2018-01-17

Estimation and Group Selection in Partially Linear Survival Models

Afzal, Arfan

UNIVERSITY OF CALGARY

Estimation and Group Selection in Partially Linear Survival Models

by

Arfan Raheen Afzal

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

January, 2018

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Estimation and Group Selection in Partially Linear Survival Models" submitted by Arfan Raheen Afzal in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Supervisor, Dr. Xuewen Lu
Department of Mathematics and Statistics
University of Calgary

Supervisory Committee Member,
Dr. Rohana Ambagaspitiya
Department of Mathematics and Statistics
University of Calgary

Supervisory Committee Member,
Dr. Hua Shen
Department of Mathematics and Statistics
University of Calgary

Internal/Examiner, Dr. Rob Deardon
Department of Production Animal Health
Department of Mathematics and Statistics
University of Calgary

External Examiner, Dr. Yichun Zhao
Department of Mathematics and Statistics
Georgia State University

Date

Abstract

In survival analysis, different regression models are available to estimate the effects of covariates on the censored survival outcome. The proportional hazards (PH) model has been the most popular model among them because of its simplicity and desirable theoretical properties. However, the PH model assumes that the hazard ratio is constant over observed time. When this assumption is not met or we are interested in the risk difference, the additive hazards (AH) model is a useful alternative. On the other hand, assuming linear structure of covariate effects on survival in these models may be too strict. As a remedy to that, partially linear survival models are getting increasingly popular as it combines the flexibility of nonparametric modeling with the parsimony and easy interpretability of parametric modeling. Nonetheless, building these models becomes a challenging problem when predictors or covariates are high-dimensional and grouped. Consequently, it becomes crucial to select important groups and important individual variables within groups by the so called bi-level variable selection method to reduce the dimension of the data and build a sensible and useful semiparametric model for applications as the methods for individual variable selection in such cases may perform inefficiently by ignoring the information present in the grouping structure.

To fill gaps in estimation and group selection in partially linear survival models with high-dimensional data, in this thesis, we propose new methods for estimation and group selection in two partially linear survival models, namely, the partially linear AH model and the partially linear PH model.

In the first part of this thesis, we consider estimation in a partially linear AH model with left-truncated and right-censored data when the dimension of covariates is fixed and the risk function has a partially linear structure. We construct a pseudo-score function to estimate the coefficients of the linear covariates and the B-spline basis functions. The

proposed estimators are asymptotically normal under the assumption that the true nonlinear functions are B-spline functions whose knot locations and the number of knots are held fixed.

In the second and third parts, we study group variable selection in the partially linear AH model and the partially linear PH model with right censored data. In such regression models with a grouping structure among the explanatory variables, variable selection at the group and within group individual variable level is important to improve model accuracy and interpretability. Motivated by the hierarchical grouped variable selection in the linear PH model and the linear AH model, we propose a hierarchical bi-level variable selection approach for high-dimensional covariates in the linear part of the partially linear AH model and the partially linear PH model, respectively. The proposed methods are capable of conducting simultaneous group selection and individual variable selection within groups in the presence of nonparametric risk functions of low-dimensional covariates. For group selection in the partially linear AH model, the rates of convergence and selection consistency of the proposed estimators are established using martingale and empirical process theory; after reducing the dimension of the covariates, we suggest the use of the method in the first part for inference in the partially linear AH model. For group selection in the partially linear PH model, similar theoretical results of the proposed estimators are obtained, and the oracle properties such as asymptotic normality of the estimators are discussed.

Finally, computational algorithms and programs are developed for utilizing the proposed methods. Simulation studies indicate good finite sample performance of the methods. For each model, real data examples are provided to illustrate the application of the methods.

Acknowledgements

I would like to express my profound gratitude to my supervisor, Dr. Xuewen Lu for his invaluable support, continuous guidance, infinite patience, and constant encouragement. Dr. Lu's intelligence and deep insight has always fascinated me and inspired me to critically think about a problem while his kind-heartedness gave me the strength when I faced challenges. He was always there for me under any circumstances, beyond any situations. I am extremely fortunate to have Dr. Lu as my supervisor. I admire him the most and he is my academic, research and personality role model.

I would also like to thank my committee members, Drs. Rohana Ambagaspitiya, Rob Deardon, Hua Shen, and Yichun Zhao for inputting their valuable time and insightful comments on my research which will benefit my future research. In addition, I would like to thank all of my professors in the Department of Mathematics and Statistics for their advice, inspiration, and for providing me with an excellent education. I am also thankful to all of the staffs for their various support and my sincere gratitude to all of my friends for their friendship and emotional support. Thanks to Shomoita for helping me with Westgrid use. Most importantly, I am thankful to the Department for providing me with the financial support in my program which helped me to keep my focus on my research. I am also grateful to the Faculty of Nursing for their encouragement and support in finishing my research.

Last but not the least, my heartfelt thanks and love goes to my beloved family, my mom and my siblings for their unconditional love and confidence in me through all these years. Special thanks to my sister Shima who would make time to talk to me every week and was always there for me whenever I needed emotional support.

Table of Contents

List of Tables

List of Figures and Illustrations

Chapter 1

Introduction

1.1 Survival Analysis and Semiparametric Models

Survival analysis is an umbrella term to denote a wide variety of statistical methods to describe, explain, or predict the timing and duration until the occurrence of an event (Kleinbaum, 1998). Originally, the name "survival analysis" was developed by the biostatisticians to analyze the occurrence of deaths in medical science, however, these methods are also applied in a number of areas including social science, engineering, economics, actuarial science, etc. In medical science, examples of survival times can be time to death, time until tumor recurrence; in social science, examples include time to change jobs, divorce; in engineering, survival data comes from studying the life of a machine; in actuarial science, it is time to life insurance claim, and so on. Different names have been used to refer survival analysis due to its adaptation in other research fields, namely, time to event analysis, failure time analysis, duration analysis, reliability analysis, however, survival analysis remains the most widely used and recognized name (Lee and Wang, 2003). One unique feature of survival data is the presence of censoring, that is, survival data are not fully observed in all subjects. For example, in a medical science study, censoring can occur if a subject chooses to quit participating in the study, moves away from the study area and cannot be followed anymore, or die from some unrelated event. On the other hand, truncation in the survival data occurs when only part of the population is included and observed in the study and nothing at all is known about the unobserved part. For example, in actuarial science, if an automobile physical damage policy has a deductible of \$500 per claim, then any losses below \$500 will not appear in the data sets. While censoring data is from the whole population, truncated data is from a subpopulation who experience some event that is not of our interest.

We introduce the counting process notations here which will be used in our research for estimation and establishing theoretical properties.

1.1.1 Counting Processes and Martingales

In survival analysis our interest is in the time to occur a specific event. This data can be represented as the standard counting process notation, $N(t)$, which is simply a random function of time t and counts the number of events observed in a time interval $[0, t]$. It is zero at time zero and constant over time except it jumps at each time point where an event occurs, with jumps of size 1. The following definitions will be used in our study:

Definition 1. *(Counting Process) A counting process is a stochastic process* $\{N(t), t \geq 0\}$ *with jumps of size one and values which are positive integers and increasing, such that*
(1) $N(t) \geq 0$,
(2) $N(t)$ *is an integer,*
(3) If $s \leq t$ *then* $N(s) \leq N(t)$.

Definition 2. *(Uniformly Integrable) A stochastic process* $X(t)$ *is called uniformly integrable (UI) if there exists a* $K \in [0, \infty)$ *such that* $\sup_{t \in [0,\infty)} E\left(|X(t)| I(|X(t)| \geq K)\right) \leq \infty$, *where* $I(\cdot)$ *is an indicator function.*

Theorem 1. *(Doob-Meyer) Let* N *be a submartingale of class D with* $N(0) = 0$. *Then there exists a unique, increasing, integrable, predictable process* A *with* $A(0) = 0$ *such that* $M(t) = N(t) - A(t)$ *is a uniformly integrable martingale.*

Because of the third condition of the counting process in Definition 1, a counting process is increasing and hence, a submartingale. Thus, by Theorem 1, it can be decomposed into two parts as follows

$$N(t) = A(t) + M(t),$$

where $M(t)$ is the martingale associated with the counting process $N(t)$ and $A(t)$ is a predictable increasing process. $A(t)$ is known as the compensator or the cumulative intensity of $N(t)$. Such a representation of semiparametric survival models is widely used because central limit theorem (CLT) is available for martingales and makes it possible to derive large sample properties of the estimators (Theorem 2).

We will use predictable variation of martingale for computing the variance of the counting process martingale $M(t) = N(t) - A(t)$, and the variance of integrals with respect to $M(t)$. Also, the martingale CLT conditions are formulated using predictable variation process.

We assume that only one event can occur at a given point of time, so we are only dealing with untied observations. The behavior of $N(t)$ is controlled by its intensity process, $f_N(t)$. The intensity process is given as $f_N(t)dt = P(\text{event occurs in } [t, t+dt] \mid \mathcal{F}_{t-})$, where $\mathcal{F}_{t-}$ is a filtration representing all the available information just before time t. Based on Aalen et al. (2008) the intensity process of $N(t)$ is assumed to take the following form

$$f_N(t) = \nu(t)Y(t),$$

where $\nu(t)$ is a nonnegative function indicating the hazard rate and $Y(t)$ is an observable process that indicates the number of at risk individuals just before time t.

Definition 3. *(Predictable Variation of a Martingale) Let M be a right-continuous martingale with respect to a right-continuous filtration $\{\mathcal{F}_t : t \geq 0\}$ and assume $E[M^2(t)] < \infty$ for any $t \geq 0$. Then there exists a unique increasing right-continuous predictable process $\langle M, M\rangle = \langle M\rangle$, called the predictable variation of M, which is defined as*

$$\langle M\rangle(t) = \int_0^t E[\{dM(u)\}^2 | \mathcal{F}_{u^-}].$$

It can be shown that $\langle M\rangle(0) = 0$ a.s., $E[\langle M\rangle(t)] < \infty$ for each t, and $\{M^2(t) - \langle M\rangle(t) : t \geq 0\}$ is a right-continuous martingale.

Theorem 2. *The predictable variation process of the integration with respect to the martingale $M(t)$, assuming that $H(t)$ is a predictable process, is given by*

$$\langle \int_0^t H(s)dM(s) \rangle = \int_0^t H^2(s)d\langle M(s) \rangle.$$

In other words, the integrand $H(s)$ acts like a constant after conditioning. This is due to the predictability of $H(t)$.

For more details about counting processes and martingales, see Hall and Heyde (1980) and Fleming and Harrington (2011).

Next we outline the main survival models on which our proposed research is built on.

1.1.2 Cox Proportional Hazards Model

The proportional hazards model (PHM) proposed by Cox (1972) is probably the oldest and the most prominent regression model used in survival analysis to study the association between the survival time and risk factors. It estimates the relative risk of experiencing an event of interest between two groups of subjects. To describe the PHM, let T_i represent the survival time for the i^{th} subject ($i = 1, \ldots, n$), X_i be the associated d-dimensional vector of covariates, and C_i denote the censoring time for the i^{th} subject. Let $Z_i = \min(T_i, C_i)$ be the observed time and δ_i denote the event indicator, i.e., $\delta_i = I(T_i \leq C_i)$, which takes value 1 if the event occurs, or 0 if the event time is censored. The data, therefore, consist of n observations $(Z_i, \delta_i, X_i),\ i = 1, \ldots, n$, which are assumed to be an independent and identically distributed (i.i.d.) sample from (Z, δ, X). The conditional hazard function for a subject with a d-dimensional covariate vector X under the PHM is given as

$$h(t|X) = h_0(t) \exp(\beta^\top X), \tag{1.1}$$

where β is a d-dimensional regression parameter vector and $h_0(t)$ is the baseline hazard, which is usually left unspecified.

One of the main reasons for the popularity of the PHM is the estimation of the regression parameters does not depend on the unspecified baseline hazard function $h_0(t)$. Cox (1975) proposed this partial likelihood estimates of β which is a technique developed to make inference about the regression parameters in the presence of nuisance parameters.

Let $N_i(t) = 1\{T_i \leq t,\ \delta = 1\}$ be the right continuous counting process and $Y_i(t) = 1\{T_i \geq t\}$ be the left continuous at-risk process for the ith individual. We only consider events over a finite time interval $[0, \tau]$. Then the partial likelihood in counting process notation is written as

$$L(\beta) = \prod_{t \geq 0} \prod_{i=1}^{n} \left(\frac{Y_i(t) \exp\left(\beta^\top X_i\right)}{\sum_{j=1}^{n} Y_j(t) \exp\left(\beta^\top X_j\right)} \right)^{dN_i(t)},$$

where $dN_i(t)$ is the increment of $N_i(t)$ over a small interval dt around time t. Consequently, the partial log-likelihood is given as

$$l(\beta) = \sum_{i=1}^{n} \int_0^\tau \left\{ \beta^\top X_i - \log \sum_{j=1}^{n} Y_j(t) \exp\left(\beta^\top X_j\right) \right\} dN_i(t).$$

The estimates are obtained by maximizing $l(\beta)$ which is done by solving the set of d nonlinear equations $U_h(\beta) = 0,\ h = 1, \ldots, d$, where

$$U_h(\beta) = \frac{\partial l(\beta)}{\partial \beta_h} = \sum_{i=1}^{n} \int_0^\tau \left\{ X_{ih} - \frac{\sum_{j=1}^{n} Y_j(t) X_{jh} \exp\left(\beta^\top X_j\right)}{\sum_{j=1}^{n} Y_j(t) \exp\left(\beta^\top X_j\right)} \right\} dN_i(t).$$

The resulting maximum partial likelihood estimators possess asymptotic properties similar to those of the standard maximum likelihood estimator (Tsiatis, 1981; Andersen and Gill, 1982). Such desirable theoretical properties, simple interpretation of the results and extensively available computer programs, have made the PHM the default method of choice in survival analysis (Lin and Ying, 1994).

1.1.3 Additive Hazards Model

The PHM has one important assumption, the proportional hazards assumption, which assumes the hazards ratio is constant over the observed survival times. However, when this assumption

is not met or when we are interested in the absolute hazards change instead of hazards ratio, the additive hazards model (AHM) serves as a useful alternative to the PHM. Aalen et al. (2008) in their book pointed out a number of reasons why using the AHM is advantageous sometimes. Rothman (2012) focused on the importance of additive models when evaluating independent risk factors. As mentioned in Aalen (1989), the PHM is sensitive to adding or removing covariates from the model while the AHM handles this situation better due to its linear nature. We can consider the hazard of an event as the sum of the baseline hazard and an excess hazard. When the interest is to study how excess hazard depends on covariates, the AHM is specifically useful as the excess hazard may easily become negative, where the PHM may not be very practical (Hall and Müller, 2003). Tang and Dickinson (1998) referred that, relative risk or hazards ratio as a measure of actual effect of a risk factor could be misleading sometime, while the difference measure may be better in many circumstances, especially, useful for comparing groups with different baseline rates.

The hazard function for the failure time T associated with a d-vector of covariates X in the AHM (Cox and Oakes, 1984; Thomas, 1986; Breslow and Day, 1987) takes the form

$$h(t|X) = h_0(t) + \beta^\top X. \tag{1.2}$$

The AHM in various forms have been advocated and successfully utilized by many authors (Aalen, 1980; Pocock et al., 1982; Buckley, 1984; Pierce and Preston, 1984; Thomas, 1986; Breslow and Day, 1987; Aalen, 1989; Huffer and McKeague, 1991), however, none were able to directly use the partial likelihood approach to eliminate the nuisance baseline function $h_0(t)$ when estimating β.

Lin and Ying (1994) proposed the pseudoscore function to estimate β in the AHM that mimics the martingale feature of the partial likelihood score function for the PHM where the estimating equation does not involve $h_0(t)$,

$$U(\beta) = \sum_{i=1}^{n} \int_0^\tau \left\{ Xi - \bar{X}(t) \right\} \left\{ dN_i(t) - Y_i(t)\beta^\top X_i dt \right\},$$

where $\bar{X}(t) = \sum_{i=1}^{n} Y_i(t)X_i / \sum_{i=1}^{n} Y_i(t)$. The resulting estimator takes the explicit form

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^{\tau} Y_i(t) \left\{X_i - \bar{X}(t)\right\}^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\tau} \left\{X_i - \bar{X}(t)\right\} dN_i(t)\right],$$

where $a^{\otimes 2} = aa^{\top}$.

The pseudoscore method has been frequently used by researchers because of the closed form solution of the estimators, where the estimators are consistent and asymptotically normal with an easily estimated covariance matrix.

In our research, we have extended the linear PHM and linear AHM by incorporating covariates that can have a nonlinear effect on the survival probability and called them partially linear proportional hazards model (PL-PHM) and partially linear additive hazards model (PL-AHM). In the next section, we illustrate the concept of partially linear models.

1.2 Partially Linear Models

In real data, all covariates might not necessarily be linearly related with the response, some of them might have a nonlinear relationship. Partially linear models (PLMs), which include both linear and nonlinear components in the model, are flexible extension of linear models and have been systematically studied in recent years (Härdle et al., 2012). A PLM usually takes the form,

$$Y = \alpha + \beta^{\top} X + \sum_{q=1}^{Q} \phi_q(W_q) + \varepsilon,$$

where Y is a response variable and $W = (W_1, \ldots, W_Q)$ is a Q dimensional vector of covariates, α is the intercept, ϕ_q's are unknown smooth functions with zero means, i.e., $\mathrm{E}\phi_q(W_q) = 0$, and ε is the random error term with mean zero and a finite variance σ^2.

Using the same idea, the PL-PHM is obtained by extending model (1.1) as follows:

$$h(t|X, W) = h_0(t) \exp\left\{\beta^{\top} X + \sum_{q=1}^{Q} \phi_q(W_q)\right\}. \tag{1.3}$$

Similarly, the PL-AHM which is an extension of model (1.2), takes the form:

$$h(t|X,W) = h_0(t) + \beta^\top X + \sum_{q=1}^{Q} \phi_q(W_q). \tag{1.4}$$

Model (1.3) and model (1.4) now contain both the linear component $\beta^\top X$ and the nonlinear components $\phi_q(W_q);\ q = 1, \ldots, Q$. These models extend a purely linear model given in (1.1) and (1.2) and avoids the curse of dimensionality of a purely nonparametric model.

1.2.1 B-Splines

To approximate the nonparametric functions in a PLM, B-splines have been extensively used by many researchers. B-splines are well-known for their ability to provide good approximations to smooth functions (De Boor, 1978; Schumaker, 1981), they are numerically stable, computationally faster, and has broad application in nonparametric smoothing (Stone et al., 1997).

To approximate an unknown function of variable x on the interval $[a, b]$, a B-spline is defined as a piecewise polynomial function of degree k in the domain of the variable. Let $a = u_0 \leq u_1 \leq \cdots \leq u_{m+1} = b$ be a sequence of $m + 2$ ascending real numbers which are considered as a subdivision of $m + 2$ distinct points on the interval $[a, b]$ on which the x variable is valued. These points are called knots or break points. The interval $[u_i, u_{i+1})$ is considered as the i-th knot span for $i = 0, 1, \ldots, m$. The significance of B-splines is in the fact that any spline function of degree k on a given set of knots, $U = \{u_0, u_1, \ldots, u_{m+1}\}$, can be expressed as a linear combination of its basis functions. Here, we use normalized B-splines where the i-th B-spline basis function of degree k, is written as $N_i^k(x)$ and defined recursively as follows.

Definition 4. *(Cox-de Boor Recursion Formula) For all $x \in [a, b]$,*

$$N_i^0(x) = \begin{cases} 1 & \text{if } u_i \leq x < u_{i+1} \\ 0 & \text{otherwise,} \end{cases}$$

$$N_i^k(x) = \frac{x - u_i}{u_{i+k} - u_i} N_i^{k-1}(x) + \frac{u_{i+k+1} - x}{u_{i+k+1} - u_{i+1}} N_{i+1}^{k-1}(x),$$

where $1 \leq k \leq m$. *There is a special case of* B*-splines when the knots are equally spaced, and they are usually known as uniform* B*-splines.*

B-splines have some convenient properties. A B-spline basis function of degree k has $k+1$ polynomials of degree k on $k+1$ intervals. Outside these $k+1$ intervals, the basis function is zero, which makes the basis functions local. The derivative of a B-spline of degree k is a B-spline of degree $k-1$. In general, the i^{th} k degree B-spline is nonzero only on the interval $[u_i, u_{i+k+1}]$. This property ensures that the i^{th} and $(i+j+1)^{\text{th}}$ B-splines are orthogonal for $j \geq k$. B-splines whose supports overlap are linearly independent. For each fixed sequence of interior knots $\{u_1, \ldots, u_m\}$, the set of such splines is a linear space of functions with $m + (k+1)$ (= number of interior knots + order of B-spline basis functions) free parameters (De Boor, 1978).

Another important property of splines is their smoothness. k^{th} degree splines usually have no more than $k-1$ continuous derivatives. For example, cubic splines often have two continuous derivatives. Two continuous derivatives are often sufficient to provide smooth approximations to the functions, one of the main reasons for the popularity of B-splines. In addition, third degree piecewise polynomials are usually numerically well-behaved. As shown in Gray (1992) and Cheng and Wang (2011), it is adequate to choose less than 10 knots for sufficiently smooth approximation.

1.3 Variable Selection

Variable selection is fundamental in high-dimensional statistical modeling when the number of covariates is large and grows with the sample size in analyzing data. Generally, a large number of potential predictors are measured to avoid missing an important one, however,

often a smaller number of important variables is desirable to improve statistical efficiency and model interpretability. In addition, many of the traditional variable selection methods become undesirable as a result of high computational cost and lack of stability in high-dimensional data due to noise accumulation and excessive predictors. Breiman (1996), for example, pointed out that best-subset selection can be unstable leading to poor prediction performance of the model.

To simultaneously estimate and select important variables, a family of penalized approaches is proposed. Variable selection is performed by minimizing a penalized objective function by adding a penalty function of the following form

$$\min \{\text{Loss function} + \text{Penalty}\}.$$

Popular choices of loss functions are least squares and negative log-likelihood. Plenty of different penalty functions have been used for penalized regression, such as the bridge estimator (Frank and Friedman, 1993; Fu, 1998); least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), adaptive elastic net (Zou and Zhang, 2009), the smooth integration of counting and absolute deviation (SICA) (Lv and Fan, 2009), and the minimum concave penalty (MCP) (Zhang, 2010), among others. All of these methods are designed to conduct individual variable selection.

In many applications, covariates are naturally grouped. For example, in factor analysis-of-variance (ANOVA), a factor that has several levels can be expressed through several dummy variables where the dummy variables form a natural group, i.e., for response Y with two factors α and β, intercept μ and random error ε,

$$Y = \mu + \alpha_j + \beta_k + \varepsilon, \qquad j = 1, \ldots, J, \;\; k = 1, \ldots, K,$$

where $\{\alpha_j\}_{j=1}^{J}$ and $\{\beta_k\}_{k=1}^{K}$ can be considered as two groups. Similarly, in additive models with polynomial or nonparametric components, each component may be expressed as a linear

combination of a number of basis functions of the original measured variable, which also form a natural group; for example,

$$Y = \mu + \varphi_1(W_1) + \cdots + \varphi_J(W_J) + \varepsilon,$$

where the $\varphi_j(W_j) = \sum_{\ell=1}^{m} \gamma_{\ell j} B_\ell(W_j)$, here $\{B_\ell(W_j)\}_\ell^m$ are basis functions and can be considered as a group.

Therefore, it would be reasonable to select groups of related covariates rather than individual variables in the above situations. Common group variable selection methods that are available in the literature include the group LASSO (Yuan and Lin, 2006), group SCAD (Wang et al., 2007), group bridge (Huang et al., 2009) and group MCP (Breheny and Huang, 2009) penalties.

1.3.1 Penalty Functions

Many penalty functions are available in the literature for variable selection. Fan and Li (2001) advocated penalty functions that give estimators with three properties:

1. *Sparsity*: The coefficients of insignificant variables should be estimated as zero. This achieves the purpose of the variable selection.

2. *Continuity*: The estimated coefficients should be continuous in data to enhance the model stability. This avoids unnecessary variation in the prediction.

3. *Unbiasedness*: When the true coefficients are large, they should be estimated asymptotically unbiased. This avoids unnecessary biases in the model selection steps.

We introduce the following notations first:

$$L_0 \text{ norm} : \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{d} I(|\beta_j| > 0),$$
$$L_1 \text{ norm} : \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{d} |\beta_j|,$$
$$L_2 \text{ norm} : \|\boldsymbol{\beta}\|_2 = (\sum_{j=1}^{d} |\beta_j|^2)^{1/2},$$
$$L_\gamma \text{ norm} : \|\boldsymbol{\beta}\|_\gamma = (\sum_{j=1}^{d} |\beta_j|^\gamma)^{1/\gamma}.$$

Some frequently used penalty functions are listed below where $\lambda > 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are the tuning parameters. Note that, (1.5)-(1.9) perform individual variable selection and (1.10)-(1.13) perform group variable selection.

1.3.2 The LASSO Penalty

The LASSO (Tibshirani, 1996) uses the L_1 penalty,

$$p_\lambda(|\beta_j|) = \lambda |\beta_j|. \tag{1.5}$$

Here, L_1 penalty is not differentiable at zero, and therefore, some of estimates will be obtained as exactly zero. Leng et al. (2006) showed that LASSO, in general, is not variable selection consistent.

1.3.3 The Bridge Penalty

The bridge penalty (Frank and Friedman, 1993; Fu, 1998) is defined by

$$p_\lambda(|\beta_j|) = \lambda |\beta_j|^\gamma, \tag{1.6}$$

where $0 \leq \gamma \leq 2$ is the bridge index. It includes LASSO with $\gamma = 1$, ridge regression with $\gamma = 2$ and subset selection with $\gamma = 0$ as special cases.

Frank and Friedman (1993) proposed a family of bridge penalty with $\gamma > 0$. Notice that $0 < \gamma < 1$ corresponds to a class of concave penalties, while $1 \leq \gamma \leq 2$ corresponds to a class

of convex penalties. Thus, for a convex loss function, the bridge penalty has the variable selection feature when $0 < \gamma < 1$.

1.3.4 The Elastic Net Penalty

The elastic net penalty (Zou and Hastie, 2005) has the form

$$p_\lambda(|\beta_j|) = \lambda_1\beta_j^2 + \lambda_2 |\beta_j| = (\lambda_1 + \lambda_2) \left\{\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right\}, \tag{1.7}$$

where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. The Elastic Net penalty can capture grouping effect, where strongly correlated predictors tend to be all-in or all-out of the model together. When the number of predictors is much bigger than the number of observations ($d >> n$), this penalty is particularly useful.

1.3.5 The SCAD Penalty

The SCAD penalty (Fan and Li, 2001) was proposed to reduce the bias caused by the L_1 penalty (LASSO). The SCAD penalty and its first derivative are defined as

$$p_{\lambda,a}(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if} \quad |\beta_j| \le \lambda, \\ -\frac{\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & \text{if} \quad \lambda < |\beta_j| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if} \quad |\beta_j| > a\lambda, \end{cases} \tag{1.8}$$

$$p'_{\lambda,a}(|\beta_j|) = \lambda \left\{ I(|\beta_j| \le \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda) \right\},$$

where $a > 2$ and $(s)_+ = s$ for $s > 0$ and 0 otherwise. This penalty function takes off at the origin as L_1 penalty and then gradually levels off until its derivative reaches zero.

1.3.6 The MCP Penalty

The MCP (Zhang, 2010)and its first derivative are defined by

$$p_{\lambda,a}(|\beta_j|) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2a}, & \text{if} \quad |\beta_j| \le a\lambda, \\ \frac{1}{2}a\lambda^2, & \text{if} \quad |\beta_j| > a\lambda, \end{cases} \tag{1.9}$$

$$p'_{\lambda,a}(|\beta_j|) = \lambda \frac{(a\lambda - |\beta_j|)_+}{a\lambda},$$

where $a > 1$ is a shape parameter. The MCP can handle situations where the number of covariates is greater than the number of observations ($d > n$).

1.3.7 The Group LASSO Penalty

Let $X_k = (X_{1k}, \ldots, X_{nk})^\top$, $k = 1, \ldots, d$, be the design vectors and $Y = (Y_1, \ldots, Y_n)^\top$ be the response vector, then a multiple linear regression model can be written as

$$Y = \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the error vector. Let $A_1, \ldots, A_J$ be subsets of $\{1, \ldots, d\}$ representing known groupings of the design vectors and denote the regression coefficients in the j^{th} group by $\beta_{A_j} = (\beta_k, k \subseteq A_j)^\top$.

The group LASSO penalty, proposed by Yuan and Lin (2006) is given as

$$\lambda \sum_{j=1}^{J} \|\beta_{A_j}\|_{K_j,2}, \tag{1.10}$$

where $\lambda > 0$ is the tuning parameter and K_j is a positive definite matrix and $\|\beta_{A_j}\|_{K_j,2} = (\beta_{A_j}^\top K_j \beta_{A_j})^{1/2}$. Yuan and Lin (2006) suggested the choice of K_j is $K_j = |A_j| I_j$ with I_j being the $|A_j| \times |A_j|$ identity matrix, where $|A_j|$ is the cardinality of A_j.

The group LASSO is a natural extension of the LASSO, and tends to select a larger model than the true model by including unimportant variables. Group LASSO penalty selects groups of important variables, however, it can not select individual variables within groups (Huang et al., 2009).

1.3.8 The Group Bridge Penalty

To conduct variable selection at the group and individual variable levels simultaneously, Huang et al. (2009) proposed the following group bridge penalty

$$\lambda \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^\gamma, \tag{1.11}$$

where c_j's are constants for the adjustment of the different dimensions of A_j. Huang et al. (2009) suggested a simple choice of c_j is $c_j \propto |A_j|^{1-\gamma}$ for uncensored data.

The group bridge approach has been shown to have the oracle group selection property, that is, it can correctly select important groups with probability converging to one.

1.3.9 The Group SCAD Penalty

Wang et al. (2007) introduced the the group SCAD penalty as

$$\sum_{j=1}^{J} p_{\lambda,a}(\|\boldsymbol{\beta}_{A_j}\|_2), \tag{1.12}$$

where $p_{\lambda,a}(\cdot)$ is the SCAD penalty, defined in (1.8), and $\|\boldsymbol{\beta}_{A_j}\|_2 = \sqrt{\sum_{k\in A_j} \beta_k^2}$.

1.3.10 The Group MCP Penalty

The group MCP developed by Breheny and Huang (2009) is defined as

$$\sum_{j=1}^{J} p_{\lambda,b}\left(\sum_{k\in A_j} p_{\lambda,a}(|\boldsymbol{\beta}_k|)\right), \tag{1.13}$$

where $p(\cdot)$ is the MCP penalty given by (1.9).

Huang et al. (2012) showed that the group MCP can perform bi-level selection, i.e., the penalty is capable of selecting important groups as well as important variables within selected groups.

1.4 Summary

The PLMs are an extension of linear regression models and additive nonparametric regression models which combine both, and are more flexible than parametric models and more efficient than nonparametric models. Estimation and inference for the PLMs is well studied (Ma and Yang, 2011; Härdle et al., 2012). Additionally, several authors have considered variable selection in the linear part of a PLM. They were motivated by the genetic epidemiology studies where data are usually collected on high-dimensional genomic measurements as well

as low-dimensional clinical covariates as combining both of them in disease prognosis gives better sensitivity and specificity. Such improvement has been observed in some disease studies (Rosenwald et al., 2002; Pittman et al., 2004). More examples of such studies can be found in Ma and Huang (2007). Therefore, as referred in Ma and Du (2012), two distinct sets of covariates are measured in these studies; X representing high-dimensional genomic measurements such as gene expressions or SNPs, and W denoting low-dimensional clinical and environmental risk factors such as age, gender, blood pressure etc. The high-dimensional X is typically modeled in a parametric way for better interpretability where the interest lies in identifying a small subset that is associated with the disease prediction. On the other hand, the low-dimensional W is handled in a more flexible nonparametric way as many biological processes are nonlinear.

Our research is motivated by four biological data sets. Our first data set is the South Wales nickel refiners study data set (Breslow and Day, 1987, Appendix D). Here we estimated the risk of developing carcinoma of the bronchi and nasal sinuses associated with different covariates. The three other data sets are the primary biliary cirrhosis (PBC) data (Fleming and Harrington, 2011), Wisconsin prognostic breast cancer (WPBC) data (available from the UCI machine learning repository) and mantle cell lymphoma (MCL) data (Rajabi and Sweetenham, 2015). In these data sets, we were interested in determining which groups and within-group individual variables have a significant effect on the survival time.

