either functional principal component analysis or Bernstein polynomials associated with a group of parameters that can be included or excluded together in the final model by the group coordinate descent algorithm. Let $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\eta}}$ be the estimators of $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$, where $\boldsymbol{\eta}=(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T$ can be estimated by minimizing the penalized negative log-likelihood function $\ell_p(\boldsymbol{\theta})$. In the following, for the determination of $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\eta}}$, the group coordinate descent algorithm (Yuan and Lin, 2006; Yang and Zou, 2015; Breheny and Huang, 2015; Cao *et al.*, 2016; Lv *et al.*, 2018) is used to estimate $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ alternately.

To determining $\hat{\boldsymbol{\eta}}$, we will update each $\boldsymbol{\eta}_d$ while letting all other elements of $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$ fixed at their current estimates. More specifically, to obtain the updated estimate $\hat{\boldsymbol{\eta}}_d^{(q+1)}$ at the $(q+1)^{\text{th}}$ iteration, let $\hat{\boldsymbol{\psi}}^{(q+1)}$ be the current estimate of $\boldsymbol{\psi}$ at the $(q+1)^{\text{th}}$ iteration and let $\hat{\boldsymbol{\eta}}_d^{(q+1)}=(\hat{\boldsymbol{\eta}}_1^{(q+1)},...,\hat{\boldsymbol{\eta}}_{d-1}^{(q+1)},\hat{\boldsymbol{\eta}}_d^{(q)},...,\hat{\boldsymbol{\eta}}_{P+J}^{(q)})$ be the current estimate of $\boldsymbol{\eta}$ for which the first $(d-1)$th elements have been updated at $(q+1)$th iteration, and

define $\ell'_d(\boldsymbol{\psi}, \boldsymbol{\eta}) = \partial \ell(\boldsymbol{\psi}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}_d$, $H_d^{(q)} = \partial^2 \ell(\boldsymbol{\psi}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}_d^2$, both evaluated at $\hat{\boldsymbol{\eta}}_d^{(q+1)}$. Let $\hat{\boldsymbol{\eta}}_d^{(q+1)*} = (\hat{\boldsymbol{\eta}}_1^{(q+1)}, ..., \hat{\boldsymbol{\eta}}_{d-1}^{(q+1)}, \boldsymbol{\eta}_d, \hat{\boldsymbol{\eta}}_{d+1}^{(q)}, ..., \hat{\boldsymbol{\eta}}_{P+J}^{(q)})$ and $h_d^{(q)}$ denote the largest eigenvalue or diagonal element of $H_d^{(q)}$. Note that $\ell'_d(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)})$ depends on the current estimates $\hat{\boldsymbol{\psi}}^{(q+1)}$ and $\hat{\boldsymbol{\eta}}_d^{(q+1)}$, and the objective function $\ell_p(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)*})$ can be approximated by

$$\begin{aligned}
\ell_p(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)*}) &\approx \ell(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)}) + \left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right]' \ell'_d(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)}) \\
&\quad + \frac{1}{2}\left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right]' H_d^{(q)} \left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right] + \sum_{d=1}^{P+J} P_\lambda(\boldsymbol{\eta}_d) \\
&\approx \ell(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)}) + \left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right]' \ell'_d(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)}) \\
&\quad + \frac{h_d^{(q)}}{2}\left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right]' \left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right] + \sum_{d=1}^{P+J} P_\lambda(\boldsymbol{\eta}_d)
\end{aligned}$$

with the sub-Hessian matrix $H_d^{(q)}$ being replaced by its largest eigenvalue $h_d^{(q)}$. The updated estimate $\hat{\boldsymbol{\eta}}_d^{(q+1)}$ of $\boldsymbol{\eta}_d$ can be determined by

$$\hat{\boldsymbol{\eta}}_d^{(q+1)} = \underset{\boldsymbol{\eta}_j}{\operatorname{argmin}} \left\{ \left[\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\right]' \ell'_d(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)}) + \frac{h_d^{(q)}}{2} \|\boldsymbol{\eta}_d - \hat{\boldsymbol{\eta}}_d^{(q)}\|_2^2 + P_\lambda(\boldsymbol{\eta}_d) \right\}.$$

The closed-form expressions of various penalty functions can be obtained and will be presented in the following. To simplify the notation, we let $h_d^{(q)} = h_d$, then the estimate of the parameter $\boldsymbol{\eta}$ gained from the group LASSO penalty function takes a form

$$\hat{\boldsymbol{\eta}}_d^{(q+1)} = S(\boldsymbol{c}_d; \lambda_d/h_d)\,,$$

where $\lambda_d = \lambda\sqrt{m_d + 1}$ and $S(\boldsymbol{c}_d; \lambda_d) = (1 - \lambda/\|\boldsymbol{c}_d\|)_+ \boldsymbol{c}_d$ with $\boldsymbol{c}_d = \hat{\boldsymbol{\eta}}_d^{(q)} - \ell'_d(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)})/h_d$. With the use of the group MCP penalty function, we have

$$\hat{\boldsymbol{\eta}}_d^{(q+1)} = \begin{cases} S\left(\dfrac{h_d \boldsymbol{c}_d}{h_d - 1/\gamma}; \dfrac{\lambda_d}{h_d - 1/\gamma}\right) & \text{if } \|\boldsymbol{c}_d\|_2 \le \lambda_d \gamma\,, \\ \boldsymbol{c}_d & \text{if } \|\boldsymbol{c}_d\|_2 > \lambda_d \gamma. \end{cases}$$

$$
\overset{(q+1)}{d} = \begin{cases} S(\boldsymbol{c}_d; \lambda_d/h_d) & \text{if } \|\boldsymbol{c}_d\|_2 \le \lambda_d + \lambda_d/h_d\,, \\ \dfrac{\left[h_d(\kappa-1) - \frac{\kappa\lambda_d}{\|\boldsymbol{c}_d\|_2}\right]\boldsymbol{c}_d}{h_d\kappa - h_d - 1} & \text{if } \lambda_d + \lambda_d/h_d < \|\boldsymbol{c}_d\|_2 \le \kappa\lambda_d\,, \\ \boldsymbol{c}_d & \text{if } \|\boldsymbol{c}_d\|_2 > \kappa\lambda_d. \end{cases}
$$

It is worth pointing out that all $f_p(Z_p)$'s need to be centered on avoiding the model identification issue. For this, define

$$
\hat{f}^*_{pn}(Z_p) = \sum_{l=0}^{m_p} \hat{\boldsymbol{\eta}}^*_{pl} B_{pl}(Z_p, m_p, c_p, u_p) \quad \text{and} \quad \bar{\hat{f}}^*_{pn}(Z_p) = \sum_{i=1}^{n} \hat{f}^*_{pn}(Z_{ip})/n\,.
$$

The final estimator of $f_p(Z_p)$ will be defined as

$$
\hat{f}_{pn}(Z_p) = \hat{f}^*_{pn}(Z_p) - \bar{\hat{f}}^*_{pn}(Z_p)\,, \qquad 1 \le p \le P\,.
$$

In the following, we develop a detailed estimation procedure .

Step 1: Apply the FPCA/PACE to functional covariates, then calculate estimated FPC scores $\xi_{ijr}$ of all subjects and use $\xi_{ijr}$ as new covariates.