As we introduced at the beginning of this chapter, the most popular model used in survival analysis is the PHM, which examines the covariate effects on the hazard function and requires meeting the proportional hazards assumption. Another alternative method for the analysis of survival data is the AHM that does not need this assumption and can easily be implemented and the results interpreted. Wang et al. (2009) and Huang et al. (2014) investigated group selection in the linear PHM, where, Liu et al. (2014) investigated group selection in the linear AHM. However, no group variable selection is investigated in the PL-AHM and PL-PHM. Our

research fills in this gap and investigates the method, procedure, properties and application of group selection in the PL-AHM and PL-PHM.

The rest of the thesis is organized as follows. We described the estimation procedure in the PL-AHM in Chapter 2. A hierarchical method for group selection in the PL-AHM is reported in Chapter 3. In Chapter 4, we developed a hierarchical method for group selection for the PL-PHM. Summary and future research directions are reported in Chapter 5. Chapters 2, 3, and 4 are written in manuscript style. Chapter 2 has been published in *Statistical Modeling*. We are preparing two manuscripts from Chapters 3 and 4 for submission to statistical journals for publication.

Chapter 2

Estimation of Partly Linear Additive Hazards Model with Left Truncated and Right Censored Data

2.1 Introduction

In survival analysis, the multiplicative risk model (Cox, 1972) and the additive risk model (Aalen, 1980; Lin and Ying, 1994) provide two principal frameworks for the regression analysis of censored survival data where the former estimates the risk ratio and later estimates the risk difference. The multiplicative risk model which is also called the Cox proportional hazards model (PHM), has so far been the most popular model for studying the association between risk factors and survival times. On the other hand, in contrast to the Cox PHM, the additive risk or additive hazards model (AHM) describes a different aspect of the association between covariates and failure times, and provides a useful alternative to the multiplicative risk model or PHM. Buckley (1984) pointed out that the AHM is biologically more plausible than the PHM, while O'neill (1986) found that the use of the PHM may result in serious bias when the true model is additive. Aalen et al. (2008) in their book listed a number of reasons justifying the use of the AHMs. Particularly, the AHM considered by Lin and Ying (1994) has drawn much attention in research in analyzing left truncated and right censored data, due to its easy interpretation and implementation. Their AHM has the following form:

$$\lambda(t;\mathbf{X}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{X}(t), \tag{2.1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^\top$ is a p dimensional possibly time varying covariate vector. Mimicking the estimation procedure for the Cox PHM, for left truncated and right censored data, Lin and Ying (1994) developed a conditional estimating function for $\boldsymbol{\beta}$, which does not contain the nuisance parameter

$\lambda_0(t)$. The AHM has been also applied to other types of censored data. For example, Lin et al. (1998) proposed an estimation method for current status data under the AHM. Later, Martinussen and Scheike (2002) and Lu and Song (2012) used a semiparametric efficient score function to improve estimation efficiency, at the cost of estimating $\lambda_0(t)$ separately. Some of the recent literature that considered data arising from an AHM include bivariate current status data (Tong et al., 2012), current status data with auxiliary covariates (Feng et al., 2015), informative current status data (Zhao et al., 2015), clustered interval censored data (Li et al., 2012), gap time data of recurrent events with multiple causes (Sankaran and Anisha, 2012), left truncated and right censored data (Huang and Qin, 2013), right censored data with missing covariates (Hao et al., 2014), right censored data with missing censoring indicator (Qiu et al., 2015), left truncated and case I interval censored data (Wang et al., 2015), right censored data with instrumental variable (Li et al., 2015) and error contaminated survival data with replicate measurements (Yan and Yi, 2016), among others.

All the works aforementioned concern the linear AHM defined in (2.1), which cannot handle nonlinear or nonparametric covariate effects. On the other hand, partially linear models (PLMs) are getting increasingly popular due to the fact that it combines the flexibility of nonparametric modeling with the parsimony and easy interpretability of parametric modeling. PLMs, thus, play an important role in health sciences, economics and engineering studies (Engle et al., 1986; Robinson, 1988; Speckman, 1988; Andrews, 1994; Yatchew, 1997). For uncensored data, substantial amounts of works are available in the literature on partially linear models and their generalizations. For instance, You and Chen (2006), Zhou and Liang (2009) and Ma et al. (2013) proposed estimation procedure of varying coefficient partially linear model with error-prone covariates; Liang et al. (2011) and Wang and Sun (2007) considered PLMs where covariates and responses were missing at random, respectively; Zhang et al. (2016) accounted for time series error in a PLM; Kim (2016) included locally stationary regressors in a PLM; Li and Xue (2015) made inference in a generalized PLM with random

effect for longitudinal data; Hu et al. (2014) applied a PLM to panel data; Aneiros et al. (2015) estimated the error variance in semi-functional PLMs.

In survival analysis, data are often censored, relatively, fewer works are seen in the literature regarding partly linear models. Among those works, Huang (1999) considered the efficient estimation of the partly linear additive Cox model and used B-splines to estimate the nonparametric functions. Yin et al. (2008) proposed the partially linear varying-coefficient additive hazards model with a nonlinear interaction between the covariates and an exposure variable where they used kernel method to estimate the nonlinear function. Zhang (2016) investigated semiparametric estimation of a partially linear transformation model under conditional quantile restriction and its extension to random censoring. With current status data, Ma and Kosorok (2005) and Cheng and Wang (2011) studied penalized log-likelihood estimation for the partly linear transformation model; Ma (2011) researched the partly linear cure rate AHM. Recently, Lu and Song (2015) proposed efficient estimation of the partly linear additive hazards model with current status data and used B-splines to estimate the nonparametric functions.

In this Chapter, we consider the additive hazards model with a semiparametric risk function that has a partially linear structure in the same vein as that of Huang (1999) and Lu and Song (2015). More specifically, we assume that the conditional hazard function is given by

$$\lambda(t; W, \mathbf{X}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{X}(t) + \sum_{j=1}^{q} \varphi_j(W_j), \tag{2.2}$$

where $\lambda_0(t)$ and $\mathbf{X}(t)$ are defined in model (2.1), $\mathbf{W} = (W_1, \ldots, W_q)^\top$ is a q dimensional time-independent covariate vector, φ_j $(j = 1, \ldots, q)$ are known or unknown nonlinear smooth functions. The model now contains both the linear component $\boldsymbol{\beta}^\top \mathbf{X}(t)$ and the nonlinear components $\varphi_j(W_j), \quad j = 1, \ldots, q$, where the former is a parametric component and the latter could be parametric or nonparametric components. For technical reasons, we assume that $\varphi_j(\cdot), \quad j = 1, \ldots, q$, are B-splines functions with fixed knots and orders defined in (2.3),

hence they are also of parametric forms. In practice, these nonlinear functions may not be specified and are purely nonparametric. Under this situation, we suggest to approximate them by B-splines so that the proposed method still can be used. Our simulation studies shown at the end indicate the performance is quite satisfactory and the approximation error could be ignored. In summary, this model extends a purely linear model given in (2.1) and avoids the curse of dimensionality of a purely nonparametric model; the existing computing software makes it readily available for data analysis.

The rest of the chapter is organized as follows. In Section 2.2, we illustrate the estimation method for model (2.2) and present a theorem of asymptotic normality for the proposed estimator. In Section 2.3, we discuss implementation of the algorithm using some existing R packages. In Section 2.4, we examine the finite-sample properties using simulation studies and illustrate the proposed method with a real data set. Concluding remarks are made in Section 2.5. Finally, proof of Theorem 1 is presented in the Appendix.

2.2 Estimation Method and Theory

In this section, we will construct a pseudo-score function motivated by the work of Lin and Ying (1994) to estimate the parameters of the linear part and the nonlinear functions consisting of B-splines. Under some regularity conditions, we prove a theorem of asymptotic normality for the proposed estimator. Following Lin and Ying (1994), we use notations for counting processes to denote the left truncated and right censored survival data of size n drawn from the population characterized by model (2.2). We assume the survival times or responses T_i are not completely observable due to left-truncation and right-censoring by the random variables L_i and C_i, respectively. Let $(\mathbf{X}_i(t), \mathbf{W}_i)$ be the corresponding linear and nonlinear covariate vectors. Let $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I[T_i \leq C_i]$, $I[A]$ is the indicator function of a set A. Thus $(\tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$ can be observed only when $\tilde{T}_i \geq L_i$. The data, therefore, consist of n observations $(\tilde{T}_i, L_i, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$ with $\tilde{T}_i \geq L_i, i = 1, \ldots, n$. Define the

at-risk processes $Y_i(t) = I[\tilde{T}_i > t \geq L_i]$ and the counting processes $N_i(t) = \delta_i I[L_i \leq \tilde{T}_i \leq t]$. For more details related to data subject to left-truncation and right-censoring, see Lai and Ying (1991).

In order to use the estimation method of Lin and Ying (1994), we assume the nonlinear functions $\varphi_j(W_j)$'s are B-splines. Specifically, we assume W_j has a common support $[a, b]$ where a and b are finite numbers. For each nonlinear component, $\varphi_j(W_j)$, let $\tau_0 = a < \tau_1 < \cdots < \tau_{k'} < b = \tau_{k'+1}$ be a partition of $[a, b]$ into sub-intervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, k'$ with k' internal knots. A polynomial spline of order r is a function whose restriction to each sub-interval is a polynomial of degree $r-1$ and globally $r-2$ times continuously differentiable on $[a, b]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\left\{B^*_{j1}(x), \ldots, B^*_{j\tilde{k}}(x)\right\}$ with $\tilde{k} = k' + r$. As φ_j is identifiable only up to a constant, we put a centering constraint $E\{\varphi_j(W_j)\} = 0$, we instead focus on the subspace of spline functions $S^0_j := [s : s = \sum_{k=1}^{\tilde{k}} \alpha_{jk} B_{jk}(x), \sum_{i=1}^n s(W_{ij}) = 0]$, with basis $\left\{B_{jk}(x) = B^*_{jk}(x) - \sum_{i=1}^n B^*_{jk}(W_{ij})/n, k = 1, \ldots, K = \tilde{k} - 1\right\}$ (the subspace is $\tilde{k}-1$ dimensional due to the use of normalized B-spline basis functions, i.e., $\sum_{k=1}^{\tilde{k}} B_{jk}(x) = 1$ before centering). Therefore, each nonlinear component is expressed as

$$\varphi_j(x) = \sum_{k=1}^{K} \alpha_{jk} B_{jk}(x), \ 1 \leq j \leq q. \tag{2.3}$$

Using this spline expansion, the problem of estimating φ_j is then transformed to the problem of estimating the coefficients $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK})^\top$. Model (2.2) can be written as

$$\lambda(t; W, \mathbf{X}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{X}(t) + \sum_{j=1}^{q} \sum_{k=1}^{K} \alpha_{jk} B_{jk}(W_j) = \lambda_0(t) + \boldsymbol{\gamma}^\top \mathbf{Z}(t), \tag{2.4}$$

where $\boldsymbol{\gamma}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$, $\boldsymbol{\alpha}^\top = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_q^\top)^\top$, $\mathbf{Z}(t) = (\mathbf{X}(t)^\top, \mathbf{B}_1(W_1)^\top, \ldots, \mathbf{B}_q(W_q)^\top)^\top$, and $\mathbf{B}_j(W_j) = (B_{j1}(W_j), \ldots, B_{jK}(W_j))^\top$, $j = 1, \ldots, q$. Based on the linear additive hazards model studied by Lin and Ying (1994), the intensity function for $N_i(t)$ is given by

$$Y_i(t) d\Lambda(t; \mathbf{Z}_i) = Y_i(t)\{d\Lambda_0(t) + \boldsymbol{\gamma}^\top \mathbf{Z}_i(t) \, dt\}, \tag{2.5}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u) \, du$ is the cumulative baseline hazard function.

The counting processes $N_i(t)$ can be uniquely decomposed so that for every i and t,

$$N_i(t) = M_i(t) + \int_0^t Y_i(u)\, d\Lambda(u; \mathbf{Z}_i), \tag{2.6}$$

where $M_i(t)$ is a local square integrable martingale (Andersen and Gill, 1982). The cumulative baseline hazard function $\Lambda_0(t)$ is estimated by

$$\hat{\Lambda}_0(\boldsymbol{\gamma}, t) = \int_0^t \frac{\sum_{i=1}^n \{dN_i(u) - Y_i(u)\boldsymbol{\gamma}^\top \mathbf{Z}_i(u)\, du\}}{\sum_{i=1}^n Y_i(u)}. \tag{2.7}$$

Mimicking the partial likelihood score function of the multiplicative risk model, as done by Lin and Ying (1994), we propose to estimate $\boldsymbol{\gamma}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ from the estimating function,

$$U(\boldsymbol{\gamma}) = \sum_{i=1}^n \int_0^\tau \mathbf{Z}_i(t)\{dN_i(t) - Y_i(t)d\hat{\Lambda}_0(\boldsymbol{\gamma}, t) - Y_i(t)\boldsymbol{\gamma}^\top \mathbf{Z}_i(t)\, dt\}, \quad \text{for } 0 < \tau \leq \infty.$$

In practice, τ is the study ending time satisfying $P(\tilde{T} \geq \tau) > 0$. $U(\boldsymbol{\gamma})$ is equivalent to

$$\begin{aligned} U(\boldsymbol{\gamma}) &= \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\gamma}^\top \mathbf{Z}_i(t)\, dt\} \\ &= \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\, dN_i(t) - \\ &\quad \left[\sum_{i=1}^n \int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt\right]\boldsymbol{\gamma}, \end{aligned} \tag{2.8}$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\top$ for a column vector $\mathbf{a}$ and

$$\bar{\mathbf{Z}}(t) = \sum_{j=1}^n Y_j(t)\mathbf{Z}_j(t) / \sum_{j=1}^n Y_j(t).$$

The resulting estimator takes the explicit form

$$\hat{\boldsymbol{\gamma}} = \left[\sum_{i=1}^n \int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt\right]^{-1} \left[\sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\, dN_i(t)\right]. \tag{2.9}$$

In the following theorem, we give asymptotic normality of the derived estimator and provide a basis for inference.

Theorem 1. *Assume that the nonlinear functions in* (2.3), ψ_j $(j = 1, \ldots, q)$, *are B-splines,* $\boldsymbol{\alpha}_0$ *is their true coefficient vector,* $\boldsymbol{\beta}_0$ *is the true coefficient vector of the linear part,* $\boldsymbol{\gamma}_0^\top = (\boldsymbol{\beta}_0^\top, \boldsymbol{\alpha}_0^\top)^\top$. *Let*

$$\eta_r(t) = E\{Y_1(t)\boldsymbol{Z}_1^r(t)\}, \quad r = 0, 1,$$

$$A = E\int_0^\tau Y_i(t)\left\{\boldsymbol{Z}_i(t) - \frac{\eta_1(t)}{\eta_0(t)}\right\}^{\otimes 2} dt,$$
$$\Sigma = E\int_0^\tau \left\{\boldsymbol{Z}_i(t) - \frac{\eta_1(t)}{\eta_0(t)}\right\}^{\otimes 2} dN_1(t),$$

and $V = A^{-1}\Sigma A^{-1}$. *Then, we have*

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N(0, V),$$

where $\xrightarrow{d}$ *represents convergence in distribution,* V *can be consistently estimated by* $V_n = A_n^{-1}\Sigma_n A_n^{-1}$, *with*

$$A_n = n^{-1}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}\, dt$$

and

$$\Sigma_n = n^{-1}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}\, dN_i(t).$$

We relegate the proof of Theorem 1 to the Appendix.

Suppose that the estimated covariance matrices for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}_j$, $j = 1, \ldots, q$, are $V_{n\boldsymbol{\beta}}$ and $V_{n\boldsymbol{\alpha}_j}$, respectively, then the large sample $(1-\pi)100\%$-level confidence region for $\boldsymbol{\beta}_0$ or confidence intervals for β_i $(i = 1, \ldots, p)$ based on the above asymptotic normal distribution is given by

$$R_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} : n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top V_{n\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le \chi_p^2(\pi)\}, \tag{2.10}$$

$$\hat{\beta}_i - Z_{\pi/2}\mathrm{SE}(\hat{\beta}_i) \le \beta_i \le \hat{\beta}_i + Z_{\pi/2}\mathrm{SE}(\hat{\beta}_i), \tag{2.11}$$

where $\chi_p^2(\pi)$ is the $(1-\pi)100$th quantile of the chi-squared distribution with degrees of freedom p, $Z_{\pi/2}$ is the $(1-\pi/2)100$th quantile of the standard normal distribution and SE stands for standard error.

By the similar argument the large sample $(1-\pi)100\%$-level point-wise confidence interval for $\varphi_j(w)$ at a fixed value w is given as

$$\hat{\varphi}_j(w) - Z_{\pi/2}\mathrm{SE}(\hat{\varphi}_j(w)) \le \varphi_j(w) \le \hat{\varphi}_j(w) + Z_{\pi/2}\mathrm{SE}(\hat{\varphi}_j(w)), \tag{2.12}$$

where $\hat{\varphi}_j(w) = \hat{\boldsymbol{\alpha}}_j^\top \mathbf{B}_j(w)$ and $\mathrm{SE}(\hat{\varphi}_j(w)) = \left\{\mathbf{B}_j^\top(w) V_{n\boldsymbol{\alpha}_j}\mathbf{B}_j(w)\right\}^{1/2}$.

2.3 Implementation

To estimate the partly linear additive hazards model with left truncated and right censored data, we used two publicly available `R` packages: package `fda` and package `ahaz`. `R` (http://www.r-project.org) is a free programming language and software environment for statistical computing and graphics, and is widely used by statisticians for developing statistical software and data analysis. The packages in `R` are an efficient way to maintain collections of `R` functions and data sets. The package `fda` is developed by Ramsay, Wickham, Graves and Hooker to support functional data analysis as described in Ramsay and Silverman (2006). We used the function `bsplineS` from this package to generate B-spline basis functions which we later use as an argument in the `ahaz` function from the `ahaz` package. The `bsplineS` function mainly takes three arguments: a w vector of values at which the B-spline basis functions are to be evaluated, breaks or knot positions, and the order of the B-spline basis functions. In practice, we suggest equal probability spaced quantiles as knot positions. The second package `ahaz` is developed and maintained by Anders Gorst-Rasmussen and uses Lin and Ying (1994)'s procedure for estimating semiparametric additive hazards regression model. The function `ahaz` supports left truncated and right censored survival data, takes the survival object `Surv` as the response which is formed with observation times and censoring indicators, and the design matrix created by combining the linear covariates and B-spline basis functions obtained from package `fda`, and returns estimates with associated standard errors. To improve the efficiency of the estimators, we selected optimal number of knots by implementing the BIC proposed by Gorst-Rasmussen and Scheike (2011) at each simulation run where we keep the number of knots same for all nonparametric functions in one simulation but allow it to vary between simulations.

2.4 Numerical Studies

In this section, first, we conduct four simulation studies to evaluate the finite-sample performances of the proposed model. In Example 1, we consider right censoring only. In Example 2, we consider both left-truncation and right-censoring. In these simulation studies, we mimic scenarios where the nonlinear functions $\varphi_j(w_j)$ are assumed to be unknown, and then use B-splines to approximate them in estimation. In Example 3, we set the nonlinear functions to be exact B-spline functions with left-truncated and right censored data. In Example 4, we compare the proposed B-spline approach with the kernel-based approach of Yin et al. (2008) with right censored data only. At the end, we apply the proposed method to a data set from the South Wales Nickel Refines Study, which involves both left-truncation and right-censoring. In these numerical studies, we set the order of B-splines at 4 and the knots at the quantiles of the observed w's, which implies that the B-splines are cubic B-splines. Then we selected optimal number of knots by the BIC proposed by Gorst-Rasmussen and Scheike (2011), we found that in all these numerical studies, on average, zero interior knots were selected, so that the number of basis functions were four. This also implies that in most applications, cubic B-splines without interior knots are enough to fit underlying nonlinear functions, so the model is quite parsimonious and practically useful.

2.4.1 Simulation Study

Example 1: Data are right censored without left-truncation. Event times are generated from an exponential distribution with a hazard rate given as follows:

$$\lambda(t|\mathbf{x}, \mathbf{w}) = \lambda_0(t) + \boldsymbol{\beta}^\top \mathbf{z} + \varphi_1(w_1) + \varphi_2(w_2),$$

where $\lambda_0(t) = 5$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0.3, 0.5)^\top$, and the covariates $\mathbf{z} = (z_1, z_2)^T$, which are generated from $Z_1 \sim \mathrm{U}(1, 4) - 2.5$ and $Z_2 \sim \text{Bernoulli}(p = 0.5) - 0.5$. The two nonlinear functions are $\varphi_1(w_1) = \sin\{\pi(w_1/3 - 1)\}$ and $\varphi_2(w_2) = 0.3\{(w_2 - 6)^2 - 3\}$, with both W_1 and $W_2 \overset{\text{iid}}{\sim} \text{Uniform}(3, 9)$. The covariates are centered such that $E(Z_1)$=$E(Z_2)$=

$E\{\phi_1(W_1)\}=E\{\phi_2(W_2)\}=0$. The censoring time C follows a uniform distribution $U(0, c_0)$ with c_0 chosen to obtain a censoring proportion (cp) of about 20%, 30% and 40%, respectively. The sample size has been chosen to be 50, 200 and 500 respectively. The simulation is replicated 1000 times for each combination of n and c_0. For the confidence interval with nominal level $(1-\alpha)100\% = 95\%$, the coverage probability and joint coverage probability are computed from the Wald test statistics. We approximated the nonlinear functions using B-spline functions. Since the sum of the basis functions equals 1 for any w in the support of W_1 and W_2, we tackled the overparameterization problem by removing the last basis function. The final estimates of the nonlinear functions were centered at their Monte-Carlo sample means.

Table 2.1 summarizes the results of estimation and coverage probability. We observe that the averages of parameter estimates are close to the true values, so the estimation consistency is evident. As sample size increases, the sample standard deviations and the averaged estimated standard deviations decrease dramatically and are almost identical. Even with a high censoring proportion of 40%, the estimation performance is very satisfactory. The coverage probabilities almost reach the nominal level irrespective to the sample size and censoring proportion.

Figure 2.1 and Figure 2.2 show the fitted curves and 95% point-wise confidence bands of $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$. It is evident that average estimated curves capture the true curves very well and the true curves lie in the 95% point-wise confidence bands. Moreover, with an increase of sample size, the 95% point-wise confidence interval becomes narrower.

Example 2: Data are left truncated and right censored. The setup of this simulation is the same as that in Example 1 except for the conditions that: (i.) the baseline hazard depends on time now and is taken to be $\lambda_0(t) \equiv 0.1t + 3.3$, and (ii.) the data are also left truncated. Here, survival times are subject to the left-truncation where the left-truncation times are generated from a Uniform[0, 10], then the residual survival times $\tilde{T} - L$ are right

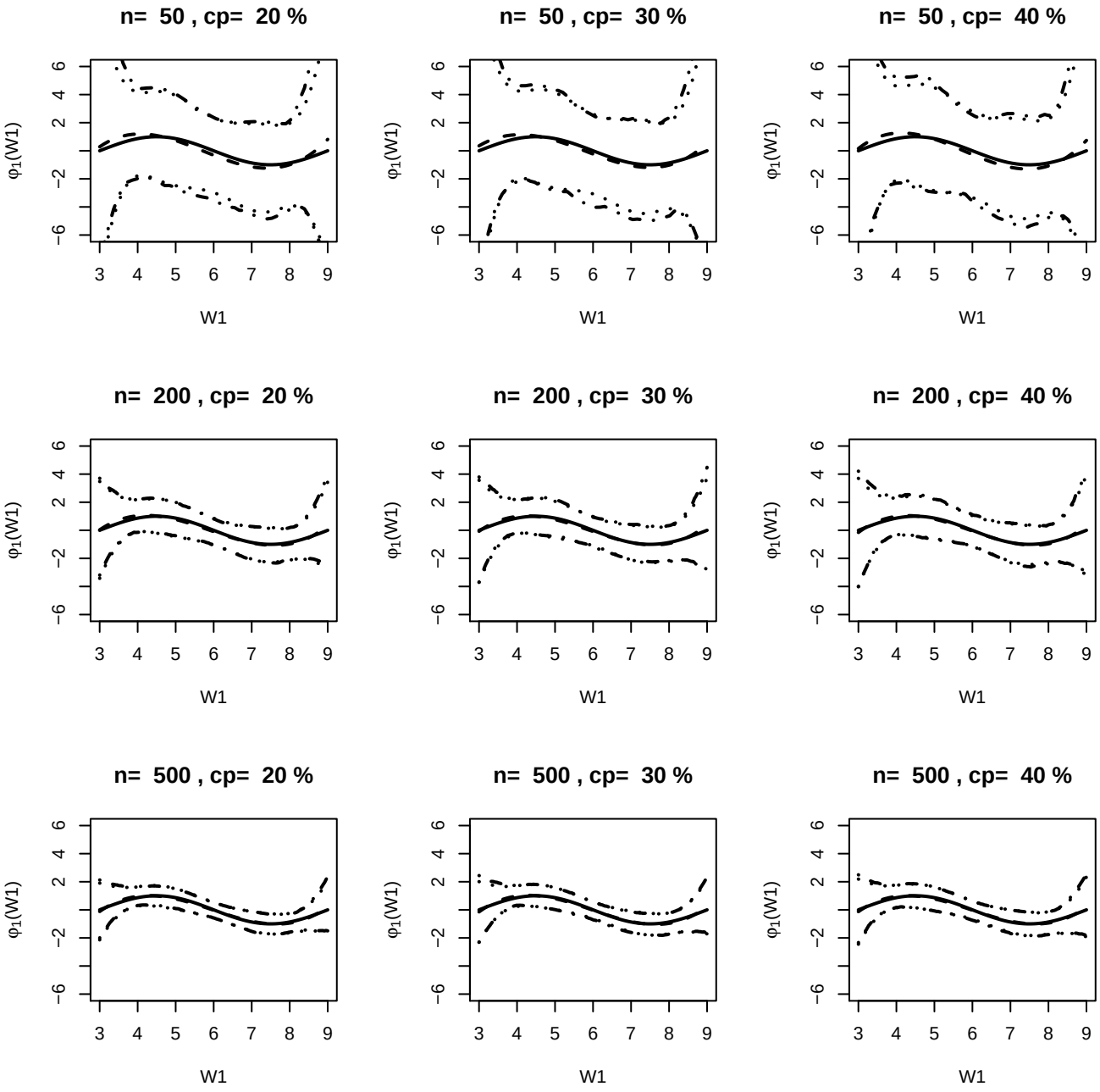


Figure 2.1: Estimation of $\varphi_1(\cdot)$ in Example 1: 95% confidence bands for function $\varphi_1(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

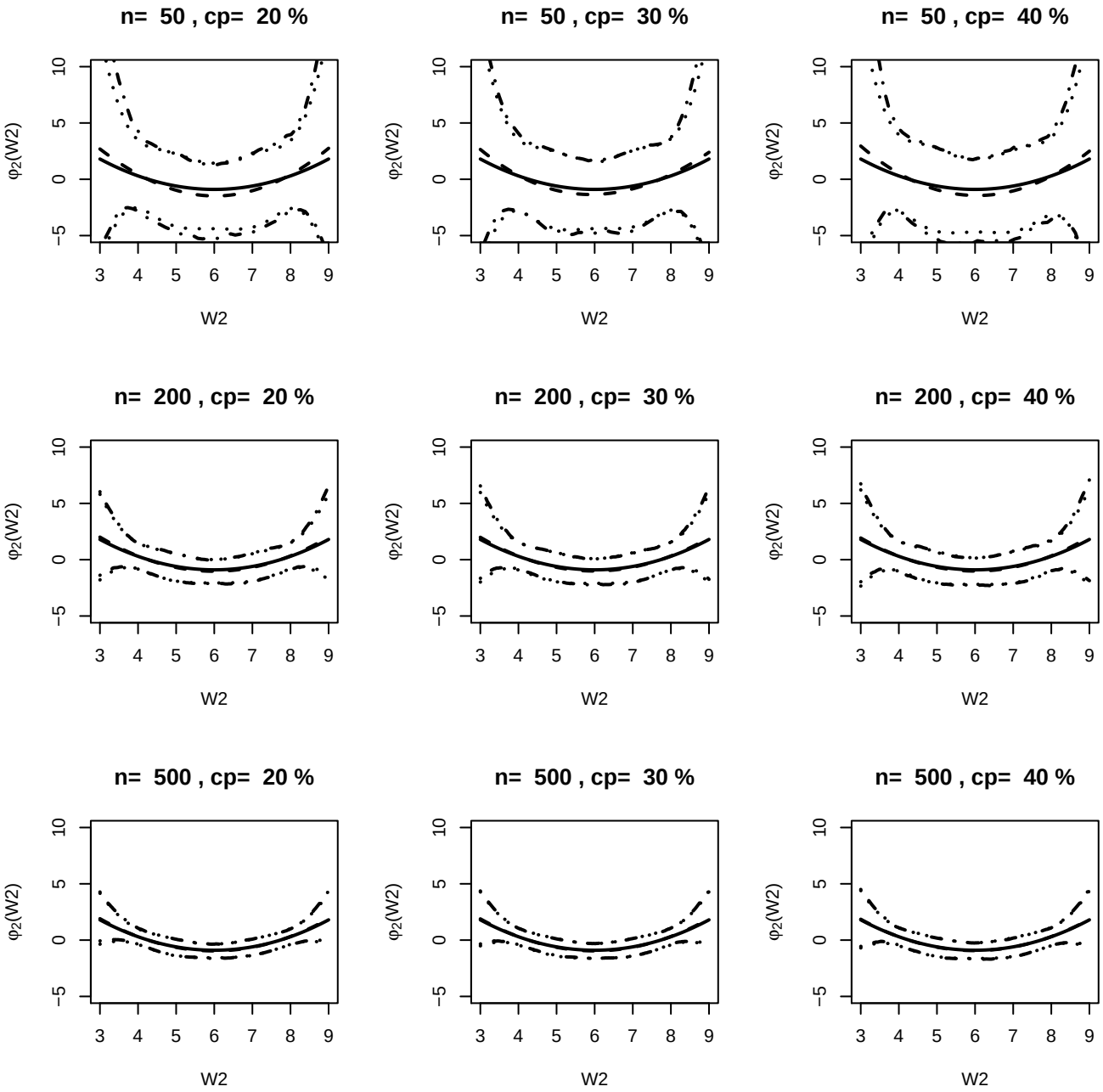


Figure 2.2: Estimation of $\varphi_2(\cdot)$ in Example 1: 95% confidence bands for function $\varphi_2(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

censored by the censoring times $C \sim$ Uniform $[0, c_0]$, where c_0's are chosen to obtain a pre-specified censoring proportion of about 20%, 30% and 40%, respectively.

Table 2.2 summarizes the results of estimation and coverage probability. Compared to the results in Example 1, even the data are left truncated, the estimation performance remains quite satisfactory. The coverage probabilities of 95% confidence intervals almost reach the nominal level irrespective to the sample size and censoring proportion. The fitted curves and 95% point-wise confidence bands for $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ are shown in Figure 2.3 and Figure 2.4. It is evident that average estimated curves capture the true curves very well for $n = 200$ or larger and the true curves lie in the 95% point-wise confidence bands. Also, the 95% point-wise confidence interval becomes narrower as the sample size increases.

Example 3: Data are left truncated and right censored and the nonlinear functions are B-spline functions.

The setup of this simulation is the same as that in Example 2 except that the two nonlinear functions $\varphi_1(w_1)$ and $\varphi_2(w_2)$ are B-spline functions. First we generate six B-spline basis functions in the interval $[3, 9]$ with two interior knots fixed at 5 and 7, which are shown in Figure 2.5. Using the first five basis functions plus an intercept, we define the two B-spline functions as

$$\tilde{\varphi}_1(w_1) = 0.03 - 0.06B_1(w_1) + 0.78B_2(w_2) + 1.58B_3(w_1) - 1.63B_4(w_1) - 0.83B_5(w_1)$$

and

$$\tilde{\varphi}_2(w_2) = 1.80 + 1.4 \times 10^{-6}B_1(w_2) - 1.20B_2(w_2) - 2.80B_3(w_2) - 2.80B_4(w_2) - 1.20B_5(w_2),$$

then center them at $E\{\tilde{\varphi}_1(W_1)\}$ and $E\{\tilde{\varphi}_2(W_2)\}$, thus we obtain two centered B-spline functions given by $\varphi_1(w_1) = \tilde{\varphi}_1(w_1) - E\{\tilde{\varphi}_1(W_1)\}$ and $\varphi_2(w_2) = \tilde{\varphi}_2(w_2) - E\{\tilde{\varphi}_2(W_2)\}$, respectively. The two expected values were estimated by Monte-Carlo sample means. The coefficients specified in $\tilde{\varphi}_1(w_1)$ and $\tilde{\varphi}_2(w_2)$ were obtained by fitting a linear model to the true values of the two nonparametric functions in Example 2 with a very small disturbance, so that these two B-spline functions behave like those two nonparametric functions.

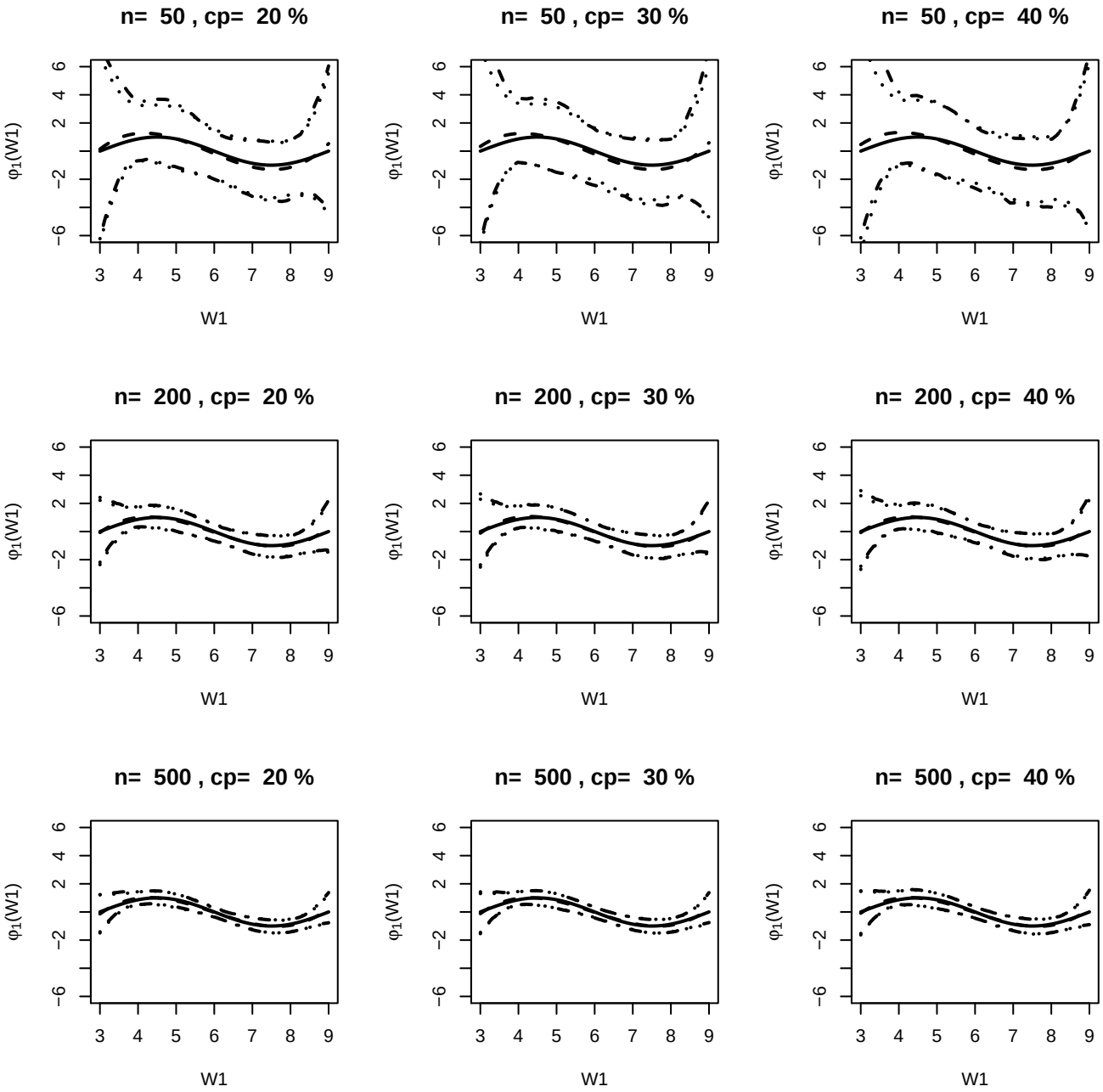


Figure 2.3: Estimation of $\varphi_1(\cdot)$ in Example 2: 95% confidence bands for function $\varphi_1(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

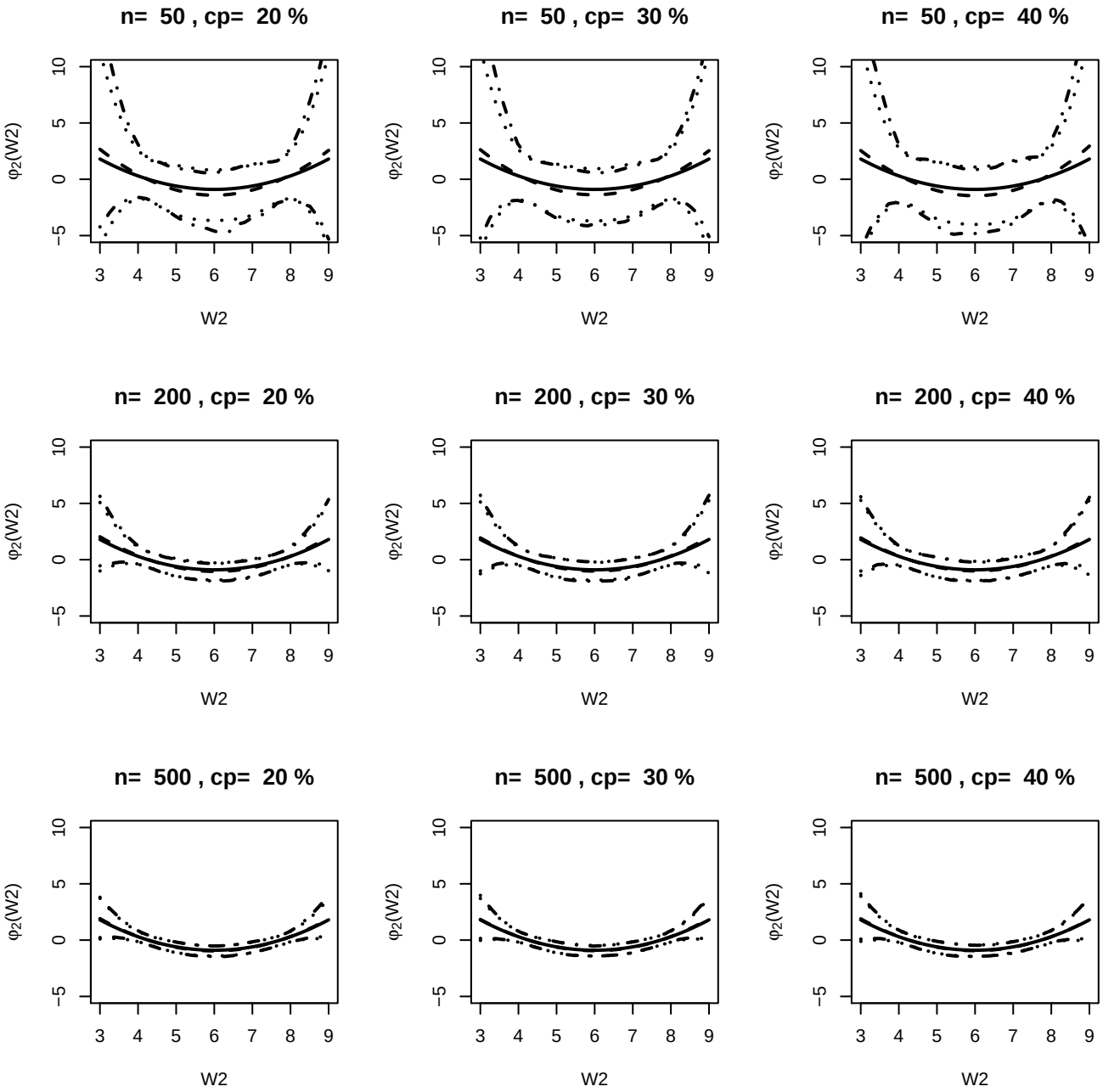


Figure 2.4: Estimation of $\varphi_2(\cdot)$ in Example 2: 95% confidence bands for function $\varphi_2(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

Table 2.3 summarizes the results of estimation and coverage probability. Compared to the results in Example 2, the performance of the parameter estimators are almost the same. The estimated B-spline functions and their 95% point-wise confidence bands shown in Figures 2.6 and 2.7 are analogous to those in Figures 2.3 and 2.4. The results in Examples 2 and 3 indicate that the proposed method works almost equally well even if the nonlinear functions are completely unknown and nonparametric.

In summary, in the partly linear additive hazards model, our simulation results indicate that even the nonlinear components are not B-splines, using our method, these functions are treated as B-splines, then they become parametric functions and can be estimated along with the linear component. The proposed method can work very well and has satisfactory performance in estimation for left truncated and right censored data. This is very important in real applications, where it is difficult to know the functional forms of nonlinear risk functions, a practical solution as the proposed method is desirable.

Example 4: Comparison with the kernel-based approach.

From a different motivation, Yin et al. (2008) studied the partially linear varying-coefficient additive hazards model using a kernel-based approach. To compare with their method, we applied the B-spline approach to both the varying-coefficient functions and the additive functions. Specifically, we used their Simulation I as an example and compare our results with their results reported in Table 1 of their paper. The data were generated from the following model

$$\lambda(t|Z,V,W) = \lambda_0(t) + \beta(W)Z + \gamma V + \alpha(W),$$

where $\beta(w) = 1.2 + \sin(2w)$, $\gamma = 1$, $\alpha(w) = 0.2w$, and $\lambda_0(t) = 0.5$. The covariate Z was generated from a uniform distribution $U[0,1]$, and the covariate V was generated from a Bernoulli random variable taking value 0 or 1 with probability 0.5. The censoring time was taken as $\min(C_0, \tau)$ with C_0 generated from uniform $U[\tau/2, 3\tau/2]$. Following their approach, we took $\tau = 0.86$ which yielded an approximate censoring rate of 25% without left-truncation,

chose 29 even partitions at $w_0 = (0.1, 0.2, \ldots, 2.9)$ and computed the estimators for $\beta(w_0)$, $\alpha(w_0)$ and γ. We considered sample sizes $n = 200$ and 400 and replicated 500 simulations.

For the confidence interval estimation of the derivative $\alpha'(w_0)$, we used the same idea of (2.12). That is, assuming $\alpha(w_0) = \boldsymbol{\theta}_1^\top \mathbf{B}(w_0)$ and $\beta(w_0) = \boldsymbol{\theta}_2^\top \mathbf{B}(w_0)$, $\mathbf{B}(w_0)$ is B-spline basis functions evaluated at w_0, then the large sample $(1-\pi)100\%$-level confidence interval for $\alpha'(w_0)$ is given as

$$\hat{\alpha}'(w_0) - Z_{\pi/2}\mathrm{SE}(\hat{\alpha}'(w_0)) \leq \alpha'(w_0) \leq \hat{\alpha}'(w_0) + Z_{\pi/2}\mathrm{SE}(\hat{\alpha}_0'(w_0)), \tag{2.13}$$

where $\hat{\alpha}'(w_0)$=$\hat{\boldsymbol{\theta}}_1^\top \mathbf{B}'(w_0)$ and $\mathrm{SE}(\hat{\alpha}'(w_0))$= $\left\{\mathbf{B}'^T(w_0) V_{n\boldsymbol{\theta}_1} \mathbf{B}'(w_0)\right\}^{1/2}$.

The results are presented in Table 2.4. We observe that the results based on the B-spline approach are comparable to or even superior to the kernel based approach, particularly for the estimate of $\alpha'(w)$, the estimate is more efficient in terms of the size of sample standard deviations.

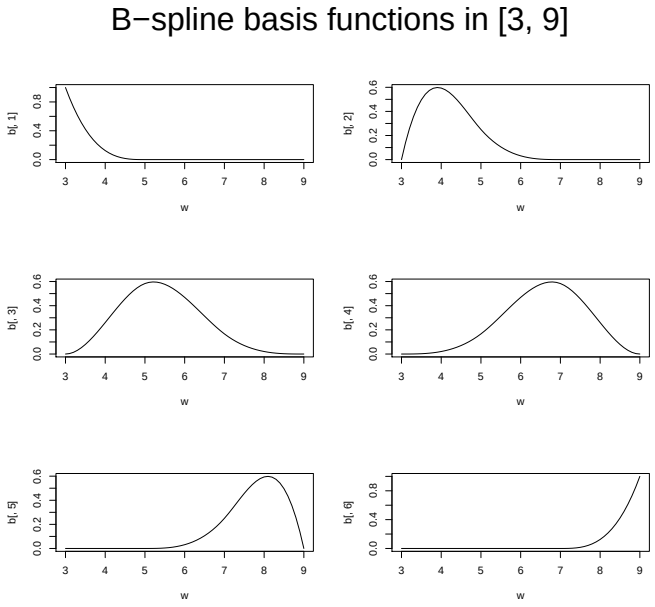


Figure 2.5: B-spline basis functions in Example 3, order=4 and two interior knots are 5 and 7.

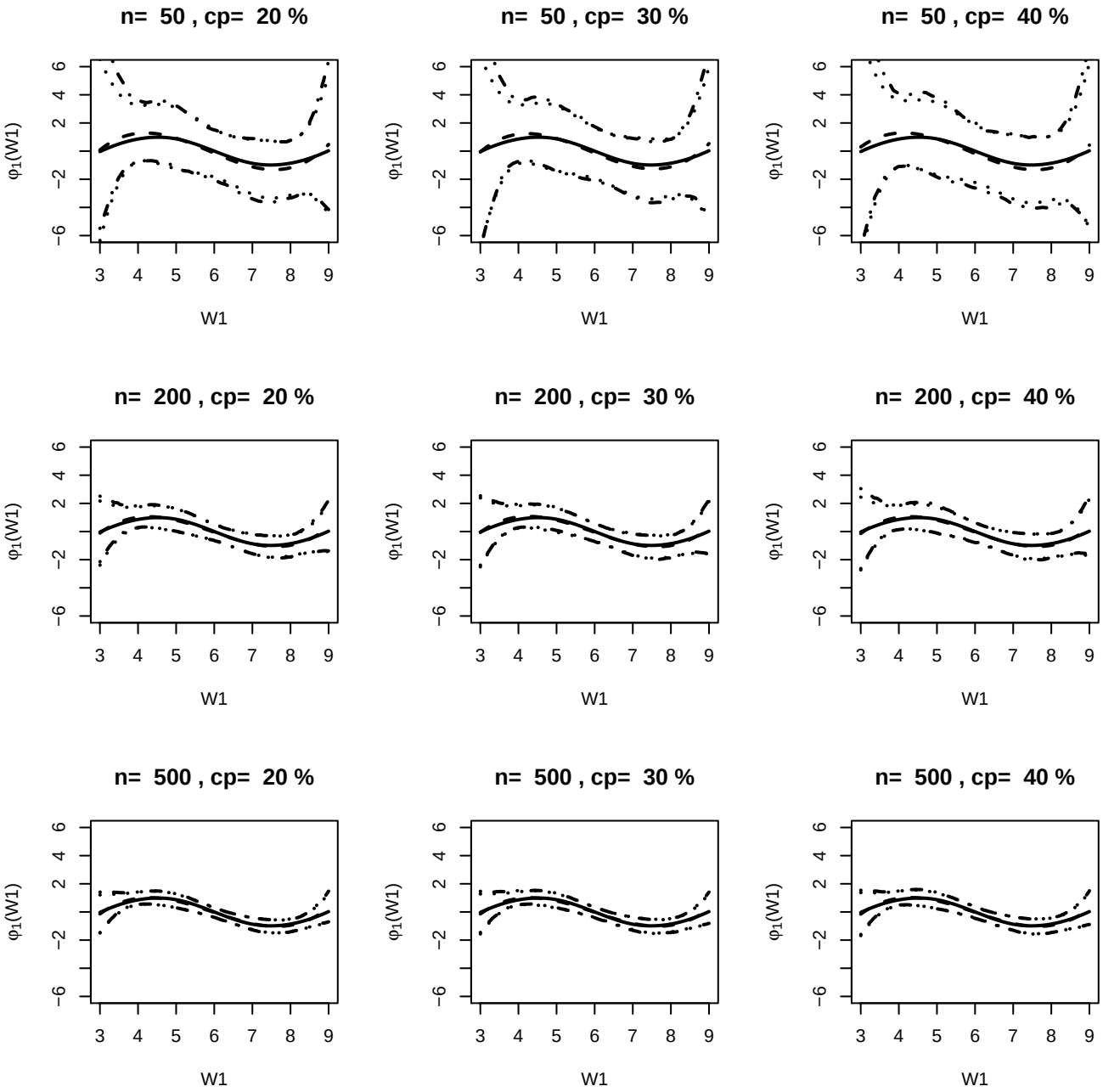


Figure 2.6: Estimation of $\varphi_1(\cdot)$ in Example 3: 95% confidence bands for function $\varphi_1(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

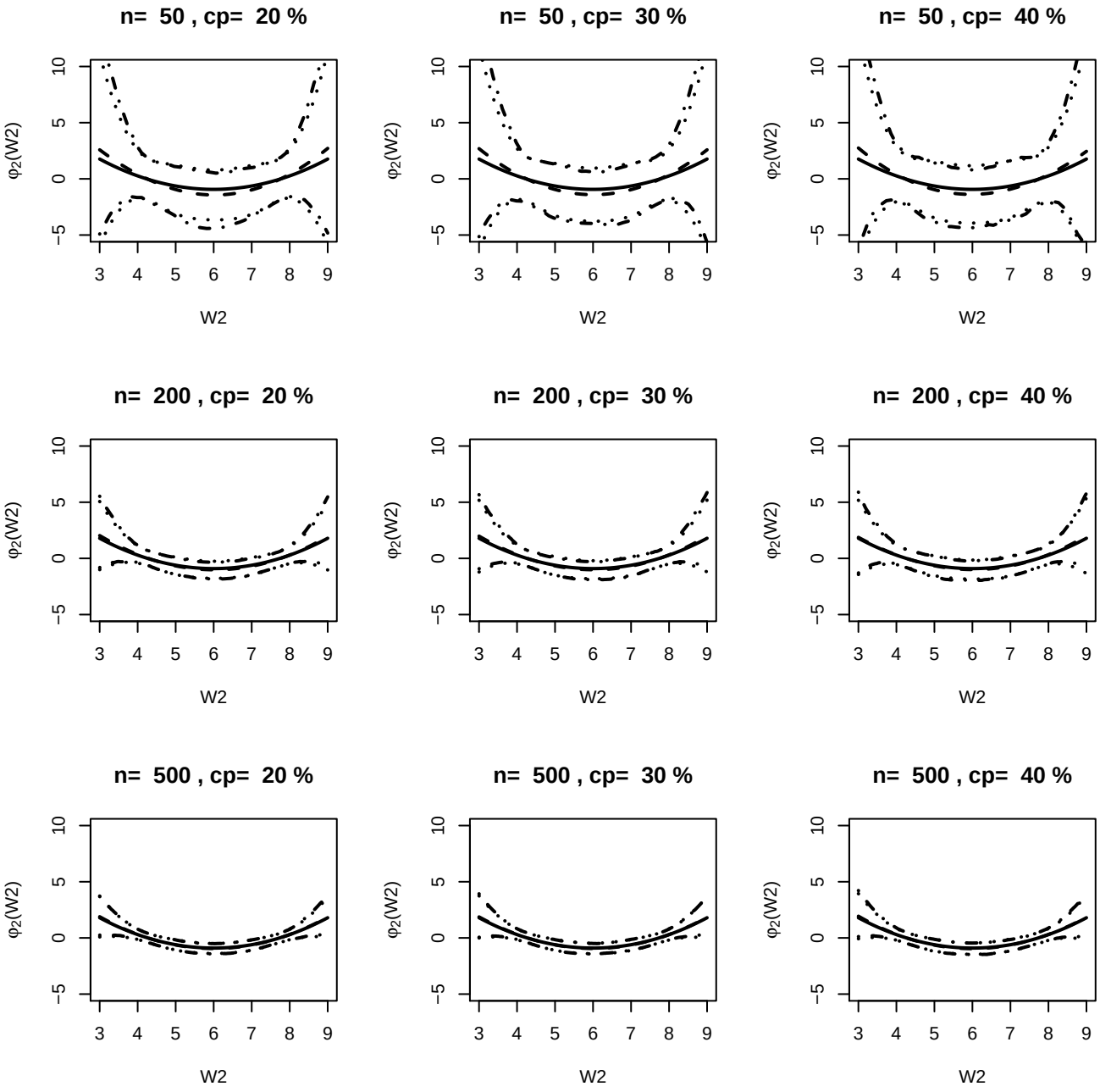


Figure 2.7: Estimation of $\varphi_2(\cdot)$ in Example 3: 95% confidence bands for function $\varphi_2(\cdot)$ based on 1000 replicates with different sample sizes and censoring proportions. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dotted lines and and the dot-dashed lines represent the 95% point-wise confidence bands based on averaged estimated standard deviation and sample quantiles, respectively.

2.4.2 Real Data Analysis: the South Wales Nickel Refiners Study

For illustration of the proposed estimation procedures, we apply them to the South Wales nickel refiners study (Breslow and Day, 1987, Appendix D). The data contained complete records for 679 men workers employed before 1925 in a nickel refinery in South Wales. The purpose of the study is to determine the risk of developing carcinoma of the bronchi and nasal sinuses associated with the refining of nickel. The follow-up through 1981 uncovered 137 lung cancer deaths among men aged $40 \sim 85$ years and 56 deaths from cancer of the nasal sinus. Since the workers had been working in the company for various periods of time before the follow-up was initiated, their survival times were subject to left-truncation. A right-censored observation arose either because the worker died from a competing cause or because he was still alive on the date of data listings. Breslow and Day (1987, pp. 222-223) and Lin and Ying (1994) fitted the conditional proportional hazards model and the conditional additive hazards model to the data respectively. They considered survival time to be years since first employment and found three significant risk factors: age at first employment, AFE, year at first employment, YFE, and exposure level, EXP. Their final estimation results based on the additive hazards model are given under the conditional estimating equation estimator $\hat{\beta}_{AHM}$ columns in Table 2.5 along with the estimates obtained from the Cox PH model.

Let $w = (YFE - 1915)/10$, both the PH model and AH model reported in Lin and Ying (1994) included w and w^2, hence, w has a quadratic effect on the hazard rates. However, the quadratic effect in their models is a subjective choice, it is interesting to see a data driven functional form of the w effect. This suggests that the partly linear AH model perfectly suits the need. Therefore, we apply the proposed model to the data. Let X_1=log(AFE-10), X_2=log(EXP+1), the new model can be expressed as follows:

$$\lambda(t; W, \mathbf{X}) = \lambda_0(t) + \beta_1 X_1 + \beta_2 X_2 + \varphi(W).$$

We use equal probability spaced knots in B-splines. The BIC proposed by Gorst-Rasmussen and Scheike (2011) selects one inner knot, the fitted nonparametric function and 95%

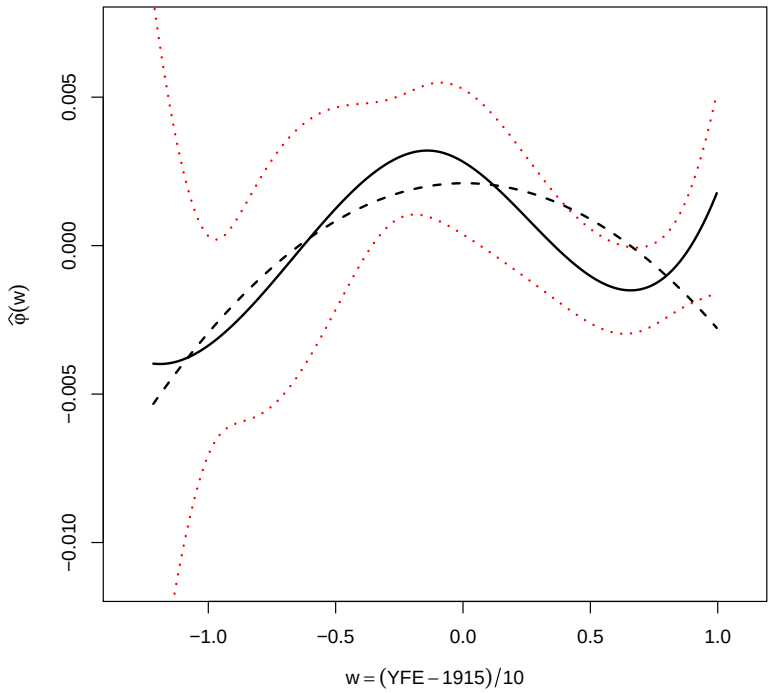


Figure 2.8: Estimated nonparametric function $\hat{\varphi}(w)$ (w=(YFE-1915)/10) in the partly linear additive hazards model for the nickel data. The solid line is the estimated curve, the dotted lines are the 95% point-wise confidence bands, the dashed line is the centered estimated quadratic polynomial curve $0.00005w - 0.00496w^2$, computed from the fitted linear additive hazard model.

point-wise confidence bands are shown in Figure 2.8, and the estimates for the linear parameters are reported in Table 2.5. Compared to the linear AH model, we observe that the partly linear AH model estimates the linear effects similarly, but for the nonlinear effect of $w = (YFE - 1915)/10$, the proposed model provides a different trajectory of estimation. The difference is shown in the tail part of the curve after $w = 0.6$, where the fitted curve shows an opposite direction from the parametric polynomial curve. This indicates that the nonlinear effect of w may not be explained simply by a quadratic polynomial.

2.5 Concluding Remarks

The additive hazards model serves as an important alternative to the proportional hazards model. To enhance modeling flexibility, we have studied the semiparametric partly linear additive hazards model and established the estimation and inference procedures. We applied polynomial splines with the computationally favorable B-spline basis, which allows reasonable approximation of smooth functions with just a small number of basis functions. The proposed model and estimation procedure are particularly attractive due to the analytic solution for the estimator and the easy computation using the existing `R` packages.

We treated the true nonlinear functions as B-spline functions and assumed the number and locations of the knots were fixed, and developed the asymptotic theory for inference. Huang and Liu (2006) took a similar approach in studying the single-index proportional hazards model for analyzing right-censored data. They used polynomial splines estimation along with a partial likelihood approach to estimate the parameters of the model. They also suggested to use BIC to select number of knots in B-splines. In practice, the true nonlinear functions are not necessarily B-splines functions, deriving large sample properties of the estimators under this setting is an open problem and is worthy of further investigation in our future research. However, our simulation studies showed that the proposed method can be used as an approximate approach to the underlying problems and such an approximate inference procedure appears to be quite accurate and can be applied to effectively solve real problems.

To specify a partially linear model, at least two possible strategies are applied in the literature. One is simply to put discrete covariates in the linear part and continuous ones in the nonlinear part. Another more reasonable approach is to first perform a preliminary univariate analysis and then separate the covariates based on the shape of the estimated nonparametric functions. In this paper, we assumed that a partially linear specification is already available one way or another and did not further consider how the partially linear

structure comes by in the first place. Apparently this is an interesting problem in itself, we leave it as a future work. Another consideration in our future research is to improve estimation efficiency when data subject to left-truncation, for that we can apply the method of Huang and Qin (2013) to combine the so-called marginal pairwise pseudo-score function and the conditional estimation function proposed by Lin and Ying (1994).

2.6 Appendix

Proof of Theorem 1. It is easy to see that

$$U(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\, dM_i(t),$$

which is a martingale integral. Following the standard counting process arguments of Andersen and Gill (1982), we have $E\{U(\boldsymbol{\gamma}_0)\} = \mathbf{0}$ and

$$\begin{aligned}
Var\{U(\boldsymbol{\gamma}_0)\} &= \sum_{i=1}^{n} Var\left[\int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\, dM_i(t)\right] \\
&= \sum_{i=1}^{n} E \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, d\langle M_i(t), M_i(t)\rangle \\
&= \sum_{i=1}^{n} E \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, Y_i(t) d\Lambda(t; \mathbf{Z}_i) \\
&= \sum_{i=1}^{n} E \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, \{dN_i(t) - dM_i(t)\} \\
&= \sum_{i=1}^{n} E \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dN_i(t) - \underbrace{\sum_{i=1}^{n} E \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dM_i(t)}_{0} \\
&= n\Sigma_{1n},
\end{aligned}$$

where $\Sigma_{1n} = E\int_0^{\tau} \{\mathbf{Z}_1(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dN_1(t) \longrightarrow \Sigma$, when $n \longrightarrow \infty$. By central limit theorem for martingales, $n^{-1/2}U(\boldsymbol{\gamma}_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$. That is, by Andersen and Gill (1982), the random vector $n^{-1/2}U(\boldsymbol{\gamma}_0)$ converges weakly to a $(p + Kq)$-variate normal with mean zero and a covariance matrix Σ, which can be consistently estimated by Σ_n. Next, expanding the score

function evaluated at the $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$ using a first order Taylor series so that

$$U(\hat{\boldsymbol{\gamma}}) = U(\boldsymbol{\gamma}_0) + \frac{\partial U(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\xi}_n} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0),$$

where ξ_n is between $\boldsymbol{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}$. Noticing that $U(\hat{\boldsymbol{\gamma}}) = 0$, we have the first order approximation

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = \left\{\frac{\partial U(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^T}\right\}^{-1}_{\boldsymbol{\gamma}=\boldsymbol{\xi}_n} U(\boldsymbol{\gamma}_0) = \left[\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt\right]^{-1} U(\boldsymbol{\gamma}_0).$$

We obtain

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) &= n\left[\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt\right]^{-1} \times n^{-1/2}U(\boldsymbol{\gamma}_0) \\ &= \left[n^{-1}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt\right]^{-1} \times n^{-1/2}U(\boldsymbol{\gamma}_0) \\ &= A_n^{-1} \times n^{-1/2}U(\boldsymbol{\gamma}_0), \end{aligned}$$

where $A_n = n^{-1}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dt$. Since $A_n \longrightarrow A$ in probability and $n^{-1/2}U(\boldsymbol{\gamma}_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$, by Slutsky's theorem and the central limit theorem for martingales, we have

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} N(0, V),$$

where $V_n = A_n^{-1}\Sigma_n A_n^{-1}$. It then follows that the estimator (2.9) converges weakly to a $(p + Kq)$-dimensional normal variate with mean zero and a covariance matrix which can be consistently estimated by V_n. The proof of Theorem 1 is completed. □

Table 2.1: Summary results of the simulation study in Example 1. Bias: bias of the parameter estimates; AESD: average estimated standard deviation of the parameter estimates; SSD: sample standard deviation of the parameter estimates; CP: coverage probability of the 95% confidence interval; *cp*: censoring proportion; n: sample size; Results are based on 1000 simulation replicates.

cp	n	True $\boldsymbol{\beta}$	Bias	AESD	SSD	CP	Joint CP
20%	50	$\beta_1 = 0.3$	0.071	1.242	1.351	0.945	0.961
		$\beta_2 = 0.5$	0.040	2.127	2.214	0.957	
	200	$\beta_1 = 0.3$	0.005	0.476	0.477	0.955	0.940
		$\beta_2 = 0.5$	0.011	0.826	0.842	0.933	
	500	$\beta_1 = 0.3$	0.004	0.288	0.285	0.957	0.962
		$\beta_2 = 0.5$	-0.005	0.498	0.499	0.954	
30%	50	$\beta_1 = 0.3$	0.007	1.317	1.292	0.960	0.955
		$\beta_2 = 0.5$	-0.035	2.247	2.375	0.952	
	200	$\beta_1 = 0.3$	0.019	0.516	0.519	0.955	0.956
		$\beta_2 = 0.5$	0.070	0.890	0.894	0.953	
	500	$\beta_1 = 0.3$	0.004	0.310	0.310	0.946	0.963
		$\beta_2 = 0.5$	-0.010	0.535	0.516	0.955	
40%	50	$\beta_1 = 0.3$	0.010	1.417	1.538	0.948	0.947
		$\beta_2 = 0.5$	0.083	2.419	2.512	0.959	
	200	$\beta_1 = 0.3$	0.020	0.558	0.554	0.959	0.955
		$\beta_2 = 0.5$	-0.001	0.961	0.962	0.954	
	500	$\beta_1 = 0.3$	0.011	0.335	0.326	0.956	0.949
		$\beta_2 = 0.5$	0.008	0.580	0.581	0.943	

Table 2.2: Summary results of the simulation study in Example 2. Bias: bias of the parameter estimates; AESD: average estimated standard deviation of the parameter estimates; SSD: sample standard deviation of the parameter estimates; CP: coverage probability of the 95% confidence interval; *cp*: censoring proportion; n: sample size; Results are based on 1000 simulation replicates.

cp	n	True $\boldsymbol{\beta}$	Bias	AESD	SSD	CP	Joint CP
20%	50	$\beta_1 = 0.3$	0.040	0.743	0.749	0.958	0.969
		$\beta_2 = 0.5$	0.190	1.295	1.295	0.963	
	200	$\beta_1 = 0.3$	0.009	0.274	0.283	0.941	0.958
		$\beta_2 = 0.5$	0.009	0.479	0.467	0.963	
	500	$\beta_1 = 0.3$	0.013	0.164	0.159	0.958	0.950
		$\beta_2 = 0.5$	0.018	0.287	0.290	0.946	
30%	50	$\beta_1 = 0.3$	0.083	0.796	0.780	0.960	0.964
		$\beta_2 = 0.5$	0.135	1.369	1.382	0.963	
	200	$\beta_1 = 0.3$	0.014	0.295	0.291	0.961	0.953
		$\beta_2 = 0.5$	0.028	0.515	0.531	0.949	
	500	$\beta_1 = 0.3$	0.007	0.176	0.168	0.961	0.960
		$\beta_2 = 0.5$	0.011	0.307	0.309	0.946	
40%	50	$\beta_1 = 0.3$	0.089	0.881	0.904	0.968	0.970
		$\beta_2 = 0.5$	0.006	1.514	1.523	0.958	
	200	$\beta_1 = 0.3$	0.022	0.323	0.314	0.952	0.955
		$\beta_2 = 0.5$	0.032	0.560	0.560	0.944	
	500	$\beta_1 = 0.3$	-0.005	0.193	0.187	0.963	0.951
		$\beta_2 = 0.5$	0.003	0.336	0.346	0.948	

Table 2.3: Summary results of the simulation study in Example 3. Bias: bias of the parameter estimates; AESD: average estimated standard deviation of the parameter estimates; SSD: sample standard deviation of the parameter estimates; CP: coverage probability of the 95% confidence interval; *cp*: censoring proportion; n: sample size; Results are based on 1000 simulation replicates.

cp	n	True $\boldsymbol{\beta}$	Bias	AESD	SSD	CP	Joint CP
20%	50	$\beta_1 = 0.3$	0.029	0.740	0.736	0.964	0.969
		$\beta_2 = 0.5$	0.133	1.265	1.207	0.973	
	200	$\beta_1 = 0.3$	0.016	0.275	0.268	0.962	0.960
		$\beta_2 = 0.5$	0.052	0.481	0.477	0.953	
	500	$\beta_1 = 0.3$	0.015	0.163	0.158	0.955	0.960
		$\beta_2 = 0.5$	0.013	0.286	0.293	0.947	
30%	50	$\beta_1 = 0.3$	0.062	0.799	0.808	0.962	0.971
		$\beta_2 = 0.5$	0.130	1.363	1.340	0.965	
	200	$\beta_1 = 0.3$	0.003	0.297	0.294	0.960	0.958
		$\beta_2 = 0.5$	0.012	0.517	0.504	0.965	
	500	$\beta_1 = 0.3$	0.007	0.176	0.175	0.953	0.958
		$\beta_2 = 0.5$	0.006	0.307	0.307	0.952	
40%	50	$\beta_1 = 0.3$	0.029	0.873	0.861	0.959	0.964
		$\beta_2 = 0.5$	0.143	1.506	1.503	0.954	
	200	$\beta_1 = 0.3$	0.037	0.321	0.316	0.957	0.958
		$\beta_2 = 0.5$	0.012	0.559	0.565	0.950	
	500	$\beta_1 = 0.3$	-0.001	0.192	0.193	0.954	0.952
		$\beta_2 = 0.5$	0.011	0.336	0.341	0.948	

Table 2.4: Summary results of the simulation study in Example 4. The results are compared with those in Table 1 of Yin et al. (2008) via the kernel-based approach. The censoring proportion was set at $cp = 25\%$.