Step 2: Set $q = 0$ and choose the initial estimates $\hat{\boldsymbol{\psi}}^{(0)}, \hat{\boldsymbol{\eta}}^{(0)} = (\hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\beta}}^{(0)})^T$.

Step 3: At the $(q+1)$th iteration, calculate the estimates $\hat{\boldsymbol{\psi}}^{(q+1)}$ by using the Nelder-Mead simplex algorithm.

Step 4: Calculate the $\hat{\boldsymbol{\eta}}^{(q+1)}$ by the group coordinate descent algorithm. The sparse estimate of $\boldsymbol{\eta}_d$ can be obtained by minimizing the following objective function

$$
\hat{\boldsymbol{\eta}}_d^{(q+1)} = \underset{\boldsymbol{\eta}_d}{\operatorname{argmin}} \left\{ l(\hat{\boldsymbol{\psi}}^{(q+1)}, \hat{\boldsymbol{\eta}}_d^{(q+1)*}) + P_\lambda(\boldsymbol{\eta}_d) \right\}
$$

for $d = 1, ..., P + J$. Then center $\hat{f}_{pn}(Z_p)$ for each $p = 1, ..., P$.

Step 5: Repeat steps 2 - 4 until convergence or exceed a pre-specified number of iterations.

Particularly, we conduct a grid search of $\lambda_1, \lambda_2$ with Bayesian information criterion

(BIC) developed by Schwarz (1978) to select the optimal values of the tuning parameter $\lambda = (\lambda_1, \lambda_2)$. The degree of the Bernstein polynomials is set as $m_0 = m_p = 3$ suggested by Wu *et al.* (2020). In the FPCA, we set the number of basis functions as 3, and the number of FPCs and associated FPC scores $R_n$ is chosen by the pre-determined percentage of variance explained (PVE), such as 95% in simulation and application.

## 4.5 A Simulation Study

An intensive simulation study is performed to evaluate the finite-sample performance of the proposed variable selection approach. In the simulation, the nonfunctional covariates $Z_{n\times P}$ were generated from the multivariate normal distribution with mean zero and the covariance matrix $\Sigma_X$ with the $(j,k)$th element being $0.2^{|j-k|}$. We consider to generate 100 simulated data with the sample size $n = (500, 800)$ and the number of covariates $Z_p$ is $P = (10, 50)$ in each scenario. Suggested by Fan and Li (2001) and Zhang (2010), we set parameters $\gamma = 2.7$ and $\kappa = 3.7$ in the group MCP and SCAD penalty functions, respectively. To simplify the setting, we assume that among all $Z_p$, the first $P_{imp}$ of them are important, with covariates having linear or nonlinear forms. We consider setting true functions of $Z_p$ in the following three scenarios:

- Case (a) has two important $Z_p$, and one is linear covariate effects and another one is nonlinear covariate effects. $f_1(Z_1) = -4Z_1 + 2$, $f_2(Z_2) = 2sin(\pi(Z_2 - 0.5))$ and $f_p(Z_p) = 0, p = 3, ..., P$.

- Case (b) has two important $Z_p$ and all of them have nonlinear covariate effects. $f_1(Z_1) = 4sin(\pi(Z_1 - 0.5))$, $f_2(Z_2) = 4cos(\pi Z_2)$ and $f_p(Z_p) = 0, p = 3, ..., P$.

- Case (c) has three important $Z_p$ and all of them have nonlinear covariate effects. $f_1(Z_1) = 4sin(\pi(Z_1 - 0.5))$, $f_2(Z_2) = 4cos(\pi Z_2)$, $f_3(Z_3) = -4.5Z_3^2 - 3Z_3 + 2.7$ and

$f_p(Z_p) = 0, p = 4, ..., P.$

In other words, the first $P_{imp}$ of $Z_p$ have true covariate effects that are nonzero, and the rest of $Z_p$ are unimportant covariates with zero covariate effects.

For the generation of the longitudinal covariates and corresponding covariate effects, we simulate the following three underlining true trajectories $X_j(s), j = 1, .., 3$, where

- $X_1(s) = \cos(2\pi(s - a_1)) + a_2 s, a_1 \sim N\left(-1, 2^3\right), a_2 \sim N(2, 1),$
- $X_2(s) = b_1 \sin(2\pi s) + b_2, b_1 \sim U(3, 5), b_2 \sim N(0, 1),$
- $X_3(s) = c_1 s^2 + c_2 s + c_3, c_1 \sim N\left(2, 0.5^2\right), c_2 \sim N(-2, 1), c_3 \sim N\left(2, 1.5^2\right),$

and error term $\epsilon \sim N(0, 0.01)$. The true coefficient functions $\beta(s)$'s are given by $\beta_1(s) = \sin(2\pi s)$, $\beta_2(s) = \sin(\pi s)$, $\beta_j(s) = 0, j = 3, ..., J$, which implies that the first two coefficient functions $\beta_1(s)$ and $\beta_2(s)$ are nonzero, and others are all equal to zero. To generate $s_i$, we first let the observation time window $s_i$ be eight equally spaced time points from 0 to 1, and then randomly select six to eight time points for each subject $i$. Then the true failure times $T_i$'s are generated from model (4.1) with $H_0(t) = \nu t^{\zeta}$.

To generate interval-censoring, we set the observation time window to be ten equally spaced time points from $(0, \tau)$ with a fifty percent probability of observing the time point for each subject $i$, where $\tau$ is the length of study. The parameters of $\tau$, $\zeta$, and $\nu$ are controlled by obtaining desirable left- and right- censored rates in different cases. We set $\zeta = 3, \nu = 0.5$ and $\tau = 4$, yielding a right-censored rate of 15%. For subject $i$, $L_i$ is specified as the largest observation time point less than $T_i$, and $R_i$ is specified as the smallest observation time point greater than $T_i$. For all Bernstein polynomials approximations, the degree is set to be $m_0 = m_p = 3$. In the FPCA method, the number of FPCs is determined by the proportion of variance explained not smaller than 95%.

Owing to the primary interest being to select significant covariates with actual nonzero covariate effects, we will calculate the following measurements and focus on the variable

selection performance so that the proposed method can correctly select the significant variables and exclude insignificant variables. The true positive ($\text{TP}_Z$) and false positive ($\text{FP}_Z$) represent the number of correctly selected important variables with actual covariate effects being nonzero and the number of incorrectly selected unimportant variables with actual covariate effects being zero for covariates $Z$. Similarly, $\text{TP}_X$ and $\text{FP}_X$ stand for the true positive and false positive for covariates $X$, respectively. $\text{TP}_{Z_p}$ and $\text{TP}_{X_j}$ are denoted as the number of times of the covariate is selected among 100 replicates, for $p = 1, ..., P_{imp}$ and $j = 1, ..., J_{imp}$. The average mean square error of important scalar covariates is defined as $\text{MSE}_Z = \frac{1}{P_{imp}} \sum_{p=1}^{P_{imp}} \int \{f_p(Z_p) - \hat{f}_p(Z_p)\}^2 dx$ for function $f_p$, $p = 1, ..., P_{imp}$. The average mean square error of important functional covariates is defined as $\text{MSE}_\beta = \frac{1}{J_{imp}} \sum_{j=1}^{J_{imp}} \int \{\beta_j(s) - \hat{\beta}_j(s)\}^2 ds$ for coefficient functions $\beta_j(s), j = 1, ..., J_{imp}$.