		Kernel based approach				B-spline based approach			
		$\beta(w) = 1.2 + \sin(2w)$				$\beta(w) = 1.2 + \sin(2w)$			
n	w_0	Bias	AESD	SSD	CP(%)	Bias	AESD	SSD	CP(%)
200	0.5	0.190	1.226	1.186	95.4	0.163	1.036	1.072	93.8
	1.0	0.075	1.171	1.205	96.0	0.005	1.067	1.058	96.0
	1.5	0.021	1.123	1.078	95.6	0.009	0.855	0.836	96.6
	2.0	0.064	0.925	0.948	96.8	0.082	0.896	0.915	95.6
	2.5	0.016	1.048	0.994	95.8	0.005	0.908	0.913	95.6
400	0.5	-0.061	0.798	0.790	94.6	0.111	0.705	0.729	94.8
	1.0	-0.047	0.824	0.810	94.4	-0.037	0.733	0.736	95.4
	1.5	0.069	0.768	0.734	95.8	-0.038	0.586	0.580	95.6
	2.0	0.104	0.643	0.638	93.2	0.019	0.609	0.602	94.6
	2.5	0.109	0.674	0.660	94.4	-0.031	0.607	0.591	95.6
		$\alpha'(w) = 0.2$				$\alpha'(w) = 0.2$			
n	w_0	Bias	AESD	SSD	CP(%)	Bias	AESD	SSD	CP(%)
200	0.5	-0.069	1.787	1.740	95.8	-0.069	1.281	1.329	95.0
	1.0	0.095	1.644	1.565	95.8	0.009	0.650	0.704	95.0
	1.5	0.112	1.623	1.559	95.0	0.028	0.829	0.866	95.2
	2.0	0.023	1.561	1.518	95.2	0.050	0.660	0.658	95.0
	2.5	0.071	1.780	1.804	95.2	0.018	1.420	1.428	95.0
400	0.5	0.065	1.229	1.156	94.2	-0.041	0.844	0.868	95.4
	1.0	-0.102	1.094	1.069	94.6	0.001	0.431	0.471	94.2
	1.5	0.030	0.970	1.045	96.0	0.036	0.559	0.560	95.0
	2.0	-0.052	1.037	1.024	94.2	0.025	0.442	0.464	96.2
	2.5	0.098	1.197	1.192	96.4	-0.088	0.935	0.937	94.0
		$\gamma = 1$				$\gamma = 1$			
n		Bias	AESD	SSD	CP(%)	Bias	AESD	SSD	CP(%)
200		0.081	0.370	0.311	91.6	0.035	0.330	0.313	96.2
400		0.025	0.223	0.218	95.2	0.014	0.227	0.231	95.4

Table 2.5: Summary statistics of the Nickel Data Analysis with three different models: AH model ($\hat{\beta}_{AHM}$), PH model ($\hat{\beta}_{PHM}$) and partly linear AH model ($\hat{\beta}_{PLAHM}$), including estimate (Est) and estimated standard error (SE).

	$\hat{\beta}_{AHM}$		$\hat{\beta}_{PHM}$		$\hat{\beta}_{PLAHM}$	
Parameter	Est	SE	Est	SE	Est	SE
log(AFE-10)	0.00431	0.00083	2.22	0.44	0.00426	0.00086
log(EXP+1)	0.00373	0.00093	0.77	0.17	0.00380	0.00096
(YFE-1915)/10	0.00005	0.00102	-0.09	0.32	B-splines	NA
(YFE-1915)^2/100	-0.00496	0.00209	-1.26	0.51		

Chapter 3

Hierarchically Penalized Partially Linear Additive Hazards Model with a Diverging Number of Parameters

3.1 Introduction

As a useful alternative to the Cox proportional hazards model (PHM), the additive hazards model (AHM) assumes that the hazard function is the sum of the regression function of covariates and the baseline hazard function, and describes a different aspect of the relationship between survival time and covariates than the PHM. The AHM addresses the risk difference while the PHM deals with the risk ratio. When the excess risk is the quantity of interest, the AHM is more reasonable than the PHM. Buckley (1984) pointed out that the AHM is biologically more plausible than the PHM, while O'neill (1986) found that the use of the PHM may result in serious bias when the true model is additive. Aalen et al. (2008) in their book listed a number of reasons justifying the use of the AHMs. The AHM proposed by Lin and Ying (1994) has the following form:

$$h(t|X) = h_0(t) + \beta^\top X, \tag{3.1}$$

where $h_0(t)$ is a completely unspecified baseline hazard function, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients, and $X = (X_1, \ldots, X_p)^\top$ is a p-dimensional covariate vector.

Here we consider a partially linear additive hazards model (PL-AHM), presented in Chapter 2, which was proposed by Afzal et al. (2017) with right censored data. More specifically, we assume that the conditional hazard function is given by

$$h(t|W, X) = h_0(t) + \phi(W) + \beta^\top X, \tag{3.2}$$

where $\phi(W) = \sum_{q=1}^{Q} \phi_q(W_q)$, $W = (W_1, \ldots, W_Q)^\top$ is a Q dimensional covariate vector, ϕ_q $(q = 1, \ldots, Q)$ are known or unknown nonlinear smooth functions. This model combines the flexibility of nonparametric modeling with the parsimony and easy interpretability of parametric modeling, and avoids the curse of dimensionality of a purely nonparametric model.

In this chapter, we investigate the group variable selection problem in the linear part of the PL-AHM given in (3.2), where the covariates in X can be naturally grouped. The data and model settings are partly motivated by cancer prognosis studies reported in Ma and Huang (2007) and the variable selection method introduced by Ma and Du (2012) in the partly linear accelerated failure time (AFT) model with diverging dimensions in X for right censored data. In their studies, two distinct sets of covariates are measured. The first set X represents high-dimensional genomic measurements such as microarray gene expression or SNPs. The second set W represents low-dimensional clinical and environmental risk factors where the dimension Q of W was fixed and low. For better interpretability and easier computation, the effect of X is usually modeled in a parametric way and the effect of W is modeled with more flexible additive nonparametric functions, since many biological processes are nonlinear. However, variable selection based on such model settings mainly focuses on individual variables such as that in Ma and Huang (2007). In some applications, groups of measurements may be taken in the hopes of capturing unobservable latent variables or for measuring different aspects of complex entities (Breheny and Huang, 2009). Examples include measurements of gene expression, which can be grouped by gene pathways, and genetic markers, which can be grouped by the gene or haplotype (a set of genetic determinants located on a single chromosome) that they belong to. For example, as Wang et al. (2009) explained, when analyzing microarray gene expression data, one can group genes into functionally similar sets as in The Gene Ontology Consortium (2000), or into known biological pathways such as the Kyoto encyclopedia of genes and genomes pathways (Kanehisa and Goto, 2000). In these settings, methods for individual variable selection may perform inefficiently by ignoring the

information present in the grouping structure, while making use of the group information, as shown in Wang et al. (2009) and Huang et al. (2014), can help to identify both pathways and genes within the pathways related to the phenotypes, and hence improves understanding of biological processes.

Since grouping structures are natural in many important practical problems, several authors tackled the problem of variable selection in linear regression models with grouped covariates. Examples include group LASSO (Yuan and Lin, 2006), adaptive group LASSO (Wang and Leng, 2008; Wei and Huang, 2010), group SCAD (Wang et al., 2007) etc. All of these methods select variables in an all-in-all-out fashion. That is, a group of predictors are either all selected or all deleted from the model, and hence, these methods are not capable of differentiating important variables from the unimportant ones within a group.

Lately, bi-level group selection has attracted much attention since it can identify important groups as well as important variables within each selected group. Such a technique can be very useful in gene expression data where a biological pathway can be related to a certain biological outcome although some genes in that pathway may be not related to the biological outcome. Accordingly, it is sensible to identify important pathways, and important genes within important pathways, simultaneously. Popular bi-level group selection methods are group bridge (Huang et al., 2009), group MCP (Breheny and Huang, 2009), sparse group LASSO (Simon et al., 2013), adaptive sparse group LASSO (Fang et al., 2015), group exponential LASSO (Breheny, 2015) etc. Based on perspectives different from Huang et al. (2009), Zhou and Zhu (2010) proposed a hierarchically penalized method, which is a special case of the group bridge method in the linear regression model studied by Huang et al. (2009).

A few authors studied variable selection in the AHM, see Leng and Ma (2007); Martinussen and Scheike (2009); Lin and Lv (2013), among others. Their methods were for individual variable selection only. To conduct group selection, Liu et al. (2014) extended the hierarchical penalty to the AHM and established the oracle property of the estimators.

Recently, many authors have considered variable selection in partially linear models (PLMs). For example, variable selection in the linear part of a PLM has been extensively studied for uncensored data. Examples include Xie and Huang (2009), Ni et al. (2009), Liang and Li (2009), Zhao and Xue (2010), Kai et al. (2011), Xia and Yang (2016), Lv et al. (2016) and Yang et al. (2017), among others. Relatively fewer works are seen on variable selection in partially linear survival models (Johnson, 2009; Du et al., 2010; Long et al., 2011; Hu and Lian, 2013; Lian et al., 2014; Jicai et al., 2016; Liu et al., 2017). However, those authors focused on individual variable selection in the partially linear accelerated failure time model (PL-AFT) and partially linear proportional hazards model (PL-PHM), variable selection in the PL-AHM is yet to be investigated.

To the best of our knowledge, in the literature, group selection has not been investigated for the PL-AHM and PL-PHM. To bridge this gap, in this chapter, specifically, we propose a bi-level group variable selection in the PL-AHM with a diverging number of covariates X, assuming a group structure in the linear part and a fixed and low dimensional W for clinical and/or environmental covariates in the nonparametric part. Similar approach could be applied to other types of partially linear survival models, such as PL-PHM in the form of $h(t|W,X) = h_0(t)\exp\left\{\phi(W) + \beta^\top X\right\}$, in contrast to the PL-AHM given by (3.2), which will be addressed elsewhere in a different chapter. In this work, we consider the number of zero coefficients is diverging with the sample size. Generally, although the number of covariates collected is large, only a subset of covariates are important in predicting the event times. Therefore, we assume the numbers of non-zero coefficients and non-zero groups are fixed. Such an assumption is often reasonable with high dimensional data.

The remainder of the chapter is organized as follows. In Section 3.2, we describe the variable selection procedure for the PL-AHM. Asymptotic theories and further improvements are discussed in Section 3.3. In Section 3.4, numerical results are presented. Concluding remarks are made in Section 3.5. All technical proofs are contained in the Appendix.

3.2 Grouped Variable Selection in the PL-AHM

Suppose that a random sample of n subjects is observed. For the i-th subject, let T_i^e and T_i^c be the event time and the censoring time respectively, where the hazard function of T_i^e is given by (3.2). Assume that T_i^e and T_i^c are independent given the covariates, and the censoring mechanism is noninformative. The true nonparametric functions and parameters will be denoted using a superscript 0. The i.i.d observable random variables are $(T_i, \Delta_i, W_i, X_i)$ where $T_i = \min(T_i^e, T_i^c)$ and $\Delta_i = I[T_i^e \leq T_i^c]$, ($I[A]$ is the indicator function of a set A), $W_i = (W_{i1}, \ldots, W_{iQ})^\top \in \mathbb{R}^Q$, and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ are the covariates in the nonparametric and the parametric parts, respectively. Define the at-risk processes $Y_i(t) = I[T_i > t]$ and the counting processes $N_i(t) = \Delta_i I[T_i \leq t]$. Note that, ϕ_q is identifiable only up to a constant and thus we assume $E\{\phi_q(W_q)\} = 0$.

Following similar strategy of Wang et al. (2009), we assume that the p variables in the linear part X can be divided into G groups. Let the g-th group have p_g variables. We use $X_{i,(g)} = \left(X_{i,g1}, \ldots, X_{i,gp_g}\right)^\top$ to denote the p_g variables in the g-th group for the i-th observation, $X_i = \left(X_{i,(1)}^\top, \ldots, X_{i,(G)}^\top\right)^\top$ to denote the total p variables, and $\beta_{(g)} = \left(\beta_{g1}, \ldots, \beta_{gp_g}\right)^\top$ to represent the regression coefficients for the g-th group. We assume that the G groups do not overlap, i.e., each variable belongs to only one group.

Thus, the partially linear additive hazards model (3.2) can be written as

$$h(t|W, X) = h_0(t) + \phi(W) + \sum_{g=1}^{G} \sum_{j=1}^{p_g} \beta_{gj} X_{gj} = h_0(t) + \phi(W) + \beta_{(1)}^\top X_{(1)} + \cdots + \beta_{(G)}^\top X_{(G)}. \quad (3.3)$$

Next, we use polynomial splines to approximate the nonparametric components. Without loss of generality, we assume W_q ($q = 1, \ldots, Q$) has a support $[0, 1]$. For each non-parametric component, $\phi_q(W_q)$, let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{k'} < 1 = \tau_{k'+1}$ be a partition of [0, 1] into subintervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, k'$ with k' internal knots. A polynomial spline of order r is a function whose restriction to each subinterval is a polynomial of degree $r-1$ and globally $r-2$ times continuously differentiable on [0, 1]. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\left\{\tilde{B}_{q1}(x), \ldots, \tilde{B}_{q\tilde{k}}(x)\right\}$ with $\tilde{k} = k' + r$. As

ϕ_q is identifiable only up to a constant, we put a centering constraint $E\{\phi_q(W_q)\} = 0$, and use the subspace of spline functions: $S_q^0 := [s : s = \sum_{k=1}^{\tilde{k}-1} \alpha_{qk} B_{qk}(x), \sum_{i=1}^n s(W_{iq}) = 0]$, with basis $\left\{B_{qk}(x) = \sqrt{K}(\tilde{B}_{qk}(x) - \sum_{i=1}^n \tilde{B}_{qk}(W_{iq})/n), k = 1, \ldots, K = \tilde{k} - 1\right\}$ (the subspace has a degree $= \tilde{k} - 1$ due to the normalization constraint $\sum_{k=1}^{\tilde{k}} \tilde{B}_{qk}(x) \equiv 1$). The multiplicative constant $\sqrt{K}$ is incorporated in the basis definition to simplify some expression later in the proofs, as done in Wang et al. (2011). Using spline expansions, we can approximate the nonparametric components by $\phi_q(x) \approx \sum_{k=1}^K \alpha_{qk} B_{qk}(x), 1 \le q \le Q$. Therefore, the problem of estimating ϕ_q is now transformed to the problem of estimating the coefficients $\alpha_q = (\alpha_{q1}, \ldots, \alpha_{qK})^\top$.

Let $Z = (B_{11}(W_1), \ldots, B_{1K}(W_Q), \ldots, B_{Q1}(W_1), \ldots, B_{QK}(W_Q))^\top$ denote the QK basis functions and $\alpha = (\alpha_{11}, \ldots, \alpha_{1K}, \ldots, \alpha_{Q1}, \ldots, \alpha_{QK})^\top$ denote the corresponding coefficients. Since the q-th nonparametric component can be approximated by $\sum_{k=1}^K \alpha_{qk} B_{qk}(x)$ $(q = 1, \ldots, Q)$, it is reasonable to assume that $B_{q1}(x), \ldots, B_{qK}(x)$ are K variables belonging to one group. Therefore, the QK variables in Z can be divided into Q groups, where each of the q-th group has K variables. We use $Z_{i,(q)} = (B_{i,q1}, \ldots, B_{i,qK})^\top$ $(q = 1, \ldots, Q;\ k = 1, \ldots, K)$, to denote the K basis functions in the q-th group for the i-th observation. Similarly, we use $Z_i = \left(Z_{i,(1)}^\top, \ldots, Z_{i,(Q)}^\top\right)^\top$ to denote the total QK variables for the i-th observation, and $\alpha_{(q)} = (\alpha_{q1}, \ldots, \alpha_{qK})^\top$ to represent the regression coefficients for the q-th group. We assume that the number of variables in each group is K, i.e., we consider the same number of basis functions to approximate each nonparametric function. To simplify computation, since we have assumed W_q $(q = 1, \ldots, Q)$ have the same support [0,1], we can assume $B_{qk}(x) = B_{q'k}(x)$ for $q \ne q',\ 1 \le q, q' \le Q,\ 1 \le k \le K$.

The partially linear additive hazards model in (3.3) is then written as

$$h(t|Z, X) = h_0(t) + \sum_{q=1}^{Q} \alpha_{(q)}^\top Z_{(q)} + \sum_{g=1}^{G} \beta_{(g)}^\top X_{(g)}. \tag{3.4}$$

To estimate regression parameters (α^0, β^0), we propose a pseudo-score function $U_n(\alpha, \beta)$

following Lin and Ying (1994),

$$U_n(\alpha,\beta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{L_i - \bar{L}_n(t)\}\{dN_i(t) - Y_i(t)\left(\alpha^\top,\beta^\top\right)L_i\,dt\},$$

where $L_i = (Z_i^\top, X_i^\top)^\top$, $\bar{L}_n(t) = \sum_{j=1}^n Y_j(t)L_j / \sum_{j=1}^n Y_j(t)$ and τ is the study ending time with $P(T \geq \tau) > 0$. By some algebraic manipulation, $U_n(\alpha,\beta) = 0$ is equivalent to the linear equations $D_n\left(\alpha^\top,\beta^\top\right)^\top = d_n$, where

$$d_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{L_i - \bar{L}_n(t)\}\,dN_i(t) \text{ and } D_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{L_i - \bar{L}_n(t)\}^{\otimes 2}\,dt,$$

with $a^{\otimes 2} = aa^\top$ for any column vector a. Furthermore, solving the equation $U_n(\alpha,\beta) = 0$ with respect to (α,β) is equivalent to minimizing the following loss function,

$$L_n(\alpha,\beta) = \frac{1}{2}\left(\alpha^\top,\beta^\top\right)D_n\left(\alpha^\top,\beta^\top\right)^\top - d_n^\top\left(\alpha^\top,\beta^\top\right)^\top. \tag{3.5}$$

Following Hu and Lian (2013) and Lian et al. (2014), who conducted individual variable selection in the linear part of the PL-PHM, our model could be realized by minimizing the penalized loss function defined as follows:

$$pL_n(\alpha,\beta) = L_n(\alpha,\beta) + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j),$$

where $p_{\lambda_n}(\beta_j)$ is a penalty function. Let $(\hat{\alpha},\hat{\beta})$ be the minimizer of the above penalized loss function. Then, the penalized estimators of ϕ_q $(q = 1,\ldots,Q)$ and β are $\sum_{k=1}^K \hat{\alpha}_{qk}B_{qk}$ and $\hat{\beta}$, respectively. In this chapter, our focus is on group selection, and the above individual variable selection is a special case of the following group selection problem.

3.2.1 Hierarchically Penalized PL-AHM

To conduct group selection, we closely follow Wang et al. (2009) and Liu et al. (2014)'s procedure. Similar to theirs, we reparameterize β_{gj} as

$$\beta_{gj} = \gamma_g\theta_{gj}\ (g = 1,\ldots,G;\ j = 1,\ldots,p_g),$$

where $\gamma_g \geq 0$ for identifiability. This decomposition indicates that all β_{gj} $(j = 1, \ldots, p_g)$ belong to the g-th group as it treats β_{gj} hierarchically. Parameter γ_g explains β_{gj} $(j = 1, \ldots, p_g)$ at the group level and θ_{gj}'s explain differences among individuals within the g-th group. Let $\theta_{(g)} = (\theta_{g1}, \ldots, \theta_{gp_g})^\top$, then $\beta_{(g)} = \gamma_g \theta_{(g)}$. The loss function in (3.5) can be written as

$$L_n(\alpha, \gamma, \theta) = \frac{1}{2}(\alpha^\top, \gamma_1\theta_{(1)}^\top, \ldots, \gamma_G\theta_{(G)}^\top) D_n (\alpha^\top, \gamma_1\theta_{(1)}^\top, \ldots, \gamma_G\theta_{(G)}^\top)^\top \\ - d_n^\top (\alpha^\top, \gamma_1\theta_{(1)}^\top, \ldots, \gamma_G\theta_{(G)}^\top)^\top,$$

where $\gamma = (\gamma_1, \ldots, \gamma_G)^\top$ and $\theta = (\theta_{11}, \ldots, \theta_{1p_1}, \ldots, \theta_{G1}, \ldots, \theta_{Gp_G})^\top$. To select important variables in the linear part, we regularize the hierarchical parameters γ and θ by

$$\min_{\alpha_{(q)}, \gamma_g, \theta_{gj}} \left\{ L_n(\alpha, \gamma, \theta) + \lambda_\gamma \sum_{g=1}^{G} \gamma_g + \lambda_\theta \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\theta_{gj}| \right\}, \tag{3.6}$$

subject to $\gamma_g \geq 0$ $(g = 1, \ldots, G)$, where $\lambda_\gamma \geq 0$ and $\lambda_\theta \geq 0$ are two tuning parameters, which control the sparsity of the estimation at the group level and within group level, respectively. As indicated by Liu et al. (2014) in the linear AHM, for fixed (α, β) and given values of λ_γ and λ_θ, the minimizer of (3.6) with respect to (γ, θ), where $L_n(\alpha, \gamma, \theta)$ is constant, is unique.

Finally, in the same vein as Wang et al. (2009), we can combine λ_γ and λ_θ into one tuning parameter $\lambda = \lambda_\gamma \lambda_\theta$ such that (3.6) is equivalent to

$$\min_{\alpha_{(q)}, \gamma_g, \theta_{gj}} \left\{ L_n(\alpha, \gamma, \theta) + \sum_{g=1}^{G} \gamma_g + \lambda \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\theta_{gj}| \right\}, \tag{3.7}$$

subject to $\gamma_g \geq 0$ $(g = 1, \ldots, G)$. Lemma 1 illustrates the meaning of equivalence.

Lemma 1. *Let $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ be a local minimizer of (3.6). Then there exists a local minimizer $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ of (3.7) such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}_g^* \hat{\theta}_{gj}^* = \hat{\gamma}_g^\dagger \hat{\theta}_{gj}^\dagger$. Similarly, if $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ is a local minimizer of (3.7), then there exists a local minimizer $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ of (3.6) such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}_g^* \hat{\theta}_{gj}^* = \hat{\gamma}_g^\dagger \hat{\theta}_{gj}^\dagger$.*

Furthermore, criterion (3.7) can be written into an equivalent form using the regression coefficients α and β.

Lemma 2. *If $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local minimizer of (3.7), then $(\hat{\alpha}, \hat{\beta})$, where $\hat{\beta}_{gj} = \hat{\gamma}_g \hat{\theta}_{gj}$, is a local minimizer of the following objective function:*

$$Q_n(\alpha, \beta) = L_n(\alpha, \beta) + 2\lambda^{1/2} \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_g} |\beta_{gj}| \right\}^{1/2}. \tag{3.8}$$

On the other hand, if $(\hat{\alpha}, \hat{\beta})$ is a local minimizer of (3.8), then $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local minimizer of (3.7), where $\hat{\gamma}_g = (\lambda \sum_{j=1}^{p_g} |\hat{\beta}_{gj}|)^{1/2}$ and $\hat{\theta}_{gj} = \hat{\beta}_{gj}/\hat{\gamma}_g$ if $\hat{\gamma}_g \neq 0$ and zero, otherwise.

Instead of using L_2-norm which performs group LASSO (Yuan and Lin, 2006), we used L_1-norm to the within group coefficients in (3.8). In addition, the group coefficients are penalized by a bridge-type penalty (Frank and Friedman, 1993), i.e., $L_{1/2}$-norm. So, the hierarchical penalty can remove unimportant groups and some unimportant variables in the important groups.

3.3 Asymptotic Properties

We denote the true risk score by $m^0(W, X) = \phi^0(W) + \beta^{0^\top} X$ where $\phi^0(W) = \phi_1^0(W_1) + \cdots + \phi_Q^0(W_Q)$. Let $R^\top = (W^\top, X^\top)$ be all the covariates, and h be any function of R (h can be vector valued). Define

$$\begin{aligned} S_n^{(0)}(t) &= n^{-1} \sum_{i=1}^{n} Y_i(t), \\ S_n^{(1)}(t)[h] &= n^{-1} \sum_{i=1}^{n} Y_i(t) h(R_i), \\ S_n^{(2)}(t)[h] &= n^{-1} \sum_{i=1}^{n} Y_i(t) h(R_i)^{\otimes 2}. \end{aligned}$$

Let $s^{(j)}(t) = E\left\{S_n^{(j)}(t)\right\}$, $j = 0, 1, 2$, $\bar{L}_n(t) = S_n^{(1)}(t)[L]/S_n^{(0)}(t)$ $\bar{L}(t) = s^{(1)}(t)[L]/s^{(0)}(t)$, $D = E\left[\int_0^\tau Y(t) \left\{L - \bar{L}(t)\right\}^{\otimes 2} dt\right]$, and $\|a\|$ denote the l_2-norm of a.

Let us consider the penalized loss function with a general penalty function. The objective function that is to be minimized is

$$Q_{n,gen}(\alpha, \beta) = L_n(\alpha, \beta) + \sum_{g=1}^{G} p_{\lambda_n}^{(g)}(|\beta_{(g)}|), \tag{3.9}$$

where $p^{(g)}_{\lambda_n}(|\beta_{(g)}|) = p^{(g)}_{\lambda_n}(|\beta_{g1}|, \ldots, |\beta_{gp_g}|)$ is a general p_g-variate penalty function for the linear parameters in the g-th group. We let the penalty functions $p^{(g)}_{\lambda_n}(\cdot)$ $(g = 1, \ldots, G)$ in (3.9) to vary between groups as well as $p^{(g)}_{\lambda_n}(\cdot)$ to depend on the tuning parameter λ_n that differs with n.

Adopting notations of Wang et al. (2009), we write the true parameter vector in the sparse linear part as $\beta^0 = (\beta^{0\top}_{\mathcal{A}}, \beta^{0\top}_{\mathcal{B}}, \beta^{0\top}_{\mathcal{C}})^\top$, where $\mathcal{A} = \{(g,j) : \beta^0_{gj} \neq 0\}$, $\mathcal{B} = \{(g,j) : \beta^0_{gj} = 0, \beta^0_{(g)} \neq 0\}$, and $\mathcal{C} = \{(g,j) : \beta^0_{(g)} = 0\}$. Here $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ contain the indices of nonzero coefficients, indices of zero coefficients that belong to nonzero groups, and indices of zero coefficients that belong to zero groups. Thus, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are disjoint and partition the set of all indices of coefficients. We write $\mathcal{D} = \mathcal{B} \cup \mathcal{C}$, which contains the indices of all zero coefficients. We also define

$$a_n = \max_{(g,j)} \left\{ \frac{\partial p^{(g)}_{\lambda_n}(|\beta^0_{g1}|, ..., |\beta^0_{gp_g}|)}{\partial |\beta_{gj}|} : \beta^0_{gj} \neq 0 \right\},$$

$$b_n = \max_{(g,j)} \left\{ \frac{\partial^2 p^{(g)}_{\lambda_n}(|\beta^0_{g1}|, ..., |\beta^0_{gp_g}|)}{\partial |\beta_{gj}|^2} : \beta^0_{gj} \neq 0 \right\}.$$

Further, let s be the number of nonzero groups. Without loss of generality, we assume that $\beta^0_{(g)} \neq 0$ $(g = 1, \ldots, s)$ and $\beta^0_{(g)} = 0$ $(g = s+1, \ldots, G)$. Let s_g be the number of nonzero coefficients in group g $(g = 1, \ldots, s)$. Again, without loss of generality, we assume that $\beta^0_{gj} \neq 0$ $(g = 1, \ldots, s;\ j = 1, \ldots, s_g)$ and $\beta^0_{gj} = 0$ $(g = 1, \ldots, s;\ j = s_g + 1, \ldots, p_g)$.

The following regularity conditions are assumed to study the asymptotic properties:

(A1) (i) The covariate vector $R^\top = (W^\top, X^\top) = (R_1, \ldots, R_{p+Q})$ has a bounded support: without loss of generality the support of W is assumed to be $[0,1]^Q$, with the marginal density of each covariate in W being continuous and bounded away from zero and infinity, and the covariate vector X is bounded. (ii) There exist constants $M_1, M_2, \sigma > 0$ such that $P(\|R_j\| > x) \leq M_1 \exp(-M_2 x^\sigma)$ for all $x > 0$ and $j = 1, \ldots, p+Q$.

(A2) (i) Only observations with censored event times in a finite interval $[0, \tau]$ are used in the loss function. At the point τ, the baseline cumulative hazard

function $\Lambda_0(\tau) \equiv \int_0^\tau \lambda_0(s)ds < \infty$. (ii) $P(\Delta = 1|R)$ and $P(T^c > \tau|R)$ are both bounded away from zero with probability one.

(A3) Let $\mathscr{H}_d$ be the collection of all functions on support [0, 1] whose m-th order derivative satisfied the Hölder condition of order r with $d \equiv m + r$, i.e., for each $h \in \mathscr{H}_d$, there exists a constant $M_0 \in (0, \infty)$ such that $\left|h^{(m)}(s) - h^{(m)}(t)\right| \leq M_0|s - t|^r$, for any $s, t \in [0, 1]$. Assume, $\phi_q^0 \in \mathscr{H}_d$ $(q = 1, \ldots, Q)$, for some $d > 1/2$. The order of the spline satisfies $r > d + 1/2$.

(A4) $E\left[\sup_{t\in[0,\tau]} Y(t) \|L\|^2 \left\{\beta^{0\top} X + \alpha^{0\top} Z\right\}^2\right] = O(K + p)$.

(A5) The eigenvalues of D are bounded away from zero and infinity.

(A6) The p_g-variate penalty function for parameters in the g-th group satisfies the following two conditions:

$$p_{\lambda_n}^{(g)}(|\beta_{(g)}|) \geq 0 \quad (\beta_{(g)} \in \mathbb{R}^{p_g}), \quad p_{\lambda_n}^{(g)}(0) = 0; \tag{3.10}$$

$$p_{\lambda_n}^{(g)}(|\beta_{(g)}|) \geq p_{\lambda_n}^{(g)}(|\beta_{(g)}^*|) \quad (|\beta_{gj}| \geq |\beta_{gj}^*|;\ j = 1, \ldots, p_g). \tag{3.11}$$

Similar conditions to those listed above have been considered in the literature (Hu and Lian, 2013; Wang et al., 2009) and are quite reasonable. Condition (A1)(i) places the boundedness condition on the covariates. It is unpleasant, but not too restrictive because in many practical situations continuous covariates may be typically rescaled to fall between 0 and 1. (A1)(ii) controls the tail behavior of the covariates and is trivially satisfied for bounded covariates. (A2)(i) avoids the unboundedness of the loss function and pseudo-score function at the end point of the support of the observed event time. (A2)(ii) ensures that the probability of being right censored at τ and the probability of being observed events are positive and bounded away from zero regardless of the covariate values. (A3) ensures the uniform continuity of the functions. A condition similar to (A4) was considered by Bradic

et al. (2011) for diverging number of parameters following Andersen and Gill (1982). The positive-definiteness of D in (A5) is a reasonable assumption by the following discussion. The term $LL^\top$ appears in the definition of D. Under mild assumptions, Huang et al. (2010) showed that eigenvalues of $E(ZZ^\top)$ are bounded and bounded away from zero and hence, we can expect that eigenvalues of $E(LL^\top)$ are bounded and bounded away from zero if eigenvalues of $E(XX^\top)$ are, and Z and X are linearly independent. Wang et al. (2009) considered the condition (A6) about the properties of the penalty function in the hierarchical group variable selection in the linear PHM.

Theorem 1. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$. *Under regularity conditions (A1) - (A6), assume that* Q, s *and* s_g *are fixed,* $K \to \infty$, $p \to \infty$, $(K+p)/n \to 0$, $a_n = O_p(\gamma_n)$ *and* $b_n \to 0$, *there exists a local minimizer* $(\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ *of* $(\alpha^\top, \beta^\top)^\top$ *in (3.9) and* $\hat{\phi}_q(w_q) = \sum_{k=1}^{K} \hat{\alpha}_{qk} B_{qk}(w_q)$, $\hat{\phi}(w) = \sum_{q=1}^{Q} \hat{\phi}_q(w_q)$ *such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p\left(\sqrt{(K+p)/n} + K^{-d}\right)$.

Theorem 2. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$ *and* $(\hat{\alpha}^\top, \hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ *be the local minimizer of* $Q_{n,gen}(\alpha, \beta)$ *in (3.9). For* $(g, j) \in \mathcal{D}$, *i.e.,* $\beta_{gj}^0 = 0$, *under the same conditions as in Theorem 1, if* $\gamma_n^{-1} \partial p_{\lambda_n}^{(g)}(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}|)/\partial|\beta_{gj}| \to \infty$ *as* $n \to \infty$, *then we have* $\hat{\beta}_{gj} = 0$ *with probability approaching to 1.*

In the following section, we construct a penalty function $p_{\lambda_n}^{(g)}$ such that the conditions in Theorem 2 satisfies.

3.3.1 Adaptive hierarchically penalized method

The above results are obtained for any general penalty. Following Wang et al. (2009), here we will show the asymptotic results for the hierarchically penalized PL-AHM based on criterion (3.8). If we write $\lambda_n = 2\lambda^{1/2}$ in (3.8), then based on Theorems 1 and 2 we have

Corollary 1. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$. *If* $\lambda_n = O_p(\gamma_n)$, *then there exists a local minimizer* $(\hat{\alpha}^\top, \hat{\beta}^\top)^\top = (\hat{\alpha}^\top, \hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ *for the hierarchically penalized PL-AHM in (3.8)*

such that $\left\|\hat{\phi}-\phi^0\right\|+\left\|\hat{\beta}-\beta^0\right\|=O_p\left(\gamma_n\right)$*; if further* $p^{-1/2}\gamma_n^{-3/2}\lambda_n\to\infty$ *as* $n\to\infty$*, then* $\hat{\beta}_{\mathcal{C}}=0$ *with probability tending to 1.*

Comparing Corollary 1 with Theorem 2, we see that although the hierarchical penalty can effectively remove unimportant groups because $\hat{\beta}_{\mathcal{C}}=0$ with probability approaching to 1, it cannot effectively remove unimportant variables within the important groups as $\hat{\beta}_{\mathcal{D}}=0$ with probability tending to 1 may not hold. To tackle this limitation, we apply the adaptive idea used in Breiman (1995), Shen and Ye (2002), Zhang and Lu (2007), Zhao and Yu (2006), Zou (2006), Zou (2008), Wang et al. (2009), Liu et al. (2014), and others, which is to penalize different coefficients differently. To do so, we consider our objective function as

$$Q_n^*(\alpha,\beta)=L_n(\alpha,\beta)+\lambda_n\sum_{g=1}^{G}\left\{\sum_{j=1}^{p_g}w_{n,gj}|\beta_{gj}|\right\}^{1/2}, \tag{3.12}$$

where $w_{n,gj}$'s are pre-specified non-negative weights. The next theorem shows that, by controlling weights properly, the adaptive hierarchically penalized PL-AHM has the selection consistency as stated in Theorem 2.

Theorem 3. *Let us define*

$$w_{n,\max}^{\mathcal{A}}=\max\left\{w_{n,gj}:(g,j)\in\mathcal{A}\right\},\ \ w_{n,\min}^{\mathcal{A}}=\min\left\{w_{n,gj}:(g,j)\in\mathcal{A}\right\};$$

$$w_{n,\max}^{\mathcal{D}}=\max\left\{w_{n,gj}:(g,j)\in\mathcal{D}\right\},\ \ w_{n,\min}^{\mathcal{D}}=\min\left\{w_{n,gj}:(g,j)\in\mathcal{D}\right\}.$$

Let $\gamma_n=\sqrt{(K+p)/n}+K^{-d}$*. Under the same conditions as assumed in Theorem 1, if* $\gamma_n^{-1}\lambda_n w_{n,\max}^{\mathcal{A}}\left(w_{n,\min}^{\mathcal{A}}\right)^{-1/2}\to 0$*,* $\lambda_n\left(w_{n,\max}^{\mathcal{A}}\right)^2\left(w_{n,\min}^{\mathcal{A}}\right)^{-3/2}\to 0$*, and* $\gamma_n^{-1}\lambda_n w_{n,\min}^{\mathcal{D}}/(w_{n,\max}^{\mathcal{A}}+w_{n,\max}^{\mathcal{D}})^{1/2}\to\infty$ *as* $n\to\infty$*, there exists a local minimizer* $(\hat{\alpha}^\top,(\hat{\beta}_{\mathcal{A}}^\top,\hat{\beta}_{\mathcal{D}}^\top))^\top$ *of* $(\alpha^\top,(\beta_{\mathcal{A}}^\top,\beta_{\mathcal{D}}^\top))^\top$ *in (3.12) such that* $\left\|\hat{\phi}-\phi^0\right\|+\left\|\hat{\beta}-\beta^0\right\|=O_p\left(\gamma_n\right)$ *and* $\hat{\beta}_{\mathcal{D}}=0$ *with probability tending to 1.*

Finally, we specify our λ_n and the weights $w_{n,gj}$ that satisfy conditions in Theorem 3, which are given by the following corollary.

Corollary 2. *Let* $\gamma_n=\sqrt{(K+p)/n}+K^{-d}$ *and* $\tilde{\beta}_n$ *be an estimator such that,* $\left\|\tilde{\beta}_n-\beta^0\right\|=O_p(\gamma_n)$*. If* $\lambda_n=\gamma_n/\log(n)$ *and* $w_{n,gj}=1/|\tilde{\beta}_{n,gj}|^r$*, where* $r>0$*, then there exists a local*

minimizer $(\hat{\alpha}^\top, (\hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{D}}^\top))^\top$ *of* $(\alpha^\top, (\beta_{\mathcal{A}}^\top, \beta_{\mathcal{D}}^\top))^\top$ *in (3.12) such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$ *and* $\hat{\beta}_{\mathcal{D}} = 0$ *with probability tending to 1.*

In practice, we choose $(\tilde{\alpha}_n, \tilde{\beta}_n) = \arg\min_{\alpha,\beta} L_n(\alpha, \beta)$, the estimator from the unpenalized score function when p is diverging with n and $p < n$. From Corollary 1 and Corollary 2, we notice that the rates of convergence of the estimators are the same but the selection performance of the adaptive hierarchically penalized method is superior to that of the hierarchically penalized method, because the adaptive method possesses the individual variable selection consistency, while the non-adaptive method holds only group selection consistency.

3.4 Numerical Computations and Results

Direct minimization of $Q_n(\alpha, \beta)$ in 3.8 (or $Q_n^*(\alpha, \beta)$ in 3.12) is difficult because the penalty is not a convex function. Following Huang et al. (2009) and Liu et al. (2014), we formulate an easier equivalent minimization problem to solve it. We define

$$S_n(\alpha, \beta, \theta) = L_n(\alpha, \beta) + \sum_{g=1}^{G} \theta_g^{-1} \sum_{j=1}^{p_g} |\beta_{gj}| + \lambda \sum_{g=1}^{G} \theta_g, \tag{3.13}$$

where λ is a tuning parameter.

Proposition 1. *For* $\lambda_n = 2\sqrt{\lambda}$, $(\hat{\alpha}, \hat{\beta})$ *minimizes* $Q_n(\alpha, \beta)$ *in (3.8) if and only if* $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ *solves*

$$\min_{\alpha,\beta,\theta} S_n(\alpha, \beta, \theta) \text{ subject to } \theta \geq 0,$$

where $\theta \geq 0$ *means* $\theta_g \geq 0,\ g = 1, \ldots, G.$

This proposition can be considered as a special case of bridge group penalty used by Huang et al. (2009) in linear regression model and it is also used by Liu et al. (2014) in AHM. For fixed α, minimizing S_n in (3.13) with respect to (β, θ) performs variable selection at individual and group levels in the linear part. Hence, following Liu et al. (2014) and based

on Proposition 1, we propose the following iterative algorithm.

Step 1. Obtain an initial estimate $(\alpha^{(0)}, \beta^{(0)})$ by minimizing $L_n(\alpha, \beta)$. Let $s = 1$.

Step 2. Compute

$$\theta_g^{(s)} = \lambda^{-\frac{1}{2}} \left(\sum_{j=1}^{p_g} |\beta_{gj}^{(s-1)}| \right)^{\frac{1}{2}}, \ g = 1, \ldots, G.$$

Step 3. Compute

$$(\alpha^{(s)}, \beta^{(s)}) = \arg\min_{\alpha,\beta} \left\{ L_n(\alpha, \beta) + \sum_{g=1}^{G} (\theta_g^{(s)})^{-1} \sum_{j=1}^{p_g} |\beta_{gj}| \right\}.$$

Step 4. $s \leftarrow s + 1$, repeat steps 2-3 until convergence.

Step 3 is the major computational step and is efficiently solved using iterative coordinate descent algorithm (Friedman et al., 2007). Since at each step it decreases the non-negative objective function (3.13), this algorithm always converges.

For the adaptive hierarchically penalized method, we only need to replace $S_n(\alpha, \beta, \theta)$ by $S_n^*(\alpha, \beta, \theta)$, where

$$S_n^*(\alpha, \beta, \theta) = L_n(\alpha, \beta) + \sum_{g=1}^{G} \theta_g^{-1} \sum_{j=1}^{p_g} w_{gj} |\beta_{gj}| + \lambda \sum_{g=1}^{G} \theta_g.$$

Proposition 2. *For $\lambda_n = 2\sqrt{\lambda}$, $(\hat{\alpha}^*, \hat{\beta}^*)$ minimizes $Q_n^*(\alpha, \beta)$ in (3.12) if and only if $(\hat{\alpha}^*, \hat{\beta}^*, \hat{\theta}^*)$ solves*

$$\min_{\alpha,\beta,\theta} S_n^*(\alpha, \beta, \theta) \ \text{ subject to } \theta \geq 0,$$

where $\theta \geq 0$ means $\theta_g \geq 0, \ g = 1, \ldots, G$.

The computational procedure is similar to that for Proposition 1 with $\theta_g^{(s)}$ replaced by $\theta_g^{*(s)}$, where

$$\theta_g^{*(s)} = \left(\lambda^{-1} \sum_{j=1}^{p_g} w_{gj} |\beta_{gj}^{(s-1)}| \right)^{\frac{1}{2}}, \ g = 1, \ldots, G.$$

Tuning parameter selection

We select the tuning parameter λ_n in (3.12) by cross validation in the same way as done by Liu et al. (2014) in AHM. For a sequence of λ_n $(\lambda_{n1}, \ldots, \lambda_{nL})$, we apply the iterative algorithm to estimate (α, β) for each λ_n, $(\hat{\alpha}_{\lambda_{n1}}^\top, \hat{\beta}_{\lambda_{n1}}^\top)^\top, \ldots, (\hat{\alpha}_{\lambda_{nL}}^\top, \hat{\beta}_{\lambda_{nL}}^\top)^\top$. Next, we chose the tuning parameter λ_n from $\lambda_{n1}, \ldots, \lambda_{nL}$ using the M-fold cross-validation

$$\hat{\lambda}_n = \underset{\lambda_n \in \{\lambda_{n1}, \ldots, \lambda_{nL}\}}{\arg\min} \mathrm{CV}(\lambda) = \underset{\lambda_n \in \{\lambda_{n1}, \ldots, \lambda_{nL}\}}{\arg\min} \left\{ \sum_{m=1}^{M} L^{(m)}(\hat{\alpha}^{(-m)}(\lambda), \hat{\beta}^{(-m)}(\lambda)) \right\},$$

where

$$\mathrm{CV}(\lambda) = \sum_{m=1}^{M} L^{(m)}(\hat{\alpha}^{(-m)}(\lambda), \hat{\beta}^{(-m)}(\lambda))$$

is a score function of cross validation for λ_n; $L^{(m)}$ is the loss function in (3.5) using the m-th subset and $(\hat{\alpha}^{(-m)}(\lambda), \hat{\beta}^{(-m)}(\lambda))$ is the estimator evaluated without the m-th subset.

3.4.1 Simulation Studies

To evaluate the finite-sample performance of the hierarchically penalized method and its adaptive version, we conducted two simulation studies. We compared the results with those based on some existing individual variable selection methods such as LASSO, SCAD and adaptive LASSO (A-LASSO), these penalties have been used for variable selection in the PLMs (Ma and Du, 2012; Hu and Lian, 2013). In our simulation studies, we used `R` package `ahaz` for implementing computation for these penalties after linearizing the nonparametric functions $\phi(\cdot)$ using B-splines. For computation of our AHP group selection method in the PL-AHM, we constructed our `R` program where we also used some of the existing `R` packages, for example, `survival`, `ahaz`, and `fda`.

Five performance measures are used to compare these methods: number of true groups selected (TG), number of zero group selected (FG), number of true nonzero variables selected as nonzero (TP), number of true zero variables selected as nonzero (FP), and L_2- prediction error (PE) in the excess risk defined as $\left\| \left\{ \hat{\beta}^\top Z + \hat{\phi}_1(W_1) + \hat{\phi}_2(W_2) \right\} - \left\{ \beta^\top Z + \phi_1(W_1) + \phi_2(W_2) \right\} \right\|$. The optimal tuning parameter λ_n is chosen by five-fold cross-validation. As a benchmark, we

computed the oracle estimates, which are obtained by minimizing (3.5) for model (3.3) which includes only important variables and groups.

Variable selection is a computationally extensive procedure and can take a lot of time if convergence is slow. We used 'WestGrid' (https://www.westgrid.ca) to conduct our simulation studies which benefited us in terms of computational time. WestGrid is helping Compute Canada (https://www.computecanada.ca) to lead the acceleration of research and innovation by bringing together computing facilities, research data management services, and a network of technical experts to meet researchers need. It has multiple computing facilities where the researchers can send their computing codes and define parameters like computing time, memory, cores to be used based on the computational burden of their jobs. To conduct our simulation studies, we submitted all of our simulations parallelly to the computing facilities at the same time. On average, it took only an hour to conduct 500 simulations in WestGrid.

In Example 1, the number of groups is moderately large, the group sizes are equal and relatively large, and within each group the coefficients are either all nonzero or all zero. In Example 2, the group sizes vary and there are zero coefficients in a nonzero group. In each example, we set sample size $n = 200$ and baseline hazard function $h_0(t) = 1.0$. The censoring variable is generated from a uniform distribution over $[0, C_0]$, where C_0 is chosen to yield censoring rate = 30%. For each of these settings, we replicated 500 simulations.

Example 1. In this example, there are 7 groups in the linear part, each with 5 covariates, and two nonparametric functions. For the linear covariates, the covariate vector is $X^\top = (X_1^\top, \ldots, X_7^\top)$. The subvector of covariates that belong to the same group is $X_j^\top = (X_{5(j-1)+1}, \ldots, X_{5(j-1)+5});\ j = 1, \ldots, 7$. To generate the covariates $X_1, \ldots, X_{35}$, we first simulate 35 random variables $R_1, \ldots, R_{35}$ independently from the standard normal distribution. Then Z_j $(j = 1, \ldots, 7)$ are simulated from a multivariate normal distribution with mean zero and an AR(1) covariance structure such that $\text{cov}(Z_{j1}, Z_{j2}) = 0.4^{|j_1 - j_2|}$ for $j_1,\ j_2 = 1, \ldots, 7$. The covariates $X_1, \ldots, X_{35}$ are generated as $X_j = (Z_{gj} + R_j)/\sqrt{2}$ $(j = 1, \ldots, 35)$, where g_j is

the smallest integer greater than $(j-1)/5$ and the X_j's with the same value of g_j belong to the same group. This structure of correlation was considered in Huang et al. (2009). The nonparametric functions are $\phi_1(W_1) = W_1^2 - (25/12)$ and $\phi_2(W_2) = \exp(-W_2) - 2\sinh(5/2)/5$, where the covariates W's are sampled from $U(-2.5, 2.5)$. Such nonparametric functions were considered in Cui et al. (2013) in a nonparametric additive regression model. The event times in Example 1 are generated from an exponential distribution with a hazard rate given as follows:

$$h(t|X, W) = h_0(t) + \beta^\top X + \phi_1(W_1) + \phi_2(W_2),$$

where $\beta = (\underbrace{1.2, \ldots, 1.2}_{5}, \underbrace{3.6, \ldots, 3.6}_{5}, \underbrace{2.4, \ldots, 2.4}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5})^\top$.

To estimate nonparametric functions, we used B-splines, see details in Section 3.2 for centering of B-splines in general. Specifically, we center $\phi_1(W_1)$ and $\phi_2(W_2)$ such that $E\{\phi_1(W_1)\} = E\{\phi_2(W_2)\} = 0$. We approximated the nonlinear functions using cubic B-spline functions. We used data-driven method for choosing the regularization parameter λ. Lian et al. (2014) used 5 to 8 basis functions in their simulations and found similar results. They reported the results only for 6 basis functions. To ease the computational burden, we also choose $K = 6$ as the number of basis functions in B-splines. This choice of K is small enough to avoid overfitting and big enough to flexibly approximate the smooth functions (Gray, 1992; Cheng and Wang, 2011). In this example, there exists three important groups and all variables within each group are important. This example illustrates that the proposed group selection methods have the ability to identify important groups.

Example 2. In this experiment, the group size differs across groups and some groups have a mixture of important and unimportant variables. There are seven groups: three groups each of size 8 and four groups each of size 4. The covariate vector is $X^\top = (X_1^\top, \ldots, X_7^\top)$, where the seven subvectors of covariates are $X_j^\top = (X_{8(j-1)+1}, \ldots, X_{8(j-1)+8})$, for $j = 1, 2, 3$, and $X_j^\top = (X_{4(j-1)+13}, \ldots, X_{4(j-1)+16})$, for $j = 4, 5, 6, 7$. To generate the covariates $X_1, \ldots, X_{40}$, we first simulate Z_i $(i = 1, \ldots, 7)$ and $R_1, \ldots, R_{40}$ independently from the standard normal

distribution. For $j = 1, \ldots, 24$, let g_j be the largest integer less than $j/8 + 1$ and, for $j = 25, \ldots, 40$, let g_j be the largest integer less than $(j - 24)/4 + 1$. The covariates $X_1, \ldots, X_{40}$ are obtained as $X_j = (Z_{g_j} + R_j)/\sqrt{2}$ $(j = 1, \ldots, 40)$. The nonparametric functions are generated in the same way as of Example 1. Therefore, the corresponding coefficients in Example 2 are,

$$\beta = (\underbrace{1.2, \ldots, 1.2}_{8}, \underbrace{3.6, 3.4, 3.2, 3.0, 2.8, 0, 0, 0}_{8}, \underbrace{0, \ldots, 0}_{8}, \underbrace{2.4, 0, 0, 0}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{0, \ldots, 0}_{4})^\top.$$

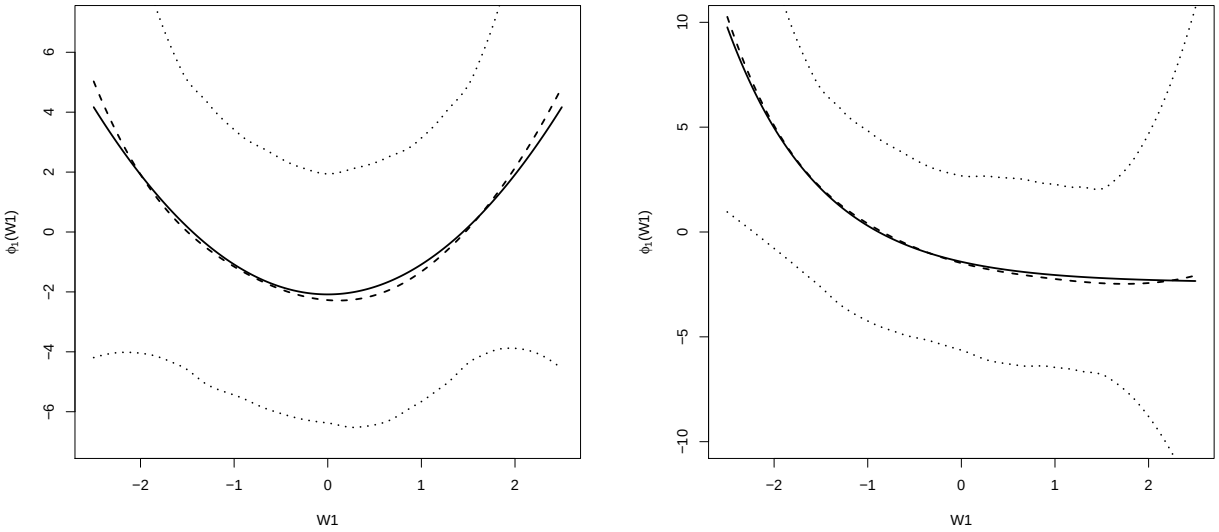
This example considers three important groups in a more complex structure than that in Example 1. These three groups represent three different settings: all variables within the group are important, many variables within the group are important and very few variables within the group are important, respectively.

Table 3.1: Simulation results with median and standard deviations (in parentheses) of L_2-PE, TG, FG, TP and FP over 500 simulations for Example 1

	L_2-PE	TG	FG	TP	FP
LASSO	124.02 (34.25)	3 (0.15)	2 (1.17)	12 (1.77)	3 (2.15)
SCAD	182.83 (46.05)	3 (0.55)	1 (0.90)	7 (2.34)	1 (1.16)
A-LASSO	118.84 (29.71)	3 (0.21)	2 (1.22)	9 (1.87)	2.5 (2.24)
HP	98.91 (26.75)	3 (0.18)	1 (0.88)	14 (1.35)	2 (2.83)
AHP	95.60 (25.73)	3 (0.23)	0 (0.80)	13 (1.50)	0 (2.44)
Oracle	94.00 (29.44)	3 (0.00)	NA	15 (0.00)	NA

Table 3.2: Simulation results with median and standard deviations (in parentheses) of L_2-PE, TG, FG, TP and FP over 500 simulations for Example 2

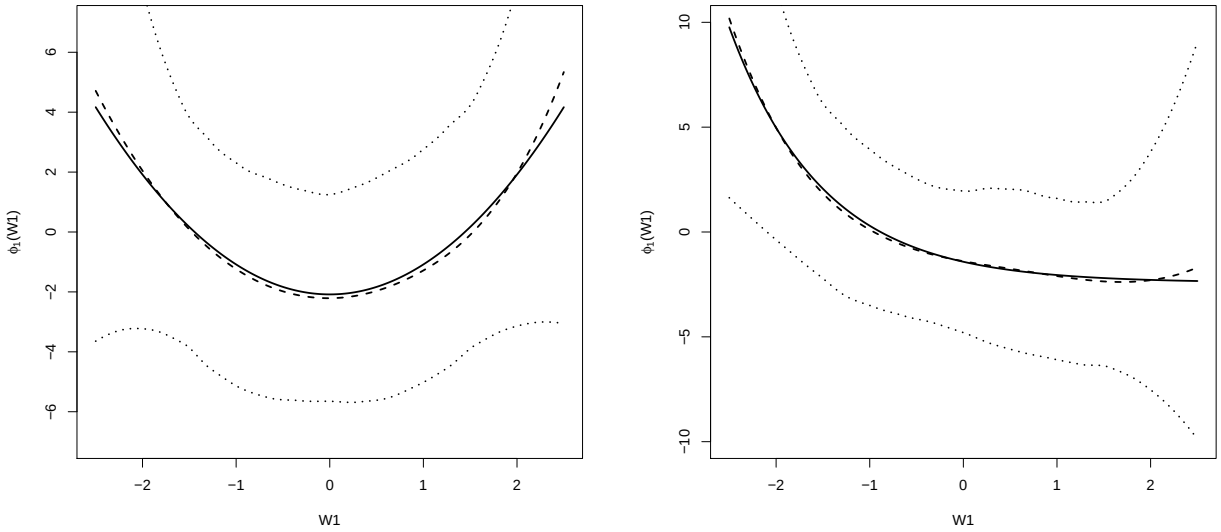
	L_2-PE	TG	FG	TP	FP
LASSO	101.50 (28.66)	3 (0.44)	2 (1.18)	10 (1.75)	4 (2.45)
SCAD	144.03 (36.94)	2 (0.59)	0 (0.80)	6 (2.24)	1 (1.25)
A-LASSO	97.16 (23.55)	3 (0.44)	2 (1.21)	8 (1.77)	3 (2.66)
HP	84.38 (22.96)	3 (0.52)	0 (0.76)	12 (1.40)	4 (3.25)
AHP	83.53 (20.81)	2 (0.50)	0 (0.58)	12 (1.40)	3 (2.72)
Oracle	78.32 (22.85)	3 (0.00)	NA	14 (0.00)	NA



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$

Figure 3.1: Estimation of $\phi(\cdot)$'s in Example 1: 95% point-wise confidence bands for $\phi(\cdot)$'s based on 500 replicates. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dot-dashed lines represent the 95% point-wise confidence bands based on 500 estimated values.

Tables 3.1 and 3.2 summarize the group and variable selection results for Examples 1 and 2 by using the LASSO, SCAD, A-LASSO, hierarchical (HP) and adaptive hierarchical penalties (AHP), respectively, where, the first three perform individual variable selection. From Table 3.1 we see the group variable selection methods perform significantly better than individual variable selection methods with lower L_2-prediction error and choose more important and less unimportant variables. Therefore, if there is known grouping structure available among the covariates, group selection methods are preferable over individual variable selection methods. Furthermore, the AHP method has the lowest L_2-prediction error and effectively removes more unimportant variables than the HP method. Hence, adaptive group selection performs superior over non-adaptive group selection method. In Table 3.2 with a complex grouping structure, the performance of the adaptive method is still better than the other methods. The fitted curves and 95% point-wise confidence bands for $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are shown in Figures 3.1 and 3.2 for Examples 1 and 2, respectively. It is evident that the



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$

Figure 3.2: Estimation of $\phi(\cdot)$'s in Example 2: 95% point-wise confidence bands for $\phi(\cdot)$'s based on 500 replicates. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dot-dashed lines represent the 95% point-wise confidence bands based on 500 estimated values.

average estimated curves capture the true curves very well and that the true curves lie in the 95% point-wise confidence bands.

3.4.2 Application

In this section, we illustrate application of our proposed variable selection methods in two real data sets.

3.4.2.1 Mantle Cell Lymphoma Data analysis

Mantle cell lymphoma (MCL) is a rare non-Hodgkin B-cell lymphoma which can be at an aggressive form or be more indolent in clinical representation (Rajabi and Sweetenham, 2015). Treatment is usually based on multiple factors including age, presence or absence of symptoms, and comorbidities. The median age at presentation is in the 60s and regardless of new assertive therapeutic approaches, the median overall survival for MCL patients is only

between 5-7 years.

Rosenwald et al. (2003) performed gene expression profiling to establish a molecular diagnosis of MCL, to clarify its pathogenesis, and to predict the length of survival of these patients. The dataset is available at http://llmpp.nih.gov/MCL. Based on established morphologic and immunophenotype criteria, 92 patients were classified as having MCL. The following variables were included in the data:

- Status: patient status at follow up (1 = death, 0 = censored);
- Time: time of follow-up in year;
- INK.ARFdeletion(X_1): deletions of INK4a/ARF (1 = yes, 0 = no);
- ATMdeletion(X_2): deletions of ATM (1 = yes, 0 = no);
- P.53deletion(X_3): deletions of P53 (1 = yes, 0 = no);
- CyclinD.1taqmanresults(X_4): cyclin D1 TaqMan result;
- BMIexpression(X_5): body mass index expression;
- Proliferation.average(X_6): proliferation signature averages.

Ma and Du (2012) performed variable selection in this data set using a partially linear accelerated failure time (AFT) regression model. They selected variables in the linear part without a grouping structure using iterated LASSO and estimated the nonlinear part using a sieve approach. They excluded the covariate Proliferation.average(X_6) from the analysis and included all other covariates X_1 – X_5 in the nonparametric part. In addition, they removed 7 records (patients) with missing covariates; with the rest 85 patients, the censoring rate was 29.4%.