Tables 4.1 - 4.3 display the variable selection results in cases (a), (b) and (c), respectively. One can see from the tables that the proposed variable selection approach seems to perform reasonably well in all cases with the use of all three penalty functions. With respect to the incorrectly selected nonzero covariate $Z$, it seems that $\text{FP}_Z$ increases when the number of important covariates $Z$ increases, which means more insignificant covariates are included in the model. In general, the variable selection accuracy is improved when the sample increases. Also, it is expected that the GLASSO penalty tends to remain more noise in the model in terms of larger $\text{FP}_Z$ and $\text{FP}_X$ than other penalty functions, which was also mentioned by Li *et al.* (2020). Table 4.4 shows the simulation results of case (b) when $m_0 = m_p = 5$. It seems that Table 4.4 performs similarly to Table 4.2 with $m_0 = m_p = 3$ .

## 4.6 An Application

Now we apply the proposed method to the motivating example of the ADNI study, which contains a rich set of neuropsychological measurements and genetic factors. It is apparent

that the time to AD conversion suffers from interval-censoring because of the periodic follow-up nature. We will consider the 407 individuals with complete information on the $327,354$ single nucleotide polymorphisms (SNPs) in the mild cognitive impairment (MCI) group. Each SNP is continuous and has values of $0, 1, 2$ corresponding to the number of Allele T. Following others (Li and Luo, 2019), we also include five longitudinal neuropsychological assessment factors in the MCI group. The five neuropsychological assessment factors are participants' Alzheimer's Disease Assessment Scale-Cognitive 13 items (ADAS-Cog 13), Rey Auditory Verbal Learning Test score of immediate recall (RAVLT-immediate), Rey Auditory Verbal Learning Test score of learning (RAVLT-learning), Functional assessment questionnaire score (FAQ) and Mini-Mental State Examination (MMSE). Since missing data are detected in the five longitudinal neurocognitive measurements, we adopt the multiple imputation approach developed by Honaker *et al.* (2011) with the available $R$ package *Amelia* to impute missing values. The right-censored rate in the study is around 45%.

Figure 4.1 displays 30 randomly selected patients from the ADNI study with longitudinal trajectories of the ADAS-Cog 13 (panel A) and RAVLT-learning (panel B). Some information can be obtained from this figure. The variables ADAS-Cog 13 and RAVLT-learning are observed repeatedly, sparely, and irregularly with missing values. Moreover, the longitudinal trajectories of each individual are heterogeneous. Therefore, the PACE method for sparse longitudinal data is well suited to accommodate such data. To model the longitudinal predictors by functional principal component analysis, we first keep the measurements of the five factors before the left examination times $L_i$ that AD has not occurred as functional data. Furthermore, to reduce the dimensionality of the SNPs, following Wu *et al.* (2020), we select the top 500 SNPs correlated to time-to-AD transition by using a sure independent screening method (Fan and Lv, 2008). Also as in the simulation, we set $m_0 = m_p = 3$ for the Bernstein polynomial approximation and scaled all covariates between 0 and 1. In the study, we are interested in distinguishing key risk factors that can

significantly affect the time to AD progression from MCI, where failure time of interest is the time from the baseline to AD conversion.

Table 4.5 and 4.6 present estimated nonzero covariate effects of neurocognitive assessments and SNPs identified by the GLASSO, GMCP and GSCAD penalty functions, respectively. It can be seen that among five longitudinal covariates, the ADAS-Cog 13 and FAQ are selected by all penalty functions. Figures 4.2 - 4.4 show the estimated effects of SNPs selected by using GLASSO, GMCP and GSCAD penalty functions. The results in these tables and figures show that the GLASSO, GMCP, and GSCAD can identify 31, 6, and 6 SNPs, respectively. Among them, 6 SNPs are commonly selected by all three penalty functions, which are rs10957236, rs11894455, rs17170619, rs1786747, rs2169089, and rs2587535. Note that 8 SNPs identified by the method in this chapter are also selected by the method in Chapter 2, which are rs12042017, rs12454238, rs1475950, rs170519, rs2428754, rs10879354, rs17170619 and rs2444907. Figures 4.5 - 4.7 display the estimated effects of neuropsychological assessment factors versus time in years selected by using the GLASSO, GMCP, and GSCAD penalty functions. The figures imply that the ADAS-Cog 13 and FAQ have significant effects on AD conversion. As discussed in Chapter 3, the similar variable selection problem was considered only for twenty-four demographic and clinical factors with the assumption of nonlinear effects, including the five neuropsychological assessment factors described above. The method in Chapter 3 identifies the same two factors as above, plus FAQ and RAVT-immediate, using all three penalty functions.

## 4.7 Discussion and Concluding Remarks

The occurrence of interval-censoring and the presence of multiple functional covariates lead to intensive computational and theoretical challenges. Direct optimization is infeasible due to the infinite dimensionality of parameters. In the chapter, we proposed a unified variable

selection technique for the partly functional additive Cox regression model under interval censorship and sparsely and irregularly sampled functional data. Specifically, we modeled the multiple longitudinal predictors as functional data with the principal analysis of conditional expectation (PACE) approach and approximated the unknown baseline hazards function and additive components of scalar predictors by the Bernstein polynomials. Therefore, a group penalized sieve estimation approach was proposed with the group coordinate descent algorithm for implementation. The penalization method can simultaneously carry out estimation and variable selection of both functional predictors and scalar predictors. Simulation studies demonstrate selection accuracy, and results suggest that the presented approach is flexible and robust in different scenarios. The results of the ADNI study indicate that the proposed variable selection method can identify six SNPs and two longitudinal neurocognitive assessments by all three group penalty functions.

Compared with functional linear models, we introduce an additive form of covariate functions, improving the flexibility of the variable selection method that can account for nonlinear covariate effects. Compared with general nonparametric additive models, the proposed partially functional model can take advantage of longitudinal data with repeated measurements. However, current work depends on several assumptions that can be relaxed in the future. The first assumption is the independent or noninformative interval censoring mechanism. It may not always be accurate and can be violated in many cases (Sun, 2006). The second assumption is the Cox model structure, and one can investigate other survival models, such as a class of transformation models with the Cox model as a particular case. The third assumption is to assume a univariate time-to-event outcome in the model. It is often to confront multivariate failure time data in practice. Among others, Sun *et al.* (2021) proposed a variable selection technique for multivariate interval-censored data under a general class of semiparametric transformation frailty models. An expectation-maximization (EM) approach that incorporates a minimum information criterion into the optimization

procedure was employed to produce the parameter estimator.

Establishing asymptotic properties of the variable selection method, such as the oracle property, is one crucial direction of future research. Nevertheless, it is not trivial to derive such theoretical property owing to the presence of interval censorship and trajectories of various functional data. Moreover, it is straightforward to deal with densely sampled functional data, such as image data, by replacing the PACE method with the standard functional principal component analysis (FPCA). Another direction is to increase the computation efficiency of the algorithm. Although the proposed method can be applied for high-dimensional cases, it is time-consuming and computationally expensive when the number of covariates is large. A parallel computation or more efficient algorithms are appealing to explore in the future.