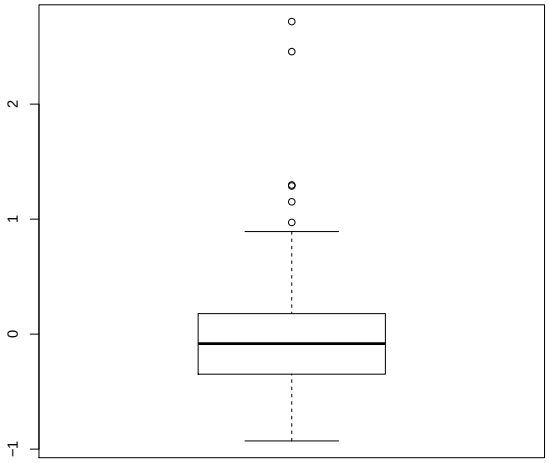
To perform group variable selection in the MCL data using a PL-AHM, we conducted some preliminary diagnosis of the data. Covariates X_1, X_2, and X_3 are binary variables, where

covariates X_4, X_5, and X_6 are continuous. Since variables belonging to the same group usually share some relationships among them, we tested for the significant correlations between the covariates. We tested the correlation between continuous variables by Pearson correlation coefficients; continuous and binary variables by Point-biserial correlation coefficient, and the association between binary variables by Fisher's exact test. The table below illustrates which variables share significant correlations where '✓' indicates significant correlations with the associated sample correlations in the parentheses:

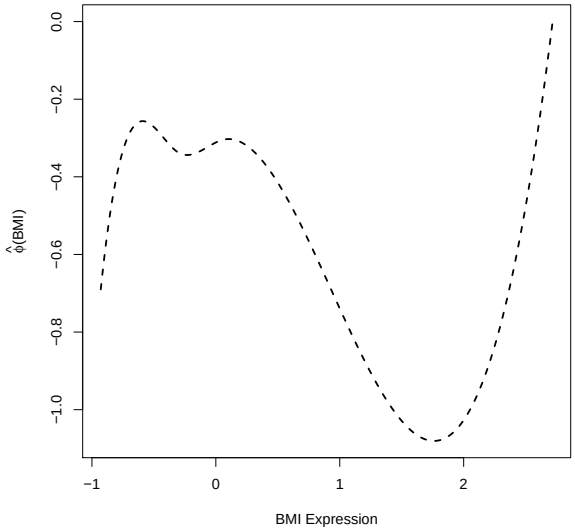
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	✓(1.00)					
X_2	✗ (0.16)	✓(1.00)				
X_3	✗ (0.10)	✗ (0.23)	✓(1.00)			
X_4	✓(0.28)	✗ (-0.10)	✗ (-0.08)	✓(1.00)		
X_5	✗ (-0.17)	✗ (-0.17)	✗ (0.00)	✗ (0.20)	✓(1.00)	
X_6	✓(0.50)	✗ (-0.05)	✓(0.23)	✓(0.41)	✗ (0.17)	✓(1.00)

From the above table we see that X_1, X_4, X_6 shares significant correlation among each other, therefore, we can consider them as a group. X_2 and X_5 do not have significant correlations with any other variables. Note that, X_6 also shares significant correlation with X_3 and they can be considered as a group as well. Thus, X_6 belongs to two overlapping groups; one with (X_1, X_4), another with X_3. However, in this chapter, we assumed covariates can only belong to one group. Therefore, we assign X_6 to the group with (X_1, X_4) based on the strength of the relationship.

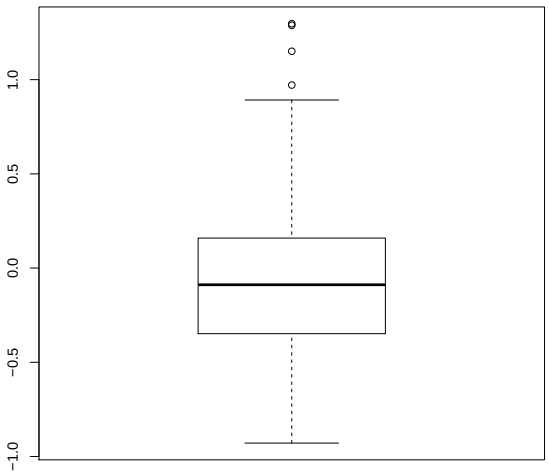
Thus, we have three groups in the linear part of our PL-AHM. Group 1 constitutes of (X_1, X_4, X_6), Group 2 has X_2 and Group 3 has X_3 in it. Similar to the common practice of putting discrete covariates in the linear part and continuous variables in the nonlinear part (Hu and Lian, 2013), we assigned the dichotomous variables in Group 2 and Group 3 in the linear part, and estimated the effect of the continuous variable X_5 on the survival of MCL patients nonparametrically.



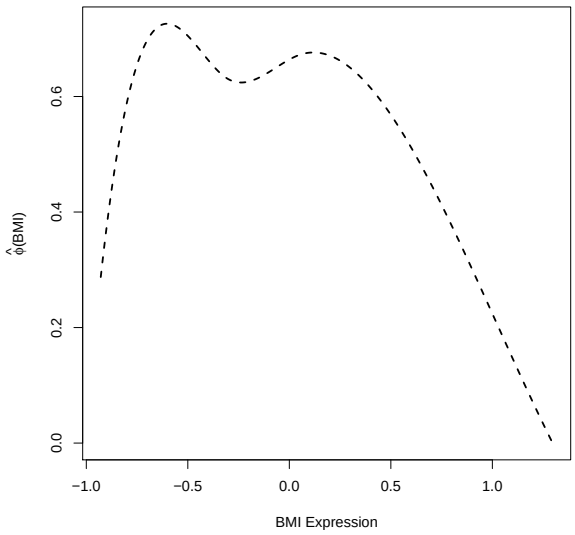
(a) Boxplot of X_5



(b) ϕ(BMI Expression)



(c) Boxplot of X_5 without 2 extreme values



(d) ϕ(BMI Expression) without 2 extreme values

Figure 3.3: Boxplot of BMI Expression and estimated curve of ϕ (BMI Expression) in the analysis of MCL data.

Table 3.3: Estimation results of MCL data

		n = 85				n=83			
Group	Covariates	LSE	LASSO	HP	AHP	LSE	LASSO	HP	AHP
G_1	INK.ARFdeletion (X_1)	0.04	0	0.06	0.06	0.06	0	0.07	0.07
	CyclinD.1taqmanresults (X_4)	0.54	0.23	0.50	0.50	0.54	0.30	0.49	0.49
	Proliferation.average (X_6)	0.46	0.36	0.45	0.45	0.45	0.38	0.43	0.43
G_2	ATMdeletion (X_2)	0.07	0	0	0	0.07	0	0	0
G_3	P.53deletion (X_3)	-0.02	0	0	0	-0.05	0	0	0

Table 3.3 shows the estimation and variable selection performance by four methods. The LSE is the least square estimates of the linear covariates using the pseudoscore method of Lin and Ying (1994), which uses the loss function (3.5), where we approximated the nonlinear function of X_5 using B-splines. For the full data set (n=85), we see that all of the variable selection methods discard Groups 2 and 3. In addition, LASSO also discards X_1. This is not surprising as from the simulations (Tables 3.1 and 3.2) we observed that LASSO tend to select less important variables by aggressively penalizing the variables. The estimates from LSE and AHP are very close in magnitude, HP and AHP are identical, where the LASSO estimates shrink more towards zero. Figures 3.3 (a) and (b) present the boxplot and nonlinear profile of BMIexpression(X_5), respectively. We found two extreme outliers which fall outside of upper inner fence ($Q_1 - 3 * IQR$) or upper outer fence ($Q_3 + 3 * IQR$) where Q_1, Q_3 and IQR are first quartile, third quartile and inter-quartile range, respectively. Figures 3.3 (c) and (d) show the boxplot and nonlinear profile of BMIexpression (X_5) after discarding the outliers. In addition, Table 3.3 also shows the variable selection performance when these two extreme values are omitted (n=83). From this comparison analysis, we see that the performance of variable selection is almost the same, but the estimated nonparametric function of X_5 is quite different in the right tail when the two large X_5 values are included. This may tell the investigators that large BMIexpression(X_5) values could increase the risk of death of MCL patients.

In our model, we included Proliferation.average(X_6) for variable selection which was not incorporated by Ma and Du (2012), and found that it has strong correlation with INK.ARFdeletion(X_1) and CyclinD.1taqmanresults(X_4). Both LASSO and AHP showed that X_6 has a strong effect, further investigation may provide additional knowledge of this effect on the survival probability of the MCL patients.

3.4.2.2 Wisconsin Prognostic Breast Cancer Data analysis

In the Wisconsin prognostic breast cancer (WPBC) data, each record represents follow-up data for one breast cancer case where cases represent invasive breast cancer or no evidence of distant metastases at the time of diagnosis. There are 198 breast cancer patients in total with no missing values. The data is available in **R** package "TH.data". The data set contains the following variables that we are interested in:

- status(δ): a factor with levels N (nonrecurrent) and R (recurrent);
- time: recurrence time (for status == "R") or disease-free time (for status == "N");
- mean-radius(X_1): radius (mean of distances from center to points on the perimeter) (mean);
- mean-texture(X_2): texture (standard deviation of gray-scale values) (mean);
- mean-perimeter(X_3): perimeter (mean);
- mean-area(X_4): area (mean);
- mean-smoothness(X_5): smoothness (local variation in radius lengths) (mean).
- mean-compactness(X_6): compactness (mean);
- mean-concavity(X_7): concavity (severity of concave portions of the contour) (mean);

- mean-concavepoints(X_8): concave points (number of concave portions of the contour) (mean);
- mean-symmetry(X_9): symmetry (mean);
- mean-fractaldim(X_{10}): fractal dimension (mean);
- tsize(X_{11}): diameter of the excised tumor in centimeters.

We set status = R (recurrent) as the event and then status = N (nonrecurrent) as the censoring indicator. The censoring rate of WPBC data was about 76%. To discover the grouping structure among the covariates, we computed the Pearson's correlation coefficient since all of the covariates are continuous. If the correlation is moderate-high (≥ 0.40) and positive, we assumed those variables belong to the same group. The table below depicts which variables share moderate to high, positive relationships with the associated sample correlations in the parentheses:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
X_1	✓(1.00)										
X_2	× (0.14)	✓(1.00)									
X_3	✓(1.00)	× (0.14)	✓(1.00)								
X_4	✓(0.99)	× (0.14)	✓(0.99)	✓(1.00)							
X_5	× (-0.05)	× (-0.17)	× (-0.01)	× (-0.06)	✓(1.00)						
X_6	× (-0.16)	× (-0.19)	× (-0.10)	×(-0.14)	✓(0.54)	✓(1.00)					
X_7	× (0.16)	× (-0.04)	× (0.24)	× (0.16)	✓(0.67)	✓(0.67)	✓(1.00)				
X_8	✓(0.47)	× (0.04)	✓(0.53)	✓(0.48)	✓(0.53)	✓(0.62)	✓(0.84)	✓(1.00)			
X_9	✓(0.66)	× (0.01)	✓(0.71)	✓(0.67)	✓(0.43)	✓(0.55)	✓(0.72)	✓(0.91)	✓(1.00)		
X_{10}	× (-0.42)	× (-0.15)	× (-0.35)	× (-0.40)	✓(0.60)	✓(0.75)	✓(0.74)	✓(0.45)	× (0.27)	✓(1.00)	
X_{11}	× (0.17)	× (0.03)	× (0.17)	× (0.18)	× (-0.15)	× (-0.08)	× (-0.06)	× (-0.01)	× (0.05)	× (-0.13)	✓(1.00)

From the above table, we observe two groups: Group 1 is consists of $(X_1, X_3, X_4, X_8, X_9)$ and Group 2 consists of $(X_5, X_6, X_7, X_8, X_9, X_{10})$. Similar to the MCL data, we have (X_8, X_9) who belong to both of the groups. Generally speaking, the magnitude of correlation between X_8 and Group 1 is higher than Group 2, whereas X_9 has higher magnitude of correlation

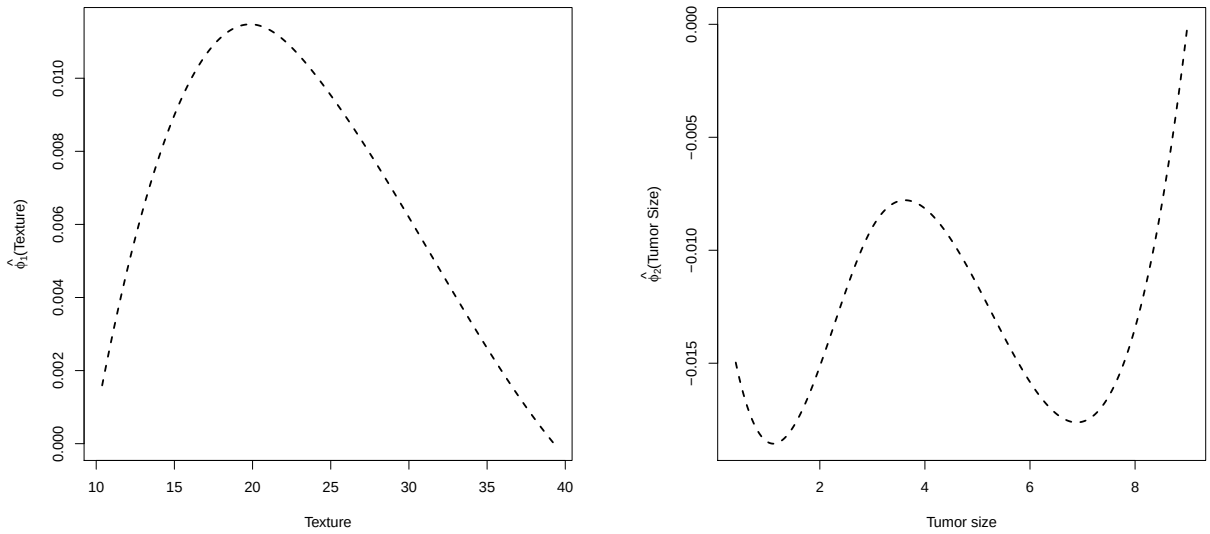
with Group 2 although the correlation between (X_9, X_{10}) is fairly small. Since the correlation between X_8 and X_9 is very strong (0.91), we decided to assign them in the same group and based on our observation above, we assigned them to Group 1. X_2 and X_{11} do not share moderate-high correlation with any other variables. Since they are the only variables in a single variable group and are continuous, we estimate them nonparametrically.

Table 3.4: Estimation results of WPBC data

Group	Covariates	LSE	LASSO	A-LASSO	HP	AHP
G_1	mean-radius (X_1)	-0.020	0	0	0	0
	mean-perimeter (X_3)	0.002	0	0	0	0
	mean-area (X_4)	8.06×10^{-5}	4.70×10^{-6}	0	0	0
	mean-concavepoints (X_8)	-0.006	0	0	0	0
	mean-symmetry (X_9)	-0.072	0.007	0	0	0.03
G_2	mean-smoothness (X_5)	-0.054	-0.022	0	-0.11	-0.11
	mean-compactness (X_6)	0.312	0	0.18	0.25	0.22
	mean-concavity (X_7)	0.053	0	0	0.11	0.09
	mean-fractaldim (X_{10})	-0.872	0	-0.41	-0.82	-0.72

The AHP method selects only one variable from Group 1 (X_9) and all of the variables from Group 2 (Table 3.4). In total, it selects five significant variables in predicting recurrence of breast cancer. HP only selects the variables in Group 2. LASSO and A-LASSO, on the other hand, select only three variables and two variables, respectively. It is surprising that LASSO does not select X_{10} which is quite big in magnitude in estimating the probability of breast cancer but selects X_4 which is very small in magnitude. To conclude, $X_5, X_6, X_7, X_9, X_{10}$ can significantly predict the recurrence of breast cancer where the effects of X_7 and X_9 are relatively small.

Figure 3.4 shows the estimated nonlinear effects of mean-texture(X_2) and tsize(X_{11}) on the recurrence of breast cancer. Overall mean-texture has a bell-shaped curve. Tumor size on the other hand has a more nonlinear profile in estimating recurrent breast cancer.



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$

Figure 3.4: Estimated curves of ϕ_1 (Texture) and ϕ_2 (Tumor size) in the analysis of WPBC data.

3.5 Concluding Remarks

In this chapter, we proposed a hierarchically penalized method for variable selection in the PL-AHM with diverging number of parameters. The hierarchically penalized method can effectively remove unimportant groups and select important variables within a group. However, the hierarchically penalized method tends to select more unimportant variables in important groups. To tackle this problem, the adaptive hierarchically penalized method is considered. We established the asymptotic convergence and selection consistency for the proposed estimators. Numerical studies indicate that the hierarchically penalized method and its adaptive version perform better than LASSO, SCAD and adaptive LASSO. Once variable selection is performed and a smaller set of important variable is selected, one can follow Afzal et al. (2017) to estimate the coefficients of PL-AHM where the asymptotic normality of the estimators is established. Our computation cost was somewhat high since the computation algorithm takes a while to converge; however, our estimators were precise

in terms of estimation accuracy and selection consistency at the cost of high computational time.

In applications, it is important to have the goodness of fit procedures available for assessing the model fit. Lin et al. (1993) proposed martingale-based residuals to graphically and numerically check the adequacy of the proportional hazards model with right censored data. Kim and Lee (1998) adopted two methods for model checking of the additive hazards model with right censored data by dividing the data into two groups and testing for the proportional hazards assumption to the additive hazards model to test the monotone departure from the additivity. One method is based on the martingale residuals and the other is based on the difference between weighted estimators of the excess risk. These model checking techniques were developed for the linear models. In our case, we can consider each B-spline basis function as a covariate in the model, then the proposed model becomes a linear model, and their methods can be applied to choose either the PL-AHM or PL-PHM in practice.

It should be noted that we did not address how to partition the covariates into linear and nonlinear parts. Few possible strategies are available in the literature to specify a PLM. Researchers might also follow the strategy suggested by Ma and Du (2012), where they put low dimensional clinical covariates in the nonparametric part and high dimensional gene expressions in the parametric part. We have not investigated the theoretical properties for ultra high dimensional data, i.e., $p \gg n$. Also, we only considered time-independent covariates where the groups are disjoint, i.e., there are no overlapping groups. All of these need further investigations in our future research.

In contrary to the frequentist approach, in a Bayesian framework, the variable selection problem can be viewed as the identification of nonzero regression parameters based on the posterior distributions. Bayesian models attempt to avoid the over-fitting problems of frequentist methods by basing predictions on modes of posterior distributions rather than estimators. In uncensored data, bi-level group variable selection using Bayesian selection

method has been investigated by Zhang et al. (2014), Xu and Ghosh (2015) and Mallick and Yi (2017). Faraggi and Simon (1998) is one of the first to consider Bayesian variable selection method for censored survival data where they performed individual variable selection in the Cox PH model. Later, Sha et al. (2006) conducted individual variable selection for analyzing microarray data with the AFT model and Lee et al. (2011) performed individual variable selection in the Cox PH model where the shrinkage prior is obtained through a scale mixture representation of normal and gamma distributions and the cumulative baseline hazard function is modeled as a priori by a gamma process. Recently, Lee et al. (2015) performed group variable selection in the Cox PH model. As it appears in the literature, no group variable selection has been investigated on the additive hazard model or its extensions using Bayesian methods and can be worthy of future research.

3.6 Appendix

In this Appendix, we prove the lemmas, theorems, and propositions that are presented in the previous sections and introduce some lemmas that will be used in the proofs.

We adopt the standard empirical process notation. For any measurable function f, we denote $\mathbb{P}_n(f)$ and $P(f)$ the expectations of f under the empirical measure $\mathbb{P}_n$ and the probability measure P, respectively. Let $\|\cdot\|_{P,r}$ denote the usual $L_r(P)$-norm. The "size" of a class $\mathcal{F}$ of functions is measured by the *bracketing number* $N_{[\,]}(\varepsilon, \mathcal{F}, L_r(P))$, the minimum number of ε-brackets in $L_r(P)$ needed to cover $\mathcal{F}$, and the *covering number* $N(\varepsilon, \mathcal{F}, L_2(Q))$, the minimum number of $L_2(Q)$-balls of radius ε needed to cover $\mathcal{F}$. The logarithms of the bracketing number and covering number are called *entropy with bracketing* and *entropy*, respectively. The *bracketing integral* and *uniform entropy integral* are defined as

$$J_{[\,]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P))}d\varepsilon, \text{ and}$$
$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}d\varepsilon,$$

respectively, where F is an envelope of $\mathcal{F}$, that is, $|f| \leq F$ for all $f \in \mathcal{F}$, and the supremum is taken over all probability measures Q with $\|F\|_{Q,2} > 0$. Interested readers can find more definitions and concepts about empirical processes in van der Vaart (1998, p. 127) and Kosorok (2007, p. 18).

The proofs of Theorems 1 and 2 use Lemmas 1 and 2, and the following lemmas:

Lemma 3. *Under Conditions (A1)-(A5), there exist constants $C, K_0 > 0$ such that*

$$P\left(\sup_{t\in[0,\tau]}\left|S_n^{(0)}(t) - s^{(0)}(t)\right| \geq Cn^{-1/2}(1+x)\right) \leq \exp(-K_0x^2), \tag{3.14}$$

$$P\left(\sup_{t\in[0,\tau]}\left|S_{nj}^{(1)}(t)[L] - s_j^{(1)}(t)[L]\right| \geq Cn^{-1/2}(1+x)|\Omega_V\right) \leq \exp(-K_0x^2/V^2), \tag{3.15}$$

$$P\left(\sup_{t\in[0,\tau]}\left|S_{nij}^{(2)}(t)[L] - s_{ij}^{(2)}(t)[L]\right| \geq Cn^{-1/2}(1+x)|\Omega_V\right) \leq \exp(-K_0x^2/V^4), \tag{3.16}$$

for all $x > 0$ and $i, j = 1, \ldots, (p + KQ)$, where $S_{nj}^{(1)}(\cdot)$ is the jth component of $S_n^{(1)}(\cdot)$, $S_{nij}^{(2)}(\cdot)$ is the (i, j)th entry of the matrix $S_n^{(2)}$, and Ω_V denotes the event that $\max_{j=1}^{p+KQ}|L_j| \leq V$ for $V > 0$.

Lemma 4. *Under Condition (A1)-(A5), there exist constants $C, M, K_1 > 0$ such that*

$$P(|D_{nij} - D_{ij}| \geq Cn^{-1/2}(1+x)|\Omega_V) \leq M\exp(-K_1\frac{x^2 \wedge n}{V^4})$$

for all $x > 0$ and $i, j = 1, \ldots, (p + KQ)$, where D_{nij} and D_{ij} are the (i, j)-th entries of the matrices D_n and D, respectively.

The proofs of Lemmas 1 and 2 closely follow those of Wang et al. (2009), respectively.

Proof of Lemma 1. Let $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$ denote the criterion that we would like to minimize in equation (3.6), let $Q^\dagger(\lambda, \alpha, \gamma, \theta)$ denote the corresponding criterion in equation (3.7), and let $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ denote a local minimizer of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$. We will prove that $(\hat{\alpha}^\dagger = \hat{\alpha}^*, \hat{\gamma}_g^\dagger = \lambda_\gamma\hat{\gamma}_g^*, \hat{\theta}_{(g)}^\dagger = \hat{\theta}_{(g)}^*/\lambda_\gamma)$ is a local minimizer of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$.

Replacing $\gamma^* = \gamma^\dagger/\lambda_\gamma$ and $\theta^* = \theta^\dagger\lambda_\gamma$ in (3.6), we immediately have $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta) = Q^\dagger(\lambda, \alpha, \lambda_\gamma\gamma, \theta/\lambda_\gamma)$. Since $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ is a local minimizer of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$, therefore, by the definition of local minimizer there exists $\delta > 0$ such that if $(\alpha^{'}, \gamma^{'}, \theta^{'})$ satisfies $|\alpha^{'} - \hat{\alpha}^*| + |\gamma^{'} - \hat{\gamma}^*| + |\theta^{'} - \hat{\theta}^*| < \delta$, then $Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*) \leq Q^*(\lambda_\gamma, \lambda_\theta, \alpha^{'}, \gamma^{'}, \theta^{'})$. We choose $\delta^{'}$ such that $\delta^{'}/\min(\lambda_\gamma, 1/\lambda_\gamma) \leq \delta/2$. Then, $\min(\lambda_\gamma, 1/\lambda_\gamma) \leq 1$ and $\delta^{'} \leq \min(\lambda_\gamma, 1/\lambda_\gamma)\delta/2 \leq \delta/2$. Thus, for any $(\alpha^{''}, \gamma^{''}, \theta^{''})$ satisfying $|\alpha^{''} - \hat{\alpha}^\dagger| + |\gamma^{''} - \hat{\gamma}^\dagger| + |\theta^{''} - \hat{\theta}^\dagger| < \delta^{'} \leq \delta/2$, we have, $|\alpha^{''} - \hat{\alpha}^\dagger| = |\alpha^{''} - \hat{\alpha}^*| \leq \delta/2$. Also,

$$\begin{aligned}\left|\frac{\gamma^{''}}{\lambda_\gamma} - \hat{\gamma}^*\right| + \left|\lambda_\gamma\theta^{''} - \hat{\theta}^*\right| &\leq \frac{\lambda_\gamma\left|\frac{\gamma^{''}}{\lambda_\gamma} - \hat{\gamma}^*\right| + \frac{1}{\lambda_\gamma}\left|\lambda_\gamma\theta^{''} - \hat{\theta}^*\right|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} = \frac{\left|\gamma^{''} - \lambda_\gamma\hat{\gamma}^*\right| + \left|\theta^{''} - \frac{\hat{\theta}^*}{\lambda_\gamma}\right|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} \\ &= \frac{|\gamma^{''} - \hat{\gamma}^\dagger| + |\theta^{''} - \hat{\theta}^\dagger|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} < \frac{\delta^{'}}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} \leq \frac{\delta}{2}.\end{aligned}$$

Therefore, $\left|\alpha^{''} - \hat{\alpha}^*\right| + \left|\gamma^{''}/\lambda_\gamma - \hat{\gamma}^*\right| + \left|\lambda_\gamma\theta^{''} - \hat{\theta}^*\right| < \delta/2 + \delta/2 = \delta$. Hence,

$$Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*) \leq Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}^{''}, \hat{\gamma}^{''}/\lambda_\gamma, \lambda_\gamma\hat{\theta}^{''}),$$

which gives us

$$Q^\dagger(\lambda, \hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger) \leq Q^\dagger(\lambda, \hat{\alpha}^{''}, \hat{\gamma}^{''}, \hat{\theta}^{''}).$$

So, $(\hat{\alpha}^\dagger = \hat{\alpha}^*, \hat{\gamma}^\dagger = \lambda_\gamma\hat{\gamma}^*, \hat{\theta}^\dagger = \hat{\theta}^*/\lambda_\gamma)$ is a local minimizer of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$.

Similarly, we can prove that for any local minimizer $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$, there is a corresponding local minimizer $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$ such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}^*_g\hat{\theta}^*_{g_j} = \hat{\gamma}^\dagger_g\hat{\theta}^\dagger_{g_j}$. □

Proof of Lemma 2. Suppose $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local minimizer of (3.7). Let $\hat{\beta}$ satisfy $\hat{\beta}_{gj} = \hat{\gamma}_g\hat{\theta}_{gj}$, then, $\hat{\theta}_{(g)} = \hat{\beta}_{(g)}/\hat{\gamma}_g$. It is trivial that $\hat{\gamma}_g = 0$ if and only if $\hat{\theta}_{(g)} = 0$. Hence, if $\hat{\gamma}_g \neq 0$, then $|\hat{\beta}_{(g)}| \neq 0$.

Let (α, β) be fixed at $(\hat{\alpha}, \hat{\beta})$. Then minimizing $Q^\dagger(\lambda, \alpha, \gamma, \theta)$ in (3.7) only depends on the penalty. For some g with $|\hat{\beta}_{(g)}| \neq 0$, the corresponding penalty term is $\gamma_g + \lambda\sum_{j=1}^{p_g}|\hat{\beta}_{gj}|/\gamma_g$, which is minimized at $\hat{\gamma}_g = (\lambda|\hat{\beta}_{(g)}|)^{1/2}$.

Let $Q(\lambda, \alpha, \beta)$ be the corresponding criterion to be minimized in equation (3.8). By Lemma 1, the local minimizers $\hat{\alpha}$ of α in (3.6) and (3.7) are the same, so we only need to consider other parameters, e.g., β, and fix α at $\hat{\alpha}$ in both (3.6) and (3.7). We first show that $(\hat{\alpha}, \hat{\beta})$ is a local minimizer of $Q(\lambda, \alpha, \beta)$, i.e., there exists a $\delta^{'} > 0$ such that if $|\Delta\alpha| + |\Delta\beta| < \delta^{'}$, then $Q(\lambda, \hat{\alpha}, \hat{\beta}) \leq Q(\lambda, \hat{\alpha} + \Delta\alpha, \hat{\beta} + \Delta\beta)$. Particularly, taking $\Delta\alpha = 0$, it becomes $|\Delta\beta| < \delta^{'}$, then $Q(\lambda, \hat{\alpha}, \hat{\beta}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta)$. Denote $\Delta\beta = \Delta\beta^{(1)} + \Delta\beta^{(2)}$, where $\Delta\beta^{(1)}_{(g)} = 0$ if $|\hat{\beta}_{(g)}| = 0$ and $\Delta\beta^{(2)}_{(g)} = 0$ if $|\hat{\beta}_{(g)}| \neq 0$. We thus, have $|\Delta\beta| = |\Delta\beta^{(1)} + \Delta\beta^{(2)}| = |\Delta\beta^{(1)}| + |\Delta\beta^{(2)}|$.

We first show $Q(\lambda, \hat{\alpha}, \hat{\beta}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$ for some $\delta^{'}$. We already have $\hat{\gamma}_g = (\lambda|\hat{\beta}_{(g)}|)^{1/2}$ and $\hat{\theta}_{(g)} = \hat{\beta}_{(g)}/\hat{\gamma}_g$ if $|\hat{\gamma}_g| \neq 0$, and $\hat{\theta}_{(g)} = 0$ if $|\hat{\gamma}_g| = 0$. Let $\hat{\gamma}^{'}_g = (\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2}$ and $\hat{\theta}^{'}_{(g)} = (\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)})/\hat{\gamma}^{'}_g$ if $|\hat{\gamma}_g| \neq 0$, and let $\hat{\gamma}^{'}_g = 0$ and $\hat{\theta}^{'}_{(g)} = 0$ if $|\hat{\gamma}_g| = 0$. Then we have $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}^{'}, \hat{\theta}^{'}) = Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$ and $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta}) = Q(\lambda, \hat{\alpha}, \hat{\beta})$. Hence, we only need to show $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta}) \leq Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}^{'}, \hat{\theta}^{'})$. As $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local minimizer of $Q^{\dagger}(\lambda, \alpha, \gamma, \theta)$, for fixed $\hat{\alpha}$, there exists a δ such that for any $(\gamma^{'}, \theta^{'})$ satisfying $|\gamma^{'} - \hat{\gamma}| + |\theta^{'} - \hat{\theta}| < \delta$, we have $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta}) \leq Q^{\dagger}(\lambda, \hat{\alpha}, \gamma^{'}, \theta^{'})$. Let $a = \min\left\{|\hat{\beta}_{(g)}| : |\beta_{(g)}| \neq 0,\ g = 1, \ldots, G\right\}, b = \max\left\{|\hat{\beta}_{(g)}| : |\beta_{(g)}| \neq 0,\ g = 1, \ldots, G\right\}$ and $\delta^{'} < a/2$. It is seen that,

$$
\begin{aligned}
\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| - |\hat{\beta}_{(g)}|\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left|(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2})^2 - (|\hat{\beta}_{(g)}|^{1/2})^2\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left|(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2})(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2})\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2}\right| &\leq \frac{\left|\Delta\beta^{(1)}_{(g)}\right|}{|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2}}.
\end{aligned}
$$

Since when $\min_g\left\{|\hat{\beta}_{(g)}|\right\} = a \neq 0$, and when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'} < a/2$, we have

$$
|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| \geq |\hat{\beta}_{(g)}| - |\Delta\beta^{(1)}_{(g)}| \geq a - \frac{a}{2} = \frac{a}{2} > 0,
$$

and

$$
|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2} \geq \left(\frac{a}{2}\right)^{1/2} + a^{1/2} = (2^{-1/2} + 1)a^{1/2} \geq 2^{1/2}a^{1/2} = (2a)^{1/2}.
$$

Therefore,

$$\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2}\right| \leq \frac{|\Delta\beta^{(1)}_{(g)}|}{(2a)^{1/2}}.$$

Hence,

$$|\hat{\gamma}_g^{'} - \hat{\gamma}_g| = \left|(\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2} - (\lambda|\hat{\beta}_{(g)}|)^{1/2}\right| \leq \frac{\lambda|\Delta\beta^{(1)}_{(g)}|}{(2\lambda a)^{1/2}}.$$

Next, if $|\hat{\gamma}_g| = 0$, then $\hat{\theta}^{'}_{(g)} = \hat{\theta}_{(g)} = 0$, and $|\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)}| = 0$. If $|\hat{\gamma}_g| \neq 0$, then

$$\begin{aligned}
\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)} &= \frac{(\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)})}{\hat{\gamma}_g^{'}} - \frac{\hat{\beta}_{(g)}}{\hat{\gamma}_g} \\
&= \frac{\hat{\beta}_{(g)}\hat{\gamma}_g + \Delta\beta^{(1)}_{(g)}\hat{\gamma}_g - \hat{\beta}_{(g)}\hat{\gamma}_g^{'}}{\hat{\gamma}_g^{'}\hat{\gamma}_g} \\
&= \frac{\hat{\beta}_{(g)}[\hat{\gamma}_g - \hat{\gamma}_g^{'}] + \Delta\beta^{(1)}_{(g)}\hat{\gamma}_g}{\hat{\gamma}_g^{'}\hat{\gamma}_g}. \qquad (3.17)
\end{aligned}$$

We already have $\hat{\beta}_{(g)} \leq b$ and $|\hat{\gamma}_g^{'} - \hat{\gamma}_g| \leq \lambda|\Delta\beta^{(1)}_{(g)}|/(2\lambda a)^{1/2}$. Consider

$$\hat{\gamma}_g^{'}\hat{\gamma}_g = (\lambda|\hat{\beta}_{(g)}|)^{1/2}(\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2}.$$

Since $|\hat{\gamma}_g| = (\lambda|\hat{\beta}_{(g)}|)^{1/2} \geq \lambda^{1/2}a^{1/2}$, when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'}$ and $\delta^{'} < a/2$, if $\hat{\gamma}_g \neq 0$, then $|\hat{\beta}_{(g)}| \neq 0$, $\Delta\beta^{(2)}_{(g)} = 0$, it implies, $|\Delta\beta^{(1)}_{(g)}| \leq |\Delta\beta^{(1)}| < \delta \Rightarrow |\Delta\beta^{(1)}_{(g)}| < \delta^{'} < a/2$ and $|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| \geq |\hat{\beta}_{(g)}| - |\Delta\beta^{(1)}_{(g)}| \geq a - a/2 = a/2 > 0$. Therefore, $|\hat{\gamma}_g^{'}| = (\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2} \geq \lambda^{1/2}(a/2)^{1/2}$ and $|\hat{\gamma}_g^{'}\hat{\gamma}_g| \geq \lambda^{1/2}a^{1/2}\lambda^{1/2}(a/2)^{1/2} = \lambda a 2^{-1/2}$. From (3.17) we have,

$$\begin{aligned}
|\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)}| &\leq \frac{|\hat{\beta}_{(g)}|}{|\hat{\gamma}_g^{'}\hat{\gamma}_g|}|\hat{\gamma}_g^{'} - \hat{\gamma}_g| + |\Delta\beta^{(1)}_{(g)}|\frac{|\hat{\gamma}_g|}{|\hat{\gamma}_g||\hat{\gamma}_g^{'}|} \\
&\leq \frac{b\lambda|\Delta\beta^{(1)}_{(g)}|}{(2\lambda a)^{1/2}(\lambda a 2^{-1/2})} + |\Delta\beta^{(1)}_{(g)}|\frac{1}{\lambda^{1/2}(a/2)^{1/2}} \\
&\leq \left[\frac{b\lambda}{(2\lambda a)^{1/2}(\lambda a)2^{-1/2}} + \frac{1}{(\lambda a/2)^{1/2}}\right]|\Delta\beta^{(1)}_{(g)}| \\
&= |\Delta\beta^{(1)}_{(g)}|\left[\frac{1}{(\lambda a/2)^{1/2}} + \frac{b}{a(\lambda a)^{1/2}}\right].
\end{aligned}$$

Therefore, we are able to choose a $\delta^{'} > 0$ satisfying $\delta^{'} < a/2$ such that $|\hat{\gamma}_g^{'} - \hat{\gamma}_g| + |\hat{\theta}^{'}_g - \hat{\theta}_g| < \delta$ when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'}$. Hence we have $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta}) \leq Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}^{'}, \hat{\theta}^{'})$ due to the local minimality, that is, $Q(\lambda, \hat{\alpha}, \hat{\beta}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$.

Next we show $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)})$. This is trivial when $\Delta\beta^{(2)} = 0$. If $\Delta\beta^{(2)} \neq 0$, then $\Delta\beta^{(1)} = 0$ and we have

$$Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)}) - Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)}) = (\Delta\beta^{(2)})^{\top} \frac{\partial L_n(\hat{\alpha}, \beta^*)}{\partial \beta} + 2\sum_{g=1}^{G} (\lambda |\Delta\beta_{(g)}^{(2)}|)^{1/2},$$

where β^* is a vector between $\hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)}$ and $\hat{\beta} + \Delta\beta^{(1)}$. Since $|\Delta\beta^{(2)}| < \delta^{'}$, for a small enough $\delta^{'}$, the second term in the above equality dominates the first term, hence we have $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)})$. Thus we have shown that there exists a $\delta^{'} > 0$ such that if $|\Delta\beta| < \delta^{'}$, then $Q(\lambda, \hat{\alpha}, \hat{\beta}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta)$, which implies that $\hat{\beta}$ is a local minimizer of $Q(\lambda, \hat{\alpha}, \beta)$.

Similarly, we can prove that if $(\hat{\alpha}, \hat{\beta})$ is a local minimizer of $Q(\lambda, \alpha, \beta)$, then $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local minimizer of $Q^{\dagger}(\lambda, \alpha, \gamma, \theta)$, where $\hat{\gamma}_g = (\lambda |\hat{\beta}_{(g)}|)^{1/2}$ and $\hat{\theta}_{(g)} = \hat{\beta}_{(g)} / \hat{\gamma}_g$ if $|\hat{\beta}_{(g)}| \neq 0$, and $\hat{\gamma}_g = 0$ and $\hat{\theta}_{(g)} = 0$ if $|\hat{\beta}_{(g)}| = 0$. □

Proof of Lemma 3. The proof follows that of Lemma A.2 in Lin and Lv (2013). We will only prove (3.15) and the other two inequalities follow similarly. Denote $B_{nj} = \sup_{t \in [0,\tau]} \left| S_{nj}^{(1)}(t)[L] - s_j^{(1)}(t)[L] \right|$. As given in Theorem 9 of Massart (2000), we want to apply a functional Hoeffding-type inequality. To do that, we need to control the term $E(B_{nj})$. And, to control the term $E(B_{nj})$, at first we will show that the class of functions $\{Y(t)L_j : t \in [0, \tau]\}$ has bounded uniform entropy integral.

Since a function of bounded variation can be expressed as the difference of two increasing functions, from Lemma 9.10 of Kosorok (2007), it follows that $\mathcal{L}_j \equiv \{L_j\}$ is a VC-hull class associated with a VC class of index 2. Therefore, by Corollary 2.6.12 of van der Vaart and Wellner (1997), the entropy of $\mathcal{L}_j$ satisfies $\log N(\epsilon \|F\|_{Q,2}, \mathcal{L}_j, L_2(Q)) \leq K'(1/\epsilon))$ for some constant $K' > 0$, and hence, $\mathcal{L}_j$ has the uniform entropy integral

$$J(1, \mathcal{L}_j, L_2) \leq \int_0^1 \sqrt{K'(1/\epsilon)} d\epsilon < \infty.$$

Next, by Example 19.16 of van der Vaart (1998), $\mathcal{Y} \equiv \{Y(t) : t \in [0, \tau]\}$ is a VC class and thus, has bounded uniform entropy integral. Therefore, by Theorem 9.15 of Kosorok (2007),

the product of two VC-hull classes, $\mathcal{YL}_j$, also has bounded uniform entropy integral.

Finally, an application of Lemma 19.38 of van der Vaart (1998) gives

$$E(B_{nj}) \leq C' n^{-1/2} J(1, \mathcal{YL}_j, L_2) \left\| F \right\|_{P,2} \leq C n^{-1/2}$$

for some constants $C', C > 0$, where the envelope F is taken as $\sup_{t \in [0,\tau]} Y(t)|L_j|$. From Theorem 9 of Massart (2000), we have

$$P\left(B_{nj} \geq C n^{-1/2}(1+x)|\Omega_V\right) \leq P\left(B_{nj} \geq E(B_{nj}) + C n^{-1/2} x|\Omega_V\right) \leq \exp(-K_0 x^2/V^2)$$

for some constant $K_0 > 0$, which concludes the proof. □

Proof of Lemma 4. We closely follow the proof of Lemma A.4 in Lin and Lv (2013) to prove Lemma 4. Let us write

$$\begin{aligned} D_{nij} - D_{ij} &= \int_0^\tau \left\{ S_{nij}^{(2)}(t)[L] - s_{ij}^{(2)}(t)[L] \right\} dt - \int_0^\tau \left\{ \frac{S_{ni}^{(1)}(t)[L] S_{nj}^{(1)}(t)[L]}{S_n^{(0)}(t)} - \frac{s_i^{(1)}(t)[L] s_j^{(1)}(t)[L]}{s^{(0)}(t)} \right\} dt \\ &\equiv T_{n1} + T_{n2}. \end{aligned}$$

From (3.16) in Lemma 3 we have $P(|T_{n1}| \geq C n^{-1/2}(1+x)|\Omega_V) \leq \exp(-K_0 x^2/V^4)$. To bound T_{n2}, let us write

$$\begin{aligned} &\frac{S_{ni}^{(1)}(t)[L] S_{nj}^{(1)}(t)[L]}{S_n^{(0)}(t)} - \frac{s_i^{(1)}(t)[L] s_j^{(1)}(t)[L]}{s^{(0)}(t)} = \frac{S_{nj}^{(1)}(t)[L]}{S_n^{(0)}(t)} \left\{ S_{ni}^{(1)}(t)[L] - s_i^{(1)}(t)[L] \right\} \\ &+ \frac{s_i^{(1)}(t)[L]}{S_n^{(0)}(t)} \left\{ S_{nj}^{(1)}(t)[L] - s_j^{(1)}(t)[L] \right\} - \frac{s_i^{(1)}(t)[L] s_j^{(1)}(t)[L]}{S_n^{(0)}(t) s^{(0)}(t)} \left\{ S_n^{(0)}(t) - s^{(0)}(t) \right\}. \end{aligned}$$

It suffices to consider the case where $\sup_{t \in [0,\tau]} |S_n^{(0)}(t) - s^{(0)}(t)| \leq \delta$ and $\sup_{t \in [0,\tau]} |S_{nj}^{(1)}(t)[L] - s_j^{(1)}(t)[L]| \leq \delta$ for some constant $\delta > 0$ and $j = 1, \ldots, (p + KQ)$. Hence, from the above representation and (3.14) and (3.15) in Lemma 3, it follows that $P(|T_{n2}| \geq C n^{-1/2}(1 + x)|\Omega_V) \leq 3 \exp(-K_0 x^2/V^2)$. Combining the bounds for T_{n1} and T_{n2} yields the desired inequality and completes the proof. □

The proofs of Propositions 1 and 2 follow closely those of Huang et al. (2009), respectively.

Proof of Proposition 1. We have $\min_{\alpha,\beta,\theta} S_n(\alpha,\beta,\theta) = \min_{\alpha,\beta} \hat{S}_n(\alpha,\beta)$, where $\hat{S}_n(\alpha,\beta) = \min_{\theta}\{S_n(\alpha,\beta,\theta) : \theta \geq 0\}$. For any (α,β), the minimizer of $S_n(\alpha,\beta,\theta)$ with respect to θ depends on β only, which is given by

$$\hat{\theta}(\beta) \equiv \arg\min_{\theta} \left\{ \sum_{g=1}^{G} \theta_g^{-1} \sum_{j=1}^{p_g} |\beta_j| + \lambda \sum_{g=1}^{G} \theta_g \right\}.$$

Solving this minimization problem, we have

$$\hat{\theta}_g(\beta) = \left(\lambda^{-1} \sum_{j=1}^{p_g} |\beta_{gj}| \right)^{\frac{1}{2}}, \ g = 1, \ldots, G.$$

Writing $\hat{S}_n(\alpha,\beta) = S_n(\alpha,\beta,\hat{\theta}(\beta))$ and substituting the expression $\hat{\theta}_g(\beta)$ into $S_n(\alpha,\beta,\hat{\theta}(\beta))$, we get,

$$\begin{aligned}
\hat{S}_n(\alpha,\beta) &= L_n(\alpha,\beta) + \sum_{g=1}^{G} \left(\lambda^{-1} \sum_{j=1}^{p_g} |\beta_{gj}| \right)^{\frac{-1}{2}} \sum_{j=1}^{p_g} |\beta_j| + \lambda \sum_{g=1}^{G} \left(\lambda^{-1} \sum_{j=1}^{p_g} |\beta_{gj}| \right)^{\frac{1}{2}} \\
&= L_n(\alpha,\beta) + \lambda^{1/2} \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_g} |\beta_{gj}|^{\frac{1}{2}} + \sum_{j=1}^{p_g} |\beta_{gj}|^{\frac{1}{2}} \right\} \\
&= L_n(\alpha,\beta) + 2\lambda^{1/2} \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_g} |\beta_{gj}| \right\}^{\frac{1}{2}} \\
&= L_n(\alpha,\beta) + \lambda_n \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_g} |\beta_{gj}| \right\}^{\frac{1}{2}}.
\end{aligned}$$

Therefore, $\hat{S}_n(\alpha,\beta) = Q_n(\alpha,\beta)$. □

Proof of Proposition 2. We have $\min_{\alpha,\beta,\theta} S_n^*(\alpha,\beta,\theta) = \min_{\alpha,\beta} \hat{S}_n^*(\alpha,\beta)$, where $\hat{S}_n^*(\alpha,\beta) = \min_{\theta}\{S_n^*(\alpha,\beta,\theta) : \theta \geq 0\}$. Similar to the proof of Proposition 1, for any (α,β),

$$\hat{\theta}^*(\beta) \equiv \arg\min_{\theta^*} \left\{ \sum_{g=1}^{G} \theta_g^{*^{-1}} \sum_{j=1}^{p_g} w_{gj}|\beta_{gj}| + \lambda \sum_{g=1}^{G} \theta_g^* \right\}.$$

Solving this minimization problem, we have

$$\hat{\theta}_g^*(\beta) = \left(\lambda^{-1} \sum_{j=1}^{p_g} w_{gj}|\beta_{gj}| \right)^{\frac{1}{2}}, \ g = 1, \ldots, G.$$

Writing $\hat{S}_n^*(\alpha,\beta) = S_n^*(\alpha,\beta,\hat{\theta}^*(\beta))$ and substituting the expression $\theta_g^*(\beta)$ into $S_n^*(\alpha,\beta,\hat{\theta}^*(\beta))$, we get,

$$\begin{aligned}
\hat{S}_n^*(\alpha,\beta) &= L_n(\alpha,\beta) + \sum_{g=1}^{G}\left(\lambda^{-1}\sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|\right)^{\frac{-1}{2}} \sum_{j=1}^{p_g} w_{gj}|\beta_j| + \lambda\sum_{g=1}^{G}\left(\lambda^{-1}\sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|\right)^{\frac{1}{2}} \\
&= L_n(\alpha,\beta) + \lambda^{1/2}\sum_{g=1}^{G}\left\{\sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|^{\frac{1}{2}} + \sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|\right\}^{\frac{1}{2}} \\
&= L(\alpha,\beta) + 2\lambda^{1/2}\sum_{g=1}^{G}\left\{\sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|\right\}^{\frac{1}{2}} \\
&= L(\alpha,\beta) + \lambda_n\sum_{g=1}^{G}\left\{\sum_{j=1}^{p_g} w_{gj}|\beta_{gj}|\right\}^{\frac{1}{2}}.
\end{aligned}$$

Thus, $\hat{S}_n^*(\alpha,\beta) = Q_n^*(\alpha,\beta)$. □

Proof of Theorem 1. Let $\alpha^0 = ({\alpha_1^0}^\top,\ldots,{\alpha_Q^0}^\top)^\top$ be a QK dimensional vector that satisfies $\left\|\phi_j^0 - {\alpha_j^0}^\top B_j\right\|_\infty = O(K^{-d}), 1 \le j \le Q$. Then, $\left\|\phi^0 - {\alpha^0}^\top B\right\|_\infty = O(K^{-d})$ and $\left\|\phi^0 - {\alpha^0}^\top B\right\| = O(K^{-d})$ since Q is fixed. Such approximation rates are possible due to our smoothness assumption (A2) and well known approximation properties of B-spline (De Boor, 1978).

Let $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$ and $u \in \mathbb{R}^{QK+p}$ where $u^\top = (u_1^\top, u_2^\top)$, u_1 is a QK-vector, and u_2 is a p-vector. To prove Theorem 1, we first show that $\left\|\hat{\phi} - {\alpha^0}^\top B\right\| = O_p(\gamma_n)$, and $\left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$ where $\hat{\phi} = {\hat{\alpha}^0}^\top B$. Then it is sufficient to show that for any $\epsilon > 0$, there exists a constant C such that

$$P\left\{\inf_{\|u\|=C} Q_{n,gen}((\alpha^0,\beta^0) + \gamma_n u) > Q_{n,gen}(\alpha^0,\beta^0)\right\} \ge 1-\epsilon, \tag{3.18}$$

when n is big enough. This implies that with probability of at least $1-\epsilon$, there exists a local minimizer in the ball $\{(\alpha^0,\beta^0) + \gamma_n u : \|u\| \le C\}$. Hence, there exists a local minimizer such that $\left\|\hat{\phi} - {\alpha^0}^\top B\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$.

Since p_{λ_n} satisfies conditions (3.10) and (3.11), we have,

$$
\begin{aligned}
&Q_{n,gen}((\alpha^0,\beta^0)+\gamma_n u)-Q_{n,gen}(\alpha^0,\beta^0)\\
&=\left\{L_n((\alpha^0,\beta^0)+\gamma_n u)-L_n(\alpha^0,\beta^0)\right\}\\
&+\sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0+\gamma_n u_{2,g1}\right|,\ldots,\left|\beta_{gs_g}^0+\gamma_n u_{2,gs_g}\right|,\left|\beta_{g(s_g+1)}^0+\gamma_n u_{2,g(s_g+1)}\right|,\ldots,\left|\beta_{gp_g}^0+\gamma_n u_{2,gp_g}\right|\right)\right.\\
&\left.-p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|,\ldots,\left|\beta_{gs_g}^0\right|,\left|\beta_{gs_g+1)}^0\right|,\ldots,\left|\beta_{gp_g}^0\right|\right)\right\}\\
&+\sum_{g=s+1}^{G}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0+\gamma_n u_{2,g1}\right|,\ldots,\left|\beta_{gp_g}^0+\gamma_n u_{2,gp_g}\right|\right)-p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|,\ldots,\left|\beta_{gp_g}^0\right|\right)\right\}\\
&\geq\left\{L_n((\alpha^0,\beta^0)+\gamma_n u)-L_n(\alpha^0,\beta^0)\right\}\\
&+\sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0+\gamma_n u_{2,g1}\right|,\ldots,\left|\beta_{gs_g}^0+\gamma_n u_{2,gs_g}\right|,\left|\beta_{g(s_g+1)}^0+\gamma_n u_{2,g(s_g+1)}\right|,\ldots,\left|\beta_{gp_g}^0+\gamma_n u_{2,gp_g}\right|\right)\right.\\
&\left.-p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|,\ldots,\left|\beta_{gs_g}^0\right|,\left|\beta_{g(s_g+1)}^0\right|,\ldots,\left|\beta_{gp_g}^0\right|\right)\right\}\\
&\geq\left\{L_n((\alpha^0,\beta^0)+\gamma_n u)-L_n(\alpha^0,\beta^0)\right\}\\
&+\sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0+\gamma_n u_{2,g1}\right|,\ldots,\left|\beta_{gs_g}^0+\gamma_n u_{2,gs_g}\right|,0\right)-p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|,\ldots,\left|\beta_{gs_g}^0\right|,0\right)\right\}\\
&=A+B.
\end{aligned}
$$

For A, denote $\omega=(\alpha,\beta)$, and $\hat{\omega}=(\hat{\alpha},\hat{\beta})$ be the estimator of $\omega^0=(\alpha^0,\beta^0)$. By Taylor expansion at $\gamma_n=0$, we have

$$
\begin{aligned}
L_n(\omega^0+\gamma_n u)&=L_n(\omega^0)+\gamma_n u^\top\left\{D_n\omega^0-d_n\right\}+\frac{1}{2}\gamma_n^2 u^\top D_n u,\\
A&=L_n(\omega^0+\gamma_n u)-L_n(\omega^0)\\
&=\gamma_n u^\top U_n(\omega^0)+\frac{1}{2}\gamma_n^2 u^\top D_n u\\
&\triangleq A_1+A_2. \qquad (3.19)
\end{aligned}
$$

By $\left\|\phi^0 - \alpha^{0^\top} B\right\|_\infty = O(K^{-d})$, we have

$$\begin{aligned}
U_n(\omega^0) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}\left\{dN_i(t) - Y_i(t)(\alpha^{0^\top}, \beta^{0^\top})L_i dt\right\} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}\left[dN_i(t) - Y_i(t)\left\{\phi^{0^\top}(W_i) + \beta^{0^\top}X_i\right\}dt\right. \\
&\left. + Y_i(t)\left\{\phi^{0^\top}(W_i) - \alpha^{0^\top}Z_i\right\}dt\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}dM_i(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}Y_i(t)\left\{\phi^0(W_i) - \alpha^{0^\top}Z_i\right\}dt \\
&= -n^{-1}\sum_{i=1}^{n}\int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}dM_i(t) + O_p(K^{-d}) \\
&= -n^{-1}\xi_n + O_p(K^{-d}),
\end{aligned}$$

where $\xi_n = \sum_{i=1}^n \int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}dM_i(t)$. Direct algebraic calculations show that, $E\left\{\|\xi_n\|^2\right\} = E\left\{tr\left(\xi_n^\top \xi_n\right)\right\} = tr\left\{E\left(\xi_n^\top \xi_n\right)\right\} = tr\left\{E\left(\|\xi_n\|^2\right)\right\}$. Let,

$$\xi_n = \sum_i \int_0^\tau \left\{L_i - \bar{L}_n(t)\right\}dM_i(t) = \sum_i \int_0^\tau H_i(t)dM_i(t), \text{ where } H_i(t) = L_i - \bar{L}_n(t).$$

Since ξ_n is a martingale integral, we have $E(\xi_n|\mathcal{F}_t^-) = 0$, where $\mathcal{F}_t^-$ denotes the past up to the beginning of the small time interval $[t, t+dt)$, and

$$\begin{aligned}
V(\xi_n|\mathcal{F}_t^-) &= E(\xi_n^{\otimes 2}|\mathcal{F}_t^-) \\
&= E(\xi_n\xi_n^\top|\mathcal{F}_t^-) \\
&= E\sum_i \int_0^\tau \text{Var}\left\{H_i(t)dM_i(t)|\mathcal{F}_t^-\right\} \\
&= E\sum_i \int_0^\tau H_i(t)^{\otimes 2} d\left\langle M\right\rangle(t) \\
&= E\int_0^\tau \sum_i \left\{L_i - \bar{L}_n(t)\right\}^{\otimes 2}\Lambda_i(t)dt,
\end{aligned}$$

where $\Lambda_i(t) = Y_i(t)\left\{h_0(t) + \phi^0(W) + X^\top\beta^0\right\}$. We can show that

$$\begin{aligned}
\sum_i \left\{L_i - \bar{L}_n(t)\right\}^{\otimes 2} Y_i(t) &= \sum_i \left\{L_i^{\otimes 2} Y_i(t)\right\} - \sum_i \left\{\bar{L}_n(t)\right\}^{\otimes 2} Y_i(t) \\
&\leq \sum_i \left\{L_i^{\otimes 2} Y_i(t)\right\}.
\end{aligned}$$

Then, $V(\xi_n|\mathcal{F}_t^-) \leq E\int_0^\tau \sum_i L_i^{\otimes 2}\Lambda_i(t)dt$. Assume $\sup_{t,W,X}|h_0(t)+\phi^0(W)+\beta^{0\top}X| \leq \tilde{M}$. Therefore,

$$\begin{aligned} E\left\{\|\xi_n\|^2\right\} &= \mathrm{tr}\left[E\left\{\int_0^\tau \sum_i \left(L_i - \bar{L}_n(t)\right)^{\otimes 2}\Lambda_i(t)dt\right\}\right] \\ &\leq \tilde{M}\left[E\left\{\int_0^\tau \sum_i \left(L_i - \bar{L}_n(t)\right)^{\otimes 2} Y_i(t)dt\right\}\right] \\ &\leq nE\left\{\mathrm{tr}L_i^{\otimes 2}Y_i(t)\right\}. \end{aligned}$$

Therefore, by Condition (A4), we have,

$$\|\xi_n\| = O_p\left(\sqrt{n(K+p)}\right),$$

and

$$\left\|U(\omega^0)\right\| = O_p(\sqrt{(K+p)/n} + K^{-d}) = O_p(\gamma_n). \tag{3.20}$$

Consequently, from (3.19),

$$A_1 = \gamma_n O_p(\gamma_n)\|u\| = O_p(\gamma_n^2)\|u\|.$$

Next, for A_2, we have,

$$A_2 = \frac{1}{2}\gamma_n^2 u^\top D_n u = \frac{1}{2}\gamma_n^2\left[u^\top Du + u^\top\left\{(D_n - D)\right\}u\right].$$

By Lemma 4, $\|D_n - D\|_2 = o_p(1)$. Since D is positive definite and its eigen values are bounded away from zero and infinity,

$$A_2 = 1/2\gamma_n^2(1+o_p(1))\|u\|^2.$$

For the penalty part, by Taylor expansion of the penalty function we have,

$$
\begin{aligned}
B &= \sum_{g=1}^{s} \left\{ p_{\lambda_n}^{(g)} \left(|\beta_{g1}^0 + \gamma_n u_{2,g1}|, \ldots, |\beta_{gp_g}^0 + \gamma_n u_{2,gs_g}|, 0 \right) - p_{\lambda_n}^{(g)} \left(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|, 0 \right) \right\} \\
&= \sum_{g=1}^{s} \left\{ \sum_{j=1}^{s_g} \frac{\partial p_{\lambda_n}^{(g)} \left(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0| \right)}{\partial |\beta_{g_j}|} \operatorname{sgn}(\beta_{gj}^0) \gamma_n u_{2,gj} \right. \\
&\left. + \frac{1}{2} \sum_{i=1}^{s_g} \sum_{j=1}^{s_g} \frac{\partial^2 p_{\lambda_n}^{(g)} \left(|\beta_{g1}^0|, \ldots, |\beta_{gp_k}^0| \right)}{\partial |\beta_{gi}| \partial |\beta_{gj}|} \gamma_n^2 u_{2,gi} u_{2,gj} \right\} + o_p \left\{ \gamma_n^2 (u_{2,g1}^2 + \cdots + u_{2,gs_g}^2) \right\} \\
&\leq q_1^{1/2} a_n \gamma_n \|u_2\| + \frac{1}{2} \gamma_n^2 b_n \|u_2\|^2 + o_p(\gamma_n^2 \|u_2\|^2) \\
&= q_1^{1/2} O_p(\gamma_n) \gamma_n \|u_2\| + o_p(\gamma_n^2 \|u_2\|^2) \text{ as } \quad b_n \to 0 \\
&= q_1^{1/2} O_p(\gamma_n^2) \|u_2\| + o_p(\gamma_n^2 \|u_2\|^2) \\
&\triangleq B_1 + B_2,
\end{aligned}
$$

where $q_1 = \sum_{g=1}^{s} s_g$. We can see that, by choosing a sufficiently large C, A_2 dominates A_1, B_1, B_2 uniformly in $\|u\| = C$. Thus, we have shown that $\|\hat{\alpha} - \alpha^0\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$. Then, $\left\|\hat{\phi} - \alpha^{0^\top} B\right\| = O_p(\gamma_n)$ and $\left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$. By $\left\|\phi^0 - \alpha^{0^\top} B\right\|_\infty = O(K^{-d})$ and the triangle inequality, we obtain

$$
\begin{aligned}
\left\|\hat{\phi} - \phi^0\right\| &\leq \left\|\hat{\phi} - \alpha^{0^\top} B\right\| + \left\|\alpha^{0^\top} B - \phi^0\right\| \\
&= O_p(\gamma_n) + O(K^{-d}) \\
&= O_p(\gamma_n).
\end{aligned}
$$

Hence, $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$. □

Proof of Theorem 2. Here we will prove the sparsity: $\text{pr}(\hat{\beta}_{\mathcal{D}} = 0) \to 1$ as $n \to \infty$. By Taylor expansion, we have

$$
\begin{aligned}
\frac{\partial Q_{n,gen}(\hat{\alpha}, \hat{\beta})}{\partial \beta_{gj}} &= \frac{\partial L_n(\hat{\alpha}, \beta^0)}{\partial \beta_{gj}} + \frac{\partial p_{\lambda_n}^{(g)} \left(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}| \right)}{\partial |\beta_{gj}|} \operatorname{sgn}(\hat{\beta}_{gj}) \\
&= C_1 + C_2.
\end{aligned} \tag{3.21}
$$

Using the result from (3.20), we have $|C_1| = O_p(\gamma_n)$. It follows from the definition of $\hat{\beta}_{gj}$ that, if $\hat{\beta}_{gj} \neq 0$,

$$\begin{aligned}\frac{\partial Q_{n,gen}(\hat{\alpha},\hat{\beta})}{\partial \beta_{gj}} &= O_p(\gamma_n) + \frac{\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|}\mathrm{sgn}(\hat{\beta}_{gj}) \\ &= \gamma_n\left\{O_p(1) + \gamma_n^{-1}\frac{\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|}\mathrm{sgn}(\hat{\beta}_{gj})\right\}. \qquad (3.22)\end{aligned}$$

Next, we show that there is a contradiction in (3.22) if $pr\left\{\hat{\beta}_{\mathcal{D}} = 0\right\}$ does not tend to 1 when $n \to \infty$, then, there exist $(g,j) \in \mathcal{D}$, such that $\hat{\beta}_{gj} \neq 0$. By the condition given in Theorem 2, that is, $\gamma_n^{-1}\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)/\partial|\beta_{gj}| \to \infty$ with probability tending to 1 as $n \to \infty$, for an arbitrary $\epsilon > 0$, when n is large we have

$$\frac{\partial Q_{n,gen}(\hat{\alpha},\hat{\beta})}{\partial \beta_{gj}} > 0,\ 0 < \hat{\beta}_{gj} < \epsilon, \quad \frac{\partial Q_{n,gen}(\hat{\alpha},\hat{\beta})}{\partial \beta_{gj}} < 0,\ -\epsilon < \hat{\beta}_{gj} < 0.$$

This is in conflict with $\partial Q_{n,gen}(\hat{\alpha},\hat{\beta})/\partial\beta_{gj} = 0$ and results in a contradiction when $\hat{\beta}_{gj} \neq 0$. Therefore, $\mathrm{pr}(\hat{\beta}_{\mathcal{D}} = 0) \to 1$ as $n \to \infty$. □

Proof of Corollary 1. We only need to check that the conditions in Theorem 1 hold for the penalty function $p^{(g)}_{\lambda_n}(|\beta_{(g)}|) = \lambda_n(|\beta_{g1}| + \cdots + |\beta_{gp_g}|)^{1/2},\ g = 1, ..., G$.

For $\beta_{gj} \in \mathcal{A}$, i.e., $\beta^0_{gj} \neq 0$, we have,

$$\begin{aligned}a_n &= \max_{(g,j)\in\mathcal{A}} \frac{\partial p_{\lambda_n}(|\beta^0_{g1}|,\ldots,|\beta^0_{gp_g}|)}{\partial|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{\partial \lambda_n(|\beta^0_{g1}| + \cdots + |\beta^0_{gp_g}|)^{1/2}}{\partial|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{2}\lambda_n(|\beta^0_{g1}| + \cdots + |\beta^0_{gp_g}|)^{-1/2} \\ &\leq \frac{1}{2}\lambda_n M^{-1/2} = O_p(\gamma_n),\end{aligned}$$

and

$$\begin{aligned} b_n &= \max_{(g,j)\in\mathcal{A}} \left| \frac{\partial^2 p_{\lambda_n}(|\beta_{g1}^0|,\ldots,|\beta_{gp_g}^0|)}{\partial|\beta_{gj}|^2} \right| \\ &= \max_{(g,j)\in\mathcal{A}} \left| \frac{\partial^2 \lambda_n(|\beta_{g1}^0|+\cdots+|\beta_{gp_g}^0|)^{1/2}}{\partial|\beta_{gj}|^2} \right| \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{4}\lambda_n(|\beta_{g1}^0|+\cdots+|\beta_{gp_g}^0|)^{-3/2} \\ &\leq \frac{1}{4}\lambda_n M^{-3/2} \to 0, \end{aligned}$$

where $M = \min_g(|\beta_{g1}^0|+\cdots+|\beta_{gp_g}^0|)$. Therefore, the rate of convergence follows from Theorem 1.