Table 4.1: Variable selection results for case (a) and $m_0 = m_p = 3$

| | $TP_{Z_1}$ | $TP_{Z_2}$ | $TP_Z$ | $FP_Z$ | $MSE_Z(SD)$ | $TP_{X_1}$ | $TP_{X_2}$ | $TP_X$ | $FP_X$ | $MSE_\beta(SD)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | n=500, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 1.03 | 0.957 (0.182) | 100 | 85 | 1.85 | 0.93 | 7.004 (26.092) |
| GMCP | 97 | 100 | 1.97 | 0.33 | 0.563 (0.204) | 96 | 79 | 1.75 | 0.05 | 9.971 (33.435) |
| GSCAD | 99 | 100 | 1.99 | 0.51 | 0.573 (0.163) | 100 | 89 | 1.89 | 0.12 | 9.978 (34.010) |
| | | | | | n=500, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 1.14 | 1.179 (0.234) | 100 | 84 | 1.84 | 0.93 | 7.314 (14.809) |
| GMCP | 99 | 100 | 1.99 | 0.12 | 0.613 (0.180) | 100 | 67 | 1.67 | 0.04 | 9.430 (23.273) |
| GSCAD | 99 | 100 | 1.99 | 0.42 | 0.606 (0.177) | 100 | 79 | 1.79 | 0.13 | 9.768 (22.613) |
| | | | | | n=800, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 0.91 | 0.910 (0.156) | 100 | 93 | 1.93 | 0.93 | 6.359 (12.291) |
| GMCP | 100 | 100 | 2 | 0.47 | 0.566 (0.122) | 100 | 90 | 1.9 | 0.03 | 8.248 (14.937) |
| GSCAD | 99 | 100 | 1.99 | 0.34 | 0.566 (0.132) | 100 | 93 | 1.93 | 0.1 | 8.316 (17.181) |
| | | | | | n=800, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 1.18 | 1.083 (0.159) | 100 | 82 | 1.82 | 0.95 | 5.992 (6.906) |
| GMCP | 95 | 100 | 1.95 | 0.08 | 0.613 (0.271) | 97 | 73 | 1.70 | 0.02 | 7.721 (8.157) |
| GSCAD | 99 | 100 | 1.99 | 1.36 | 0.609 (0.135) | 99 | 81 | 1.8 | 0.07 | 7.821 (8.169) |

Table 4.2: Variable selection results for case (b) and $m_0 = m_p = 3$

| | $TP_{Z_1}$ | $TP_{Z_2}$ | $TP_Z$ | $FP_Z$ | $MSE_Z(SD)$ | $TP_{X_1}$ | $TP_{X_2}$ | $TP_X$ | $FP_X$ | $MSE_\beta(SD)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | n=500, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.179 (0.313) | 100 | 48 | 1.48 | 0.88 | 4.060 (10.862) |
| GMCP | 100 | 100 | 2 | 0.08 | 1.163 (0.238) | 96 | 70 | 1.66 | 0.04 | 10.846 (24.542) |
| GSCAD | 100 | 100 | 2 | 0 | 1.163 (0.209) | 99 | 72 | 1.71 | 0.04 | 11.115 (24.800) |
| | | | | | n=500, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 0.01 | 3.174 (0.318) | 100 | 39 | 1.39 | 0.86 | 3.654 (11.633) |
| GMCP | 100 | 100 | 2 | 0.2 | 1.200 (0.216) | 96 | 69 | 1.65 | 0.05 | 9.634 (26.736) |
| GSCAD | 100 | 100 | 2 | 0.04 | 1.188 (0.206) | 98 | 69 | 1.67 | 0.02 | 9.153 (26.741) |
| | | | | | n=800, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.337 (0.360) | 100 | 36 | 1.36 | 0.92 | 2.795 (4.510) |
| GMCP | 100 | 100 | 2 | 0.07 | 1.215 (0.283) | 100 | 75 | 1.75 | 0.03 | 8.014 (16.027) |
| GSCAD | 100 | 100 | 2 | 0 | 1.208 (0.188) | 100 | 79 | 1.79 | 0.02 | 7.891 (16.080) |
| | | | | | n=800, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.228 (0.251) | 100 | 34 | 1.34 | 0.83 | 3.145 (6.361) |
| GMCP | 100 | 100 | 2 | 0.07 | 1.223 (0.369) | 99 | 74 | 1.73 | 0.04 | 8.197 (16.634) |
| GSCAD | 100 | 100 | 2 | 0.01 | 1.216 (0.148) | 100 | 73 | 1.73 | 0 | 8.182 (17.783) |

Table 4.3: Variable selection results for case (c) and $m_0 = m_p = 3$

| | $TP_{Z_1}$ | $TP_{Z_2}$ | $TP_{Z_3}$ | $TP_Z$ | $FP_Z$ | $MSE_Z(SD)$ | $TP_{X_1}$ | $TP_{X_2}$ | $TP_X$ | $FP_X$ | $MSE_\beta(SD)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | n=500, P=10 | | | | | |
| GLASSO | 100 | 100 | 100 | 3 | 1.54 | 2.045 (0.333) | 100 | 85 | 1.85 | 0.98 | 9.694 (18.725) |
| GMCP | 100 | 100 | 100 | 3 | 0.25 | 1.294 (0.231) | 100 | 85 | 1.85 | 0.25 | 14.241 (23.482) |
| GSCAD | 100 | 100 | 100 | 3 | 0.02 | 1.270 (0.207) | 100 | 84 | 1.84 | 0.29 | 14.299 (23.654) |
| | | | | | | n=500, P=50 | | | | | |
| GLASSO | 100 | 100 | 100 | 3 | 2.71 | 2.670 (0.476) | 100 | 78 | 1.78 | 0.97 | 6.910 (18.336) |
| GMCP | 100 | 100 | 100 | 3 | 2.6 | 1.275 (0.239) | 100 | 83 | 1.83 | 0.28 | 10.113 (27.724) |
| GSCAD | 100 | 100 | 100 | 3 | 0.52 | 1.279 (0.223) | 100 | 82 | 1.82 | 0.34 | 10.295 (27.653) |
| | | | | | | n=800, P=10 | | | | | |
| GLASSO | 100 | 100 | 100 | 3 | 0.89 | 2.119 (0.319) | 100 | 86 | 1.86 | 1 | 7.653 (18.933) |
| GMCP | 100 | 100 | 100 | 3 | 0.07 | 1.362 (0.227) | 100 | 81 | 1.81 | 0.18 | 11.173 (23.324) |
| GSCAD | 100 | 100 | 100 | 3 | 0 | 1.350 (0.208) | 100 | 88 | 1.88 | 0.33 | 10.915 (23.333) |
| | | | | | | n=800, P=50 | | | | | |
| GLASSO | 100 | 100 | 100 | 3 | 4.28 | 2.056 (0.300) | 100 | 87 | 1.87 | 0.98 | 5.554 (17.340) |
| GMCP | 100 | 100 | 100 | 3 | 0.61 | 1.313 (0.174) | 100 | 90 | 1.9 | 0.27 | 7.314 (25.299) |
| GSCAD | 100 | 100 | 100 | 3 | 0 | 1.307 (0.151) | 100 | 87 | 1.87 | 0.34 | 7.395 (25.390) |

Table 4.4: Variable selection results for case (b) and $m_0 = m_p = 5$

| | $TP_{Z_1}$ | $TP_{Z_2}$ | $TP_Z$ | $FP_Z$ | $MSE_Z(SD)$ | $TP_{X_1}$ | $TP_{X_2}$ | $TP_X$ | $FP_X$ | $MSE_\beta(SD)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | n=500, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.469 (0.314) | 100 | 31 | 1.31 | 0.86 | 3.972 (14.352) |
| GMCP | 100 | 100 | 2 | 0.43 | 1.355 (0.259) | 97 | 70 | 1.67 | 0 | 12.861 (26.799) |
| GSCAD | 100 | 100 | 2 | 0.13 | 1.358 (0.259) | 99 | 75 | 1.74 | 0.03 | 12.714 (26.551) |
| | | | | | n=500, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 0.01 | 3.500 (0.288) | 100 | 27 | 1.27 | 0.82 | 3.857 (12.914) |
| GMCP | 100 | 100 | 2 | 2.15 | 1.351 (0.261) | 98 | 64 | 1.62 | 0.03 | 10.704 (30.047) |
| GSCAD | 100 | 100 | 2 | 0.31 | 1.359 (0.256) | 97 | 69 | 1.66 | 0.03 | 10.713 (30.117) |
| | | | | | n=800, P=10 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.634 (0.274) | 100 | 37 | 1.37 | 0.89 | 3.198 (6.131) |
| GMCP | 100 | 100 | 2 | 0.36 | 1.447 (0.213) | 100 | 73 | 1.73 | 0.01 | 9.334 (17.416) |
| GSCAD | 100 | 100 | 2 | 0.08 | 1.453 (0.213) | 100 | 81 | 1.81 | 0.03 | 8.745 (17.560) |
| | | | | | n=800, P=50 | | | | | |
| GLASSO | 100 | 100 | 2 | 0 | 3.582 (0.221) | 100 | 29 | 1.29 | 0.82 | 3.690 (8.959) |
| GMCP | 100 | 100 | 2 | 0.21 | 1.459 (0.188) | 99 | 67 | 1.66 | 0 | 8.758 (20.105) |
| GSCAD | 100 | 100 | 2 | 0.06 | 1.454 (0.190) | 100 | 76 | 1.76 | 0 | 8.765 (20.227) |

Table 4.5: Selected covariates in ADNI data

| **Covariate** | **GLASSO** | **GMCP** | **GSCAD** |
|---|---|---|---|
| ADAS-Cog 13 | + | + | + |
| RAVLT-immediate | + | | |
| RAVLT-learning | | | |
| MMSE | | | |
| FAQ | + | + | + |

Table 4.6: Selected SNPs in ADNI data

| SNP | GLASSO | GMCP | GSCAD |
|---|---|---|---|
| rs10491224 | + | | |
| rs10506650 | + | | |
| rs10517034 | + | | |
| rs10850973 | + | | |
| rs10879354 | + | | |
| rs10957236 | + | + | + |
| rs11022360 | + | | |
| rs11060153 | + | | |
| rs11083215 | + | | |
| rs11229457 | + | | |
| rs11894455 | + | + | + |
| rs12042017 | + | | |
| rs12454238 | + | | |
| rs12595578 | + | | |
| rs12607795 | + | | |
| rs1475950 | + | | |
| rs1540967 | + | | |
| rs1572237 | + | | |
| rs1584033 | + | | |
| rs16315 | + | | |
| rs16893622 | + | | |
| rs170519 | + | | |
| rs17170619 | + | + | + |
| rs1786747 | + | + | + |
| rs1984823 | + | | |
| rs2169089 | + | + | + |
| rs2191849 | + | | |
| rs2428754 | + | | |
| rs2444907 | + | | |
| rs2580358 | + | | |
| rs2587535 | + | + | + |

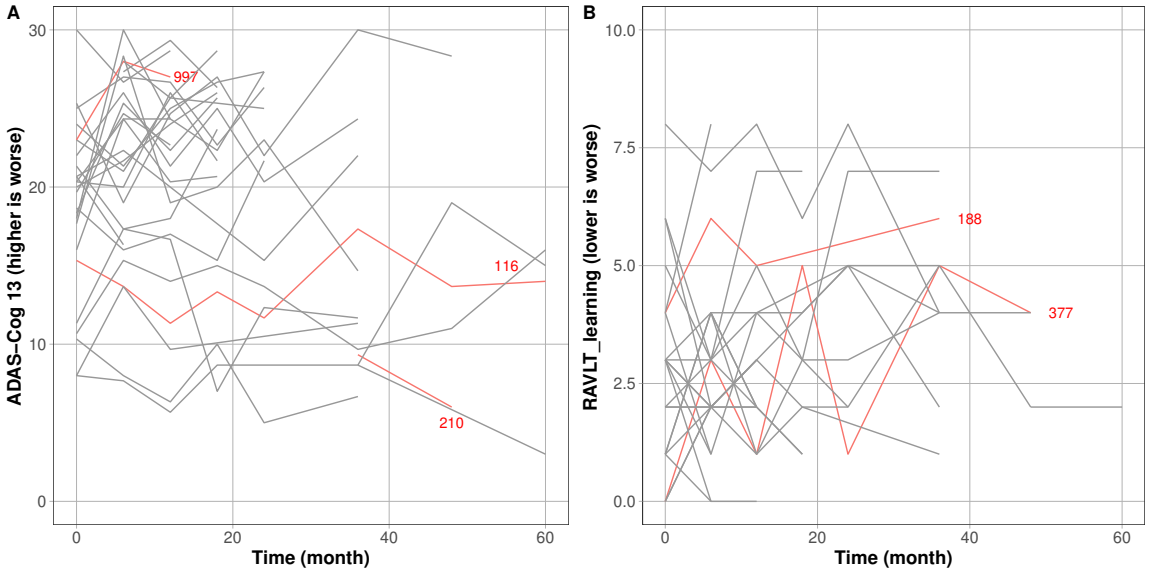



Figure 4.1: Thirty randomly selected patients in the MCI group from the ADNI study. A: Longitudinal trajectories of Alzheimer's Disease Assessment Scale-Cognitive 13 items (ADAS-Cog 13); B, Longitudinal trajectories of Rey Auditory Verbal Learning Test score of learning (RAVLT-learning)

Figure 4.2: Estimated effects of the SNPs selected by GLASSO penalty

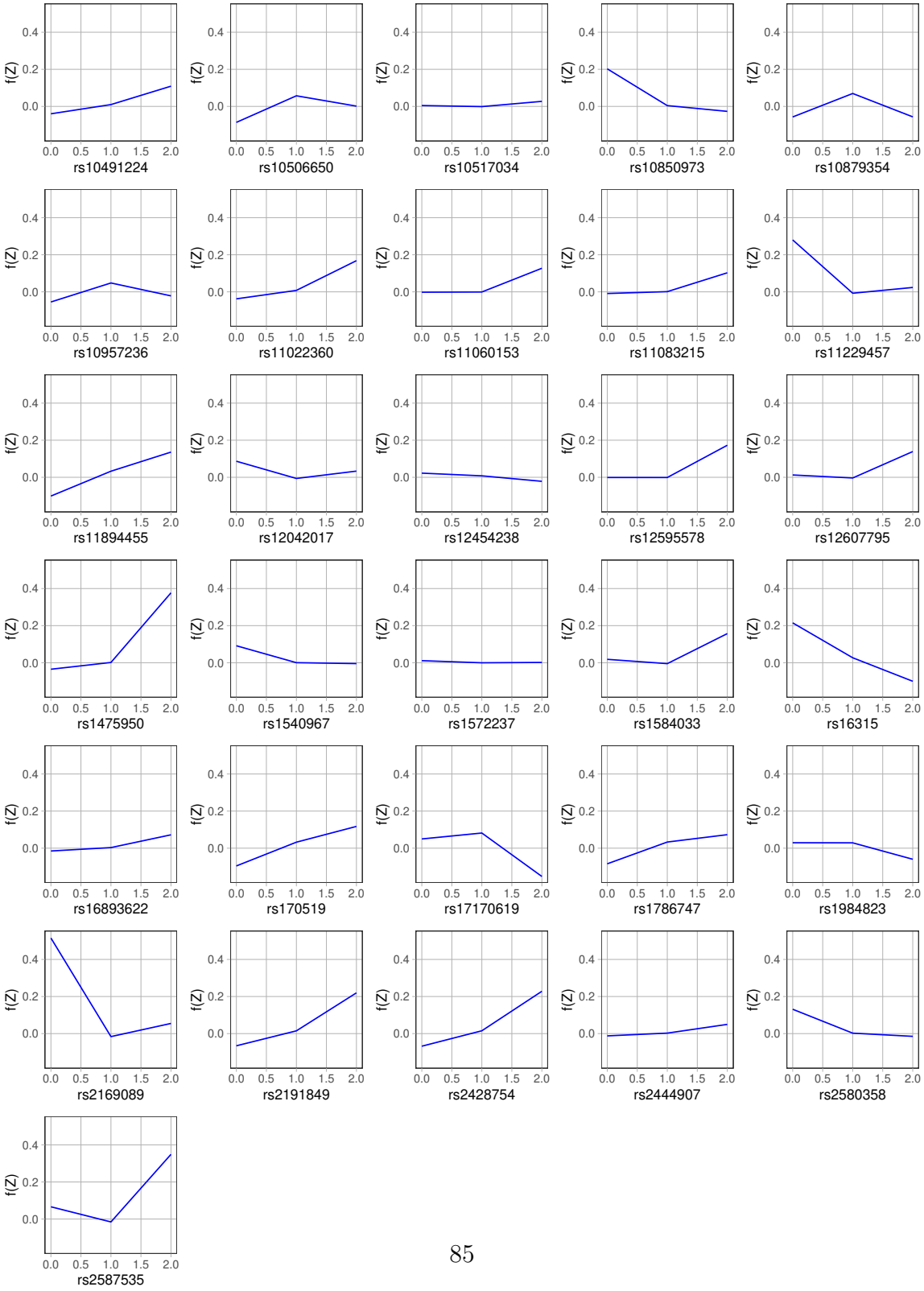


Figure 4.3: Estimated effects of the SNPs selected by GMCP penalty

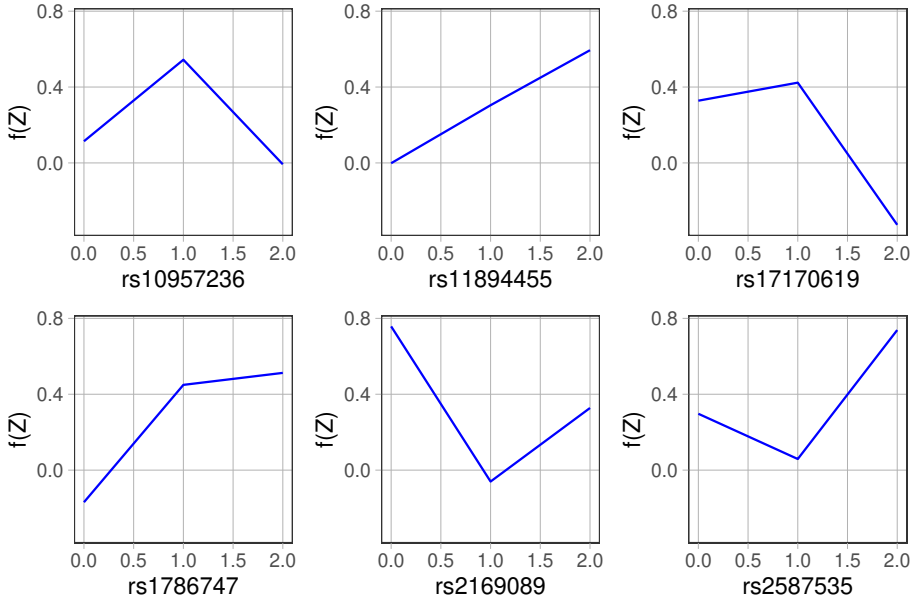



Figure 4.4: Estimated effects of the SNPs selected by GSCAD penalty

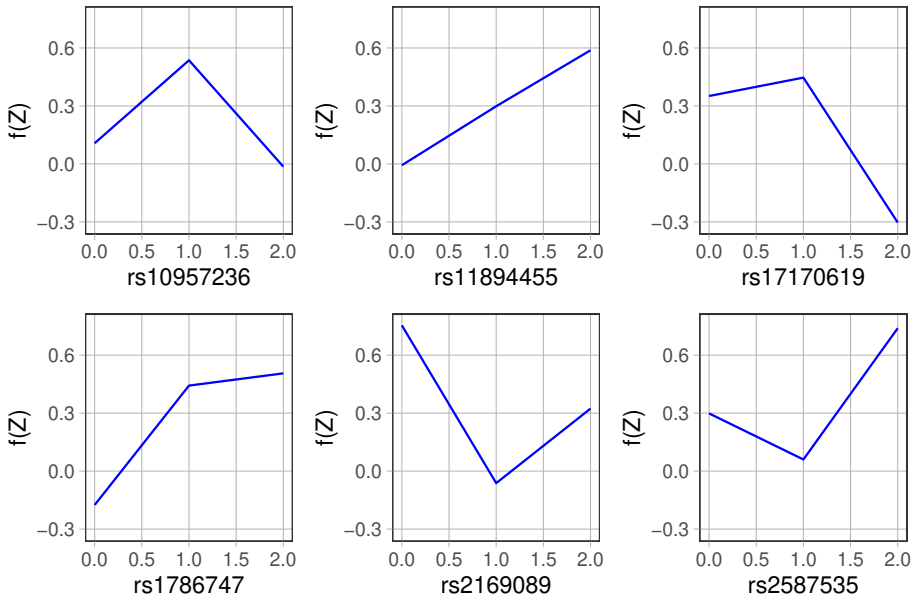


Figure 4.5: Estimated effects of the neurocognitive assessment factors selected by GLASSO penalty

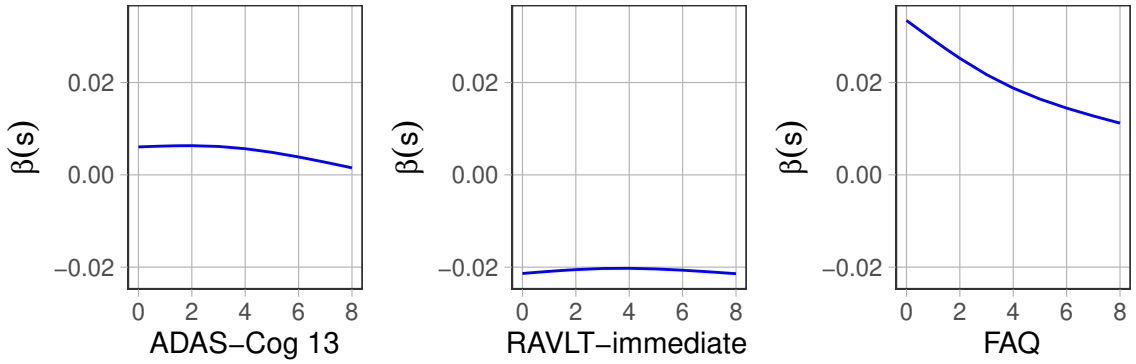



Figure 4.6: Estimated effects of the neurocognitive assessment factors selected by GMCP penalty

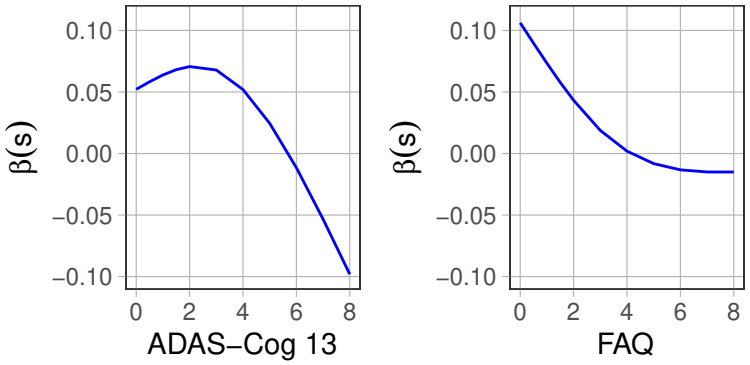
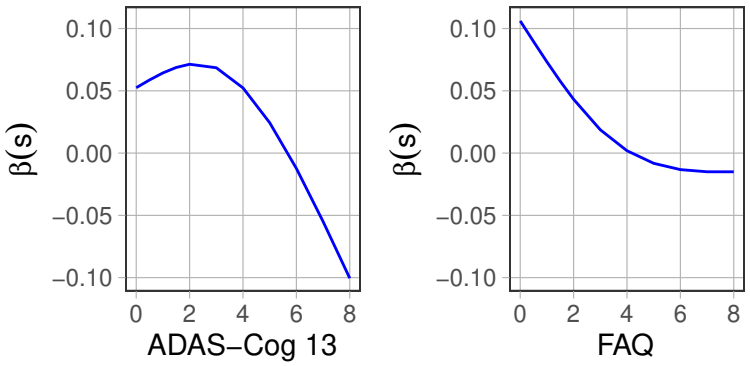



Figure 4.7: Estimated effects of the neurocognitive assessment factors selected by GSCAD penalty



# Chapter 5

# Future Research

In this chapter, we will point out several potential directions for future research that are related to the variable selection for interval-censored failure time data and functional data.

## 5.1 Variable Selection for Partially Functional Additive Transformation Models with Interval-censored Data

The current work in Chapter 4 considers the additive Cox model with functional covariates. It will be of great interest if we extend such a problem to a flexible class of transformation models, namely, partially functional additive transformation models. Under the partially functional additive transformation models, the cumulative hazard function of the $i$th subject is

$$H_i(t) = G\left[H_0(t)\exp\left\{\sum_{p=1}^{P} f_p(Z_i) + \sum_{j=1}^{J}\int_{\mathcal{S}} X_{ij}(s)\beta_j(s)ds\right\}\right], \tag{5.1}$$

where $Z_i = (Z_{i1}, \ldots, Z_{iP})^{\mathrm{T}}$, $X_j(\cdot)$ and $\beta_j(\cdot)$ are the $j$th functional predictor and corresponding coefficient function, and $T \in [0, \tau]$ for some finite $\tau > 0$. The model (5.1) is determined by the unknown functions $\beta_j(\cdot)$ and $f_p(\cdot)$, and the cumulative baseline hazard function $H_0(\cdot)$. $G(\cdot)$ is an increasing transformation function, and $H_0(t)$ represents an unknown non-decreasing and non-negative baseline cumulative hazard function. Note that when $G(X) = X$, the above model reduces to the partially functional additive Cox model introduced in Chapter 4.

Under the data structure of the interval-censored and sparely sampled functional data, we will identify significant risk factors by using the partially functional additive transformation models with several group regularization methods. The principal analysis by conditional expectation (PACE) and Bernstein polynomials will be adopted to deal with unknown functions. A penalized EM algorithm will be employed for implementation.

## 5.2 Variable Selection for Functional Mixture or Nonmixture Cure Models with Interval-censored Data

A cure subgroup is often accompanied by failure time data. Some patients could cure from an event of interest, like heart disease. It is natural to employ cure models to accommodate the cure subgroup, such as mixture and non-mixture cure models. In the ADNI study, not all individuals would eventually convert to AD from MCI. Therefore, a cure rate exists among these patients. Furthermore, the availability and existence of several trajectories of clinical outcomes and a large number of genetic factors make the variable selection desirable and necessary to select key risk factors that significantly affect the risk of AD conversion. Developing a variable selection technique for functional cure models with interval-censored and sparsely sampled longitudinal data with a cure fraction is one direction of future re-

search.

For cure models, Sun *et al.* (2019) developed a variable selection method in the semi-parametric non-mixture or promotion time cure model when interval-censored data with a cured subgroup are present. Shi *et al.* (2022) proposed a functional proportional hazard cure rate model with functional and scalar predictors for sparsely measured longitudinal data and interval-censored failure time data. One can employ the functional principal component analysis (FPCA) to deal with the functional predictors, then the functional principal components (FPCs) derived from the FPCA can be viewed as new predictors. A penalized EM algorithm can be adopted to implement the cure model.

# Bibliography


Aneiros, G., Ferraty, F., and Vieu, P. (2015). Variable selection in partial linear regression with functional covariate. *Statistics*, **49**(6), 1322–1347.

Aneiros, G., Novo, S., and Vieu, P. (2022). Variable selection in functional regression models: A review. *Journal of Multivariate Analysis*, **188**, 104871.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *The Annals of Statistics*, **25**(2), 173–187.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2007). Partially linear hazard regression for multivariate survival data. *The Journal of the American Statistical Association*, **102**(478), 538–551.

Cao, Y., Huang, J., Liu, Y., and Zhao, X. (2016). Sieve estimation of cox models with latent structures. *Biometrics*, **72**, 1086–1097.

Carnicer, J. M. and Pena, J. M. (1993). Shape preserving representations and optimality of the bernstein basis. *Advances in Computational Mathematics*, **1**(2), 173–196.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3), 759–771.

Cheng, G. and Wang, X. (2011). Semiparameric additive transformation model under current status data. *Electronic Journal of Statistics*, **5**, 1735–1764.

Cox, D. R. (1972). Regression models and life-tables(with discussion). *Journal of the Royal Statistical Society, Series B*, **74**(529), 187–220.

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian journal of statistics*, **15**(1), 1–23.

Du, M., Zhao, X., and Sun, J. (2022). Variable selection for case-cohort studies with informatively interval-censored outcomes. *Computational Statistics & Data Analysis*, **172**, 107484.

Du, P., Ma, S., and Liang, H. (2010). Penalized variable selection procedure for cox models with semiparametric relative risk. *The Annals of Statistics*, **38**(4), 2092–2117.

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

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(5), 849–911.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *Inst Math Stat Collection*, **6**, 70–86.

Fang, H., Wu, T. T., Rapoport, A. P., and Tan, M. (2016). Survival analysis with functional covariates for partial follow-up studies. *Statistical Methods in Medical Research*, **25**(6), 2405–2419.

Gellar, J. E., Colantuoni, E., Needham, D. M., and Crainiceanu, C. M. (2015). Cox regression models with functional covariates for survival data. *Statistical modelling*, **15**(3), 256–278.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, **35**(1), 70–91.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, **34**(3), 1493–1517.

Hao, M., Liu, K.-Y., Xu, W., and Zhao, X. (2021). Semiparametric inference for the functional cox model. *Journal of the American Statistical Association*, **116**(535), 1319–1329.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia ii: A program for missing data. *Journal of statistical software*, **45**, 1–47.

Hu, H., Li, H., Li, J., Yu, J., and Tan, L. (2018). Genome-wide association study identified atp6v1h locus influencing cerebrospinal fluid bace activity. *BMC Med Genet*, **19**, 75.

Huang, J. (1999). Efficient estimation of the partly linear additive cox model. *The Annals of Statistics*, **27**(5), 1536–1563.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, **27**, 481–499.

Kong, D., Xue, K., Yao, F., and Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika*, **103**(1), 147–159.

Kong, D., Ibrahim, J. G., Lee, E., and Zhu, H. (2018). Flcrm: Functional linear cox regression model. *Biometrics*, **74**(1), 109–117.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, **32**(4), 1448–1491.

Lee, E., Zhu, H., Kong, D., Wang, Y., Giovanello, K. S., and Ibrahim, J. G. (2015). Bflcrm: A bayesian functional linear cox regression model for predicting time to conversion to alzheimer's disease. *The annals of applied statistics*, **9**(4), 2153.

Li, J., Yuan, X., Li, H., Cao, X., Yu, J., Tan, L., and Chen, W. (2018). Genome-wide association study identifies two loci influencing plasma neurofilament light levels. *BMC Med Genomics*, **11**(1), 47.

Li, K. and Luo, S. (2019). Dynamic prediction of alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, **38**(24), 4804–4818.

Li, K., Chan, W., Doody, R. S., Quinn, J., Luo, S., and Initiative, A. D. N. (2017). Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. *Journal of Alzheimer's Disease*, **58**, 361–371.

Li, S., Wu, Q., and Sun, J. (2020). Penalized estimation of semiparametric transformation models with interval-censored data and application to alzheimer's disease. *Statistical Methods in Medical Research*, **29**(8), 2151–2166.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika*, **100**(4), 859–876.

Lu, M. and McMahan, C. S. (2018). A partially linear proportional hazards model for current status data. *Biometrics*, **74**(4), 1240–1249.

Lv, S., Jiang, J., Zhou, F., Huang, J., and Lin, H. (2018). Estimating high-dimensional

additive cox model with time-dependent covariate processes. *Scandinavian Journal of Statistics*, **45**(4), 900–922.

Matsui, H. and Konishi, S. (2011). Variable selection for functional regression models via the regularization. *Computational Statistics amp; Data Analysis*, **55**(12), 3304–3310.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Qu, S., Wang, J.-L., and Wang, X. (2016). Optimal estimation for the functional cox model. *The Annals of Statistics*, **44**(4).

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Scolas, S., El Ghouch, A., Legrand, C., and Oulhaj, A. (2016). Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in Medicine*, **35**, 1210–1225.

Shi, H., Ma, D., Nie, Y., Beg, M. F., Pei, J., Cao, J., Initiative, A. D. N., *et al.* (2021). Early diagnosis of alzheimer's disease on adni data using novel longitudinal score based on functional principal component analysis. *Journal of Medical Imaging*, **8**(2), 024502.

Shi, H., Ma, D., Faisal Beg, M., and Cao, J. (2022). A functional proportional hazard cure rate model for interval-censored data. *Statistical Methods in Medical Research*, **31**, 154–168.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer New York.

Sun, L., Li, S., Wang, L., and Song, X. (2019). Variable selection in semiparametric nonmixture cure model with interval-censored failure time data: An application to the prostate cancer screening study. *Statistics in Medicine*, **38**(16), 3026–3039.

Sun, L., Li, S., Wang, L., Song, X., and Sui, X. (2021). Simultaneous variable selection in regression analysis of multivariate interval-censored data. *Biometrics*.

Tian, T. and Sun, J. (2022). Variable selection for nonparametric additive cox model with interval-censored data. *Biometrical Journal*.

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

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

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S., Saykin, A., Shen, L., and Initiative, A. D. N. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, **28**(2), 229–237.

Wang, J. and Ghosh, S. K. (2012). Shape restricted nonparametric regression with bernstein polynomials. *Computational Statistics & Data Analysis*, **56**(9), 2729–2741.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016a). Functional data analysis. *Annual Review of Statistics and its application*, **3**, 257–295.

Wang, L., McMahan, C. S., Hudgens, M. G., and Qureshi, Z. P. (2016b). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, **72**(1), 222–231.

Wu, Q., Zhao, H., Zhu, L., and Sun, J. (2020). Variable selection for high-dimensional partly linear additive cox model with application to alzheimer's disease. *Statistics in medicine*, **39**(23), 3120–3134.

Wu, Y. and Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of hla markers in psoriatic arthritis. *Biometrics*, **71**(3), 782–791.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, **25**(6), 1129–1141.

Yao, F., Müller, H.-G., and Wang, J. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**(470), 577–590.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, **33**(6), 2873–2903.

Ye, J., Li, Y., and Guan, Y. (2015). Joint modeling of longitudinal drug using pattern and time to first relapse in cocaine dependence treatment data. *The Annals of Applied Statistics*, **9**(3), 1621–1642.

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

Zeng, D., Mao, L., and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, **103**(2), 253–271.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhang, H. and Lu, W. B. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, **94**, 1–13.

Zhao, H., Wu, Q., Li, G., and Sun, J. (2020). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*, **115**(529), 204–216.

Zhou, Q., Hu, T., and Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, **112**, 664–672.

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

# VITA

Tian Tian was born in Xuzhou, Jiangsu Province, China. She received her Bachelor's degree in Applied Mathematics from Nanjing Xiaozhuang University, China, in 2015. Then she obtained her Master of Arts degree from the Department of Statistics at the University of Missouri-Columbia in 2017 and then continued to pursue the Ph.D. program in statistics at the University of Missouri-Columbia in August 2017.