For sparsity, suppose there exists $(g,j) \in \mathcal{C}$ for which $\hat{\beta}_{gj} \neq 0$. Since for all $(g,j) \in \mathcal{C}$, $\beta_{gj}^0 = 0;\ j = 1,\ldots,p_g$, we have

$$\begin{aligned} \gamma_n^{-1}\frac{\partial p_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|} &= \gamma_n^{-1}\frac{\partial\lambda_n(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}}{\partial|\beta_{gj}|} \\ &= \frac{\gamma_n^{-1}\lambda_n}{2(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}}. \end{aligned}$$

According to the first conclusion of Corollary 1, there exists a γ_n^{-1} consistent local minimizer $\hat{\beta} = (\hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ for the non-adaptive hierarchically penalized loss function (3.8), which implies $\left\|\hat{\beta}_{\mathcal{C}} - \beta_{\mathcal{C}}^0\right\| \leq M^*\gamma_n$ or for $\hat{\beta}_{gj} \neq 0$, we have $|\hat{\beta}_{gj} - \beta_{gj}^0| = |\hat{\beta}_{gj}| \leq M^*\gamma_n$ for some constant M^*. Thus,

$$\begin{aligned} \frac{\gamma_n^{-1}\lambda_n}{2(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}} &\geq \frac{\gamma_n^{-1}\lambda_n}{2(M^*\gamma_n+\cdots+M^*\gamma_n)^{1/2}} \\ &= \frac{1}{2M^{*1/2}} \times \frac{\gamma_n^{-1}\lambda_n\gamma_n^{-1/2}}{p_g^{1/2}} \\ &\geq \frac{\gamma_n^{-3/2}\lambda_n p^{-1/2}}{2M^{*1/2}} \quad (\text{since } p \geq p_g). \end{aligned}$$

Therefore, for $\gamma_n^{-3/2}\lambda_n p^{-1/2} \to \infty$ when $n \to \infty$, we have, $\gamma_n^{-1}\partial\lambda_n(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}/\partial|\beta_{gj}| \to \infty$, which results in a contradiction when $\hat{\beta}_{gj} \neq 0$. So, for all $(g,j) \in \mathcal{C}$, $\hat{\beta}_{gj} = 0$. □

Proof of Theorem 3. We only need to check that the conditions in Theorem 1 hold for the penalty function $p_{\lambda_n}^{(g)}(|\beta_{(g)}|) = \lambda_n(w_{n,g1}|\beta_{g1}|+\cdots+w_{n,gp_g}|\beta_{gp_g}|)^{1/2}$.

For $\beta_{gj} \in \mathcal{A}$, i.e., $\beta_{gj}^0 \neq 0$, we have,

$$\begin{aligned} a_n &= \max_{(g,j)\in\mathcal{A}} \frac{\partial p_{\lambda_n}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\partial |\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{\partial \lambda_n (w_{n,g1}|\beta_{g1}^0| + \cdots + w_{n,gp_g}|\beta_{gp_g}^0|)^{1/2}}{\partial |\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{2}\lambda_n w_{n,gj}(w_{n,g1}|\beta_{g1}^0| + \cdots + w_{n,gp_g}|\beta_{gp_g}^0|)^{-1/2} \\ &\leq \frac{1}{2}\lambda_n w_{n,\max}^{\mathcal{A}} \left(w_{n,\min}^{\mathcal{A}}\right)^{-1/2} M^{-1/2} = O_p(\gamma_n), \end{aligned}$$

and

$$\begin{aligned} b_n &= \max_{(g,j)\in\mathcal{A}} \left| \frac{\partial^2 p_{\lambda_n}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\partial |\beta_{gj}|^2} \right| \\ &= \max_{(g,j)\in\mathcal{A}} \left| \frac{\partial^2 \lambda_n (w_{n,g1}|\beta_{g1}^0| + \cdots + w_{n,gp_g}|\beta_{gp_g}^0|)^{1/2}}{\partial |\beta_{gj}|^2} \right| \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{4}\lambda_n (w_{n,gj})^2 (w_{n,g1}|\beta_{g1}^0| + \cdots + w_{n,gp_g}|\beta_{gp_g}^0|)^{-3/2} \\ &\leq \frac{1}{4}\lambda_n \left(w_{n,\max}^{\mathcal{A}}\right)^2 \left(w_{n,\min}^{\mathcal{A}}\right)^{-3/2} M^{-3/2} \to 0, \end{aligned}$$

where $M = \min_g(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)$. Thus, the consistency follows from Theorem 1.

Next, we prove the sparsity. Assume $\hat{\beta}_{gj}$ is a local minimizer of $Q_n^*(\alpha, \beta)$ in (3.12) with $\left\|\hat{\beta}_{gj} - \beta_{gj}^0\right\| = O_p(\gamma_n)$. We can find a constant M^*, such that $|\hat{\beta}_{gj}| \leq M^*$ for all (g, j) with probability tending to 1. Then for $(g, j) \in \mathcal{D}$, i.e., $\beta_{gj}^0 = 0$, we have

$$\begin{aligned} \gamma_n^{-1} \frac{\partial p_{\lambda_n}\left(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}|\right)}{\partial |\beta_{gj}|} &= \frac{\gamma_n^{-1}\lambda_n w_{n,gj}}{2(w_{n,g1}|\hat{\beta}_{g1}| + \cdots + w_{n,gp_g}|\hat{\beta}_{gp_g}|)^{1/2}} \\ &\geq \frac{\gamma_n^{-1}\lambda_n w_{n,\min}^{\mathcal{D}}}{2M^{*1/2}(w_{n,\max}^{\mathcal{A}} + w_{n,\max}^{\mathcal{D}})^{1/2}}. \end{aligned}$$

Therefore, when $\gamma_n^{-1}\lambda_n w_{n,\min}^{\mathcal{D}}/(w_{n,\max}^{\mathcal{A}} + w_{n,\max}^{\mathcal{D}})^{1/2} \to \infty$ as $n \to \infty$, then $\hat{\beta}_{gj} = 0$ with probability approaching to 1. Hence, by Theorem 2, we have $\mathrm{pr}(\hat{\beta}_{\mathcal{D}} = 0) \to 1$. □

Proof of Corollary 2. We only need to verify that $w_{n,g_j} = |\tilde{\beta}_{n,g_j}|^{-r}$ satisfy the conditions in Theorem 3. Let $A = \max_{g,j}\left\{\beta_{g_j}^0\right\}$ and $B = \min_{g,j}\left\{\beta_{g_j}^0 : \beta_{g_j}^0 \neq 0\right\}$. Then by the

consistency of $\tilde{\beta}_n$, $w^{\mathcal{A}}_{n,\max} \to B^{-r}$ and $w^{\mathcal{A}}_{n,\min} \to A^{-r}$. Thus, if $\lambda_n = \gamma_n/\log(n)$, we have $\gamma_n^{-1}\lambda_n w^{\mathcal{A}}_{n,\max}\left(w^{\mathcal{A}}_{n,\min}\right)^{-1/2} \to 0$ and $\lambda_n\left(w^{\mathcal{A}}_{n,\max}\right)^2\left(w^{\mathcal{A}}_{n,\min}\right)^{-3/2} \to 0$, as $n \to \infty$.

For each (g, j) with $\beta^0_{n,gj} = 0$, we have $\tilde{\beta}_{gj} = O_p(\gamma_n)$. Therefore, $w^{\mathcal{D}}_{n,\min}/(w^{\mathcal{A}}_{n,\max} + w^{\mathcal{D}}_{n,\max})^{1/2} = O_p(\gamma_n^{-1/2})$. Thus, for $\lambda_n = \gamma_n/\log(n)$, we have $\gamma_n^{-1}\lambda_n w^{\mathcal{D}}_{n,\min}/(w^{\mathcal{A}}_{n,\max}+w^{\mathcal{D}}_{n,\max})^{1/2} \to \infty$. □

Chapter 4

Hierarchically Penalized Partially Linear Proportional Hazards Model with a Diverging Number of Parameters

4.1 Introduction

The proportional hazards model (Cox, 1972) is probably the oldest and the most popular model in survival analysis and has been widely used to study the relationship between multiple covariates and censored event times. The model assumes that the hazard function of a subject related to the covariates X is given by

$$h(t|X) = h_0(t) \exp\left(\beta^\top X\right), \tag{4.1}$$

where $h_0(t)$ is a completely unspecified baseline hazard function, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients, and $X = (X_1, \ldots, X_p)^\top$ is a p-dimensional covariate vector.

In practice, it is possible that not all covariates are linearly related to the hazard, i.e., some of them have a nonlinear effect on the hazard. Considering a purely parametric model is too stringent in this case, while a purely nonparametric model suffers from the so called "curse of dimensionality". Partially linear models (PLMs) in such cases combine the flexibility of nonparametric modeling with the parsimony and easy interpretability of parametric modeling, and avoids the curse of dimensionality of a purely nonparametric model (O'Sullivan, 1993; Fan et al., 1997).

To incorporate the nonlinear effect of a covariate, we consider a partially linear proportional hazards model (PL-PHM) in the same vein as that of Huang (1999). More specifically, we

assume that the conditional hazard function is given by

$$h(t|W,X) = h_0(t)\exp\left\{\phi(W) + \beta^\top X\right\}, \tag{4.2}$$

where $\phi(W) = \sum_{q=1}^{Q}\phi_q(W_q)$, $W = (W_1,\ldots,W_Q)^\top$ is a Q dimensional covariate vector, $\phi_q(\cdot)$ $(q = 1,\ldots,Q)$ are known or unknown nonlinear smooth functions. This model contains both a nonparametric component $\phi(W)$ and the parametric component $\beta^\top X$.

In reality, rarely all covariates are important in predicting the response and some components of β are in fact, zero. Efficient variable selection in such cases leads to parsimonious models with better prediction accuracy and easier interpretation.

In this chapter, we investigate the variable selection problem in the linear part of the PL-PHM given in (4.2) when covariates in X can be naturally grouped and the dimension Q of W is fixed and low. The data and model settings are partly motivated by cancer prognosis studies reported in Ma and Huang (2007) and the variable selection method introduced by Ma and Du (2012) in the partly linear accelerated failure time (AFT) model with diverging dimensions in X for right censored data. In their studies, two distinct sets of covariates are measured. The first set X represents high-dimensional genomic measurements such as microarray gene expression or SNPs. The second set W represents low-dimensional clinical and environmental risk factors. For better interpretability and easier computation, the effect of X is usually modeled in a parametric way and the effect of W is modeled with more flexible additive nonparametric functions, since many biological processes are nonlinear. However, variable selection based on such model settings mainly focuses on individual variables such as that in Ma and Huang (2007). In some applications, groups of measurements may be taken in the hopes of capturing unobservable latent variables or of measuring different aspects of complex entities (Breheny and Huang, 2009). Examples include measurements of gene expression, which can be grouped by gene pathways, and genetic markers, which can be grouped by the gene or haplotype (a set of genetic determinants located on a single chromosome) that they belong to. For example, as Wang et al. (2009) explained, when

analyzing microarray gene expression data, one can group genes into functionally similar sets as in The Gene Ontology Consortium (2000), or into known biological pathways such as the Kyoto encyclopedia of genes and genomes pathways (Kanehisa and Goto, 2000). In these settings, methods for individual variable selection may perform inefficiently by ignoring the information present in the grouping structure, while making use of the group information, as shown in Wang et al. (2009) and Huang et al. (2014), can help to identify both pathways and genes within the pathways related to the phenotypes, and hence improves understanding of biological processes.

Many variable selection methods originally proposed for uncensored data, later have been extended to the PHM. Examples include LASSO (Tibshirani, 1997), SCAD (Fan and Li, 2002), adaptive LASSO (Zhang and Lu, 2007), Elastic Net penalty (Simon et al., 2011), among others where the focus is on individual variable selection. Since grouping structures are natural in many important practical problems, several authors recently tackled the problem of variable selection with grouped covariates in the PHM. Ma et al. (2007) proposed supervised group LASSO in an attempt to select important genes and building predictive model in microarray gene expression data. Kim et al. (2012) used group LASSO in gene data to combine clinical and genomic covariates effectively. In these group selection methods, covariates belonging to the same group are either selected or deleted from the model together. However, in gene expression data, a biological pathway can be related to a certain biological outcome although some genes in that pathway may not be related to the same biological outcome. A variable selection method that can identify important pathways, and important genes within important pathways, simultaneously, is much more attractive in this case than selecting the entire group. Such a method is popularly known as a bi-level selection method and well studied in uncensored data (Huang et al., 2009; Breheny and Huang, 2009; Simon et al., 2013; Fang et al., 2015; Breheny, 2015). Zhou and Zhu (2010) proposed a hierarchically penalized method, which is a special case of the group bridge method (Huang et al., 2009) in

the linear regression model. Later, Wang et al. (2009) extended the hierarchical penalty in the PHM and established the oracle property of the estimators.

When linear models are extended to partially linear models (PLMs), variable selection in the linear part of a partially linear model has been extensively studied for uncensored data. Example includes Xie and Huang (2009), Ni et al. (2009), Liang and Li (2009), Zhao and Xue (2010), Kai et al. (2011), Xia and Yang (2016), Lv et al. (2016) and Yang et al. (2017), among others. Relatively fewer works are seen on variable selection in the PL-PHM. Du et al. (2010) performed variable selection in the linear part of a PL-PHM using SCAD and adaptive LASSO penalty where they approximated the nonparametric function using smoothing spline ANOVA. Hu and Lian (2013) and Lian et al. (2014) also performed variable selection applying SCAD penalty in diverging and ultra-high dimensional linear covariates in the PL-PHM, respectively. The latter two papers approximated the nonparametric functions using B-splines. However, all of these above researchers only considered individual variable selection in the linear part.

To the best of our knowledge, in the literature, group selection has not been considered for the partially linear survival models. To bridge this gap, in this chapter, we propose a bi-level variable selection method in the PL-PHM with a diverging number of covariates X, assuming a group structure in the linear part and a fixed and low dimensional W for clinical and/or environmental covariates in the nonparametric part. Similar approach could be applied to other types of partially linear survival models, such as the partially linear additive hazards model (PL-AHM) in the form of $h(t|W, X) = h_0(t) + \phi(W) + \beta^\top X$, in contrast to the PL-PHM given by (4.2), which will be addressed elsewhere in a different chapter. In this work, we consider the number of zero coefficients is diverging with the sample size. Typically, although the number of covariates collected is large, only a subset of covariates are important in predicting the event times. Therefore, we assume the numbers of non-zero coefficients and non-zero groups are fixed. Such an assumption is often reasonable with high dimensional

data.

The remainder of the chapter is organized as follows. In Section 4.2, we describe the group variable selection procedure for the PL-PHM. Asymptotic theories and further improvements are discussed in Section 4.3. Section 4.4 presents the numerical results. Concluding remarks are made in Section 4.5. All the technical proofs are contained in Appendix.

4.2 Grouped Variable Selection in the PL-PHM

Suppose a random sample of n subjects is observed. For the i-th subject, let T_i^e and T_i^c be the event time and the censoring time respectively, where the hazard function of T_i^e is given by (4.2). Assume that T_i^e and T_i^c are independent given the covariates, and the censoring mechanism is noninformative. The true nonparametric functions and parameters will be denoted using a superscript 0. The i.i.d observable random variables are $(T_i, \Delta_i, W_i, X_i)$ where $T_i = \min(T_i^e, T_i^c)$ and $\Delta_i = I[T_i^e \leq T_i^c]$, ($I[A]$ is the indicator function of a set A), $W_i = (W_{i1}, \ldots, W_{iQ})^\top \in \mathbb{R}^Q$, and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ are the covariates in the nonparametric and the parametric part, respectively. Define the at-risk processes $Y_i(t) = I[T_i > t]$ and the counting processes $N_i(t) = \Delta_i I[T_i \leq t]$. Note that, ϕ_q is identifiable only up to a constant and thus we assume $E\{\phi_q(W_q)\} = 0$.

Following similar strategy of Wang et al. (2009), we assume that the p variables in the linear part X can be divided into G groups. Let the g-th group have p_g variables. We use $X_{i,(g)} = \left(X_{i,g1}, \ldots, X_{i,gp_g}\right)^\top$ to denote the p_g variables in the g-th group for the i-th observation, $X_i = \left(X_{i,(1)}^\top, \ldots, X_{i,(G)}^\top\right)^\top$ to denote the total p variables, and $\beta_{(g)} = \left(\beta_{g1}, \ldots, \beta_{gp_g}\right)^\top$ to represent the regression coefficients for the g-th group. We assume that the G groups do not overlap, i.e., each variable belongs to only one group.

Thus, the partially linear proportional hazards model (4.2) can be written as

$$h(t|W,X) = h_0(t)\exp\left\{\phi(W) + \sum_{g=1}^{G}\sum_{j=1}^{p_g}\beta_{gj}X_{gj}\right\}$$
$$= h_0(t)\exp\left\{\phi(W) + \beta_{(1)}^{\top}X_{(1)} + \cdots + \beta_{(G)}^{\top}X_{(G)}\right\}. \tag{4.3}$$

Consequently, the partial likelihood is written as

$$L_n(\phi,\beta) = \prod_{i\in D}\frac{\exp\left(\phi(W_i) + \sum_{g=1}^{G}\beta_{(g)}^{\top}X_{i,(g)}\right)}{\sum_{l\in R_i}\exp\left(\phi(W_l) + \sum_{g=1}^{G}\beta_{(g)}^{\top}X_{l,(g)}\right)}, \tag{4.4}$$

where D is the set of indices of observed failures, R_i is the set of indices of the subjects who are at risk at time T_i, and $\phi(W_i) = \phi_1(W_{i1}) + \cdots + \phi_Q(W_{iQ})$. Let $Y_i(t) = I(T_i \geq t)$. The logarithm of model (4.4) in counting process notation can be written as

$$l_n(\phi,\beta) = \sum_{i=1}^{n}\Delta_i\left\{\phi(W_i) + \sum_{g=1}^{G}\beta_{(g)}^{\top}X_{i,(g)}\right.$$
$$\left. - \log\sum_{l=1}^{n}Y_l(T_i)\exp\left(\phi(W_l) + \sum_{g=1}^{G}\beta_{(g)}^{\top}X_{l,(g)}\right)\right\}. \tag{4.5}$$

To estimate parameter (ϕ,β) in the model (4.3), since ϕ is an infinitely dimensional non-parametric function, we use the Sieve method in maximizing the log-likelihood $l_n(\phi,\beta)$ with respect to (ϕ,β), and construct Sieve space for ϕ. To do that, we use polynomial splines to approximate the nonparametric components. Without loss of generality, we assume W_q $(q = 1,\ldots,Q)$ has a support $[0,1]$. For each non-parametric component, $\phi_q(W_q)$, let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{k'} < 1 = \tau_{k'+1}$ be a partition of [0, 1] into subintervals $[\tau_k,\tau_{k+1}), k = 0,\ldots,k'$ with k' internal knots. A polynomial spline of order r is a function whose restriction to each subinterval is a polynomial of degree $r-1$ and globally $r-2$ times continuously differentiable on [0, 1]. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\left\{\tilde{B}_{q1}(x),\ldots,\tilde{B}_{q\tilde{k}}(x)\right\}$ with $\tilde{k} = k' + r$. As ϕ_q is identifiable only up to a constant, we put a centering constraint $E\{\phi_q(W_q)\} = 0$, and use the subspace of spline functions: $S_q^0 := [s : s = \sum_{k=1}^{\tilde{k}-1}\alpha_{qk}B_{qk}(x), \sum_{i=1}^{n}s(W_{iq}) = 0]$, with basis $\left\{B_{qk}(x) = \sqrt{K}(\tilde{B}_{qk}(x) - \sum_{i=1}^{n}\tilde{B}_{qk}(W_{iq})/n), k = 1,\ldots,K = \tilde{k}-1\right\}$ (the subspace has a

degree $= \tilde{k} - 1$ due to the normalization constraint $\sum_{k=1}^{\tilde{k}} \tilde{B}_{qk}(x) \equiv 1$). The multiplicative constant $\sqrt{K}$ is incorporated in the basis definition to simplify some expression later in the proofs, as done in Wang et al. (2011). Using spline expansions, we can approximate the nonparametric components by $\phi_q(x) \approx \sum_{k=1}^{K} \alpha_{qk} B_{qk}(x), 1 \leq q \leq Q$. Therefore, the problem of estimating ϕ_q is now transformed to the problem of estimating the coefficients $\alpha_q = (\alpha_{q1}, \ldots, \alpha_{qK})^\top$.

Let $Z = (B_{11}(W_1), \ldots, B_{1K}(W_1), \ldots, B_{Q1}(W_Q), \ldots, B_{QK}(W_Q))^\top$ denote the QK basis functions and $\alpha = (\alpha_{11}, \ldots, \alpha_{1K}, \ldots, \alpha_{Q1}, \ldots, \alpha_{QK})^\top$ denote the corresponding coefficients. Since the q-th nonparametric component can be approximated by $\sum_{k=1}^{K} \alpha_{qk} B_{qk}(x)$ $(q = 1, \ldots, Q)$, it is reasonable to assume that $B_{q1}(x), \ldots, B_{qK}(x)$ are K variables belonging to one group. Therefore, the QK variables in Z can be divided into Q groups, where each of the q-th group has K variables. We use $Z_{i,(q)} = (B_{i,q1}, \ldots, B_{i,qK})^\top$ $(q = 1, \ldots, Q;\ k = 1, \ldots, K)$, to denote the K basis functions in the q-th group for the i-th observation. Similarly, we use $Z_i = \left(Z_{i,(1)}^\top, \ldots, Z_{i,(Q)}^\top\right)^\top$ to denote the total QK variables for the i-th observation, and $\alpha_{(q)} = (\alpha_{q1}, \ldots, \alpha_{qK})^\top$ to represent the regression coefficients for the q-th group. We assume that the number of variables in each group is K, i.e., we consider the same number of basis functions to approximate each nonparametric function. To simplify computation, since we have assumed W_q $(q = 1, \ldots, Q)$ have the same support [0,1], we can assume $B_{qk}(x) = B_{q'k}(x)$ for $q \neq q',\ 1 \leq q, q' \leq Q,\ 1 \leq k \leq K$.

The partial likelihood in (4.5) is then equivalent to

$$l_n(\alpha, \beta) = \sum_{i=1}^{n} \Delta_i \left[\sum_{q=1}^{Q} \alpha_{(q)}^\top Z_{i,(q)} + \sum_{g=1}^{G} \beta_{(g)}^\top X_{i,(g)} - \log \sum_{l=1}^{n} \left\{ Y_l(T_i) \exp\left(\sum_{q=1}^{Q} \alpha_{(q)}^\top Z_{l,(q)} + \sum_{g=1}^{G} \beta_{(g)}^\top X_{l,(g)} \right) \right\} \right]. \tag{4.6}$$

To conduct variable selection in the PL-PHM, Hu and Lian (2013) and Lian et al. (2014) considered individual variable selection in the linear part by maximizing the penalized log

partial likelihood objective function for estimating both α and β defined in the following:

$$pl_n(\alpha, \beta) = \frac{1}{n} l_n(\alpha, \beta) - \sum_{j=1}^{p} p_{\lambda_n}(\beta_j),$$

where $p_{\lambda_n}(\beta_j)$ is a penalty function. Let $(\hat{\alpha}, \hat{\beta})$ be the maximizer of the above penalized partial likelihood, then, the penalized estimators of ϕ_q $(q = 1, \ldots, Q)$ and β are $\sum_{k=1}^{K} \hat{\alpha}_{qk} B_{qk}$ and $\hat{\beta}$, respectively. In this chapter, our focus is on group selection, and the above individual variable selection is a special case of the following group selection problem.

4.2.1 Hierarchically Penalized PL-PHM

To conduct group selection, we follow Wang et al. (2009)'s procedure. Similar to theirs, we reparameterize β_{gj} as

$$\beta_{gj} = \gamma_g \theta_{gj} \ (g = 1, \ldots, G;\ j = 1, \ldots, p_g),$$

where $\gamma_g \geq 0$ for identifiability. This decomposition indicates that all β_{gj} $(j = 1, \ldots, p_g)$ belong to the g-th group as it treats β_{gj} hierarchically. Parameter γ_g explains β_{gj} $(j = 1, \ldots, p_g)$ at the group level and θ_{gj}'s explain differences among individuals within the g-th group. Let $\theta_{(g)} = (\theta_{g1}, \ldots, \theta_{gp_g})^\top$, then $\beta_{(g)} = \gamma_g \theta_{(g)}$. The partial likelihood function thus can be written as

$$L_n(\alpha, \gamma, \theta) = \prod_{i \in D} \frac{\exp(\sum_{q=1}^{Q} \alpha_{(q)}^\top Z_{i,(q)} + \sum_{g=1}^{G} \gamma_g \theta_{(g)}^\top X_{i,(g)})}{\sum_{l \in R_i} \exp(\sum_{q=1}^{Q} \alpha_{(q)}^\top Z_{l,(q)} + \sum_{g=1}^{G} \gamma_g \theta_{(g)}^\top X_{l,(g)})}.$$

where $\gamma = (\gamma_1, \ldots, \gamma_G)^\top$ and $\theta = (\theta_{11}, \ldots, \theta_{1p_1}, \ldots, \theta_{G1}, \ldots, \theta_{Gp_G})^\top$. Let $l_n(\alpha, \gamma, \theta)$ denote $\log\{L_n(\alpha, \gamma, \theta)\}$. For group selection in the linear part, we consider the penalized log partial likelihood is given as

$$\max_{\alpha_{(q)}, \gamma_g, \theta_{gj}} \left\{ \frac{1}{n} l_n(\alpha, \gamma, \theta) - \lambda_\gamma \sum_{g=1}^{G} \gamma_g - \lambda_\theta \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\theta_{gj}| \right\}, \tag{4.7}$$

subject to $\gamma_g \geq 0$ $(g = 1, \ldots, G)$, where $\lambda_\gamma \geq 0$ and $\lambda_\theta \geq 0$ are two tuning parameters, which control the sparsity of the estimation at the group level and within group level, respectively.

As shown by Wang et al. (2009) in the linear PHM, for fixed (α, β) and given values of λ_γ and λ_θ, the maximizer of (4.7) with respect to (γ, θ), where $l_n(\alpha, \gamma, \theta)$ is constant, is unique.

Finally, in the same vein as Wang et al. (2009), we can combine λ_γ and λ_θ into one tuning parameter $\lambda = \lambda_\gamma \lambda_\theta$ such that (4.7) is equivalent to

$$\max_{\alpha_{(q)}, \gamma_g, \theta_{gj}} \left\{ \frac{1}{n} l_n(\alpha, \gamma, \theta) - \sum_{g=1}^{G} \gamma_g - \lambda \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\theta_{gj}| \right\}, \tag{4.8}$$

subject to $\gamma_g \geq 0$ $(g = 1, ..., G)$. Lemma 1 illustrates the meaning of equivalence.

Lemma 1. *Let $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ be a local maximizer of (4.7). Then there exists a local maximizer $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ of (4.8) such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}_g^* \hat{\theta}_{gj}^* = \hat{\gamma}_g^\dagger \hat{\theta}_{gj}^\dagger$. Similarly, if $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ is a local maximizer of (4.8), then there exists a local maximizer $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ of (4.7) such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}_g^* \hat{\theta}_{gj}^* = \hat{\gamma}_g^\dagger \hat{\theta}_{gj}^\dagger$.*

Furthermore, criterion (4.8) can be written into an equivalent form using the regression coefficients α and β.

Lemma 2. *If $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local maximizer of (4.8), then $(\hat{\alpha}, \hat{\beta})$, where $\hat{\beta}_{gj} = \hat{\gamma}_g \hat{\theta}_{gj}$, is a local maximizer of*

$$\max_{\alpha_q, \beta_{gj}} \left\{ \frac{1}{n} l_n(\alpha, \beta) - 2\lambda^{1/2} \sum_{g=1}^{G} \left(\sum_{j=1}^{p_g} |\beta_{gj}| \right)^{1/2} \right\}. \tag{4.9}$$

On the other hand, if $(\hat{\alpha}, \hat{\beta})$ is a local maximizer of (4.9), then $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local maximizer of (4.8), where $\hat{\gamma}_g = (\lambda \sum_{j=1}^{p_g} |\hat{\beta}_{gj}|)^{1/2}$ and $\hat{\theta}_{gj} = \hat{\beta}_{gj} / \hat{\gamma}_g$ if $\hat{\gamma}_g \neq 0$ and zero otherwise.

The numerical computation is based on (4.8) while the proof of asymptotic properties is based on (4.9). Instead of using L_2-norm which performs group LASSO (Yuan and Lin, 2006), we used L_1-norm to the within group coefficients in (4.9). In addition, the group coefficients are penalized by a bridge-type penalty (Frank and Friedman, 1993), i.e., $L_{1/2}$-norm. So, the hierarchical penalty can remove unimportant groups and some unimportant variables in the important groups.

4.2.2 Computational Algorithm

To estimate α, γ and θ in (4.8), we will use an iterative algorithm. First, we fix γ and estimate (α, θ); then fixing θ, we estimate (α, γ). We iterate between these steps until convergence is achieved. Precisely, the algorithm is written as

Step 0. Center and normalize X_{gj}, and obtain an initial value $\gamma_g^{(0)}$ for each γ_g; i.e., $\gamma_g^{(0)} = 1$. Let $s = 1$.

Step 1. At the s-th iteration, let $\tilde{X}_{i,gj} = \gamma_g^{(s-1)} X_{i,gj}$ $(g = 1, \ldots, G;\ j = 1, \ldots, p_g)$ and obtain estimate $(\alpha^{(s)}, \theta^{(s)})$ by

$$(\alpha^{(s)}, \theta^{(s)}) = \underset{\alpha_{qk}, \theta_{gj}}{\arg\max} \frac{1}{n} \log \left\{ \prod_{i \in D} \frac{\exp(\sum_{q=1}^{Q} \sum_{k=1}^{K} \alpha_{qk} Z_{i,qk} + \sum_{g=1}^{G} \sum_{j=1}^{p_g} \theta_{gj} \tilde{X}_{i,gj})}{\sum_{l \in R_i} \exp(\sum_{q=1}^{Q} \sum_{k=1}^{K} \alpha_{qk} Z_{l,qk} + \sum_{g=1}^{G} \sum_{j=1}^{p_g} \theta_{gj} \tilde{X}_{l,gj})} \right\} - \lambda \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\theta_{gj}|.$$

Step 2. Let $\tilde{X}_{i,g} = \sum_{j=1}^{p_g} \theta_{gj}^{(s)} X_{i,gj}$ $(g = 1, \ldots, G)$ and obtain estimate $(\alpha^{(s)}, \gamma^{(s)})$ by

$$(\alpha^{(s)}, \gamma^{(s)}) = \underset{\alpha_{qk}, \gamma_g \geq 0}{\arg\max} \frac{1}{n} \log \left\{ \prod_{i \in D} \frac{\exp(\sum_{q=1}^{Q} \sum_{k=1}^{K} \alpha_{qk} Z_{i,qk} + \sum_{g=1}^{G} \gamma_g \tilde{X}_{i,gj})}{\sum_{l \in R_i} \exp(\sum_{q=1}^{Q} \sum_{k=1}^{K} \alpha_{qk} Z_{l,qk} + \sum_{g=1}^{G} \gamma_g \tilde{X}_{l,gj})} \right\} - \sum_{g=1}^{G} \gamma_g.$$

In this step, $\alpha^{(s)}$ is updated from $\alpha^{(s)}$ in step 1.

Step 3. Repeat Steps 1 and 2 until $\alpha^{(s)}, \gamma^{(s)}$, and $\theta^{(s)}$ converge at the m-th iteration. Let $\hat{\alpha} = \alpha^{(m)}$ and $\hat{\beta}_{(g)} = \gamma_g^{(m)} \theta_{(g)}^{(m)}$ be the final solutions.

Since at each step, the value of objective function (4.8) is non-decreasing, this algorithm always converges. Step 1 is a LASSO-type problem without penalizing α, and the algorithms proposed in Fan and Li (2002), Gui and Li (2005), Zhang and Lu (2007) or Park and Hastie (2007) can be used to solve for θ. Step 2 is a nonnegative garrote algorithm without penalizing α, and we can use Fan and Li (2002) or Yuan and Lin (2007)'s algorithm to solve for γ.

4.3 Asymptotic Theory

For theoretical analysis, we will consider the counting process representation of the partial likelihood. We denote the true risk score by $m^0(W, X) = \phi^0(W) + \beta^{0\top} X$ where $\phi^0(W) = \phi_1^0(W_1) + \cdots + \phi_Q^0(W_Q)$. Let $R^\top = (W^\top, X^\top)$ be all the covariates and g, h be any functions

of R (h can be vector valued). We define

$$\begin{aligned}
S_n^{(0)}(g,t) &= n^{-1}\sum_{i=1}^{n} Y_i(t)\exp[g(R_i)],\\
S_n^{(1)}(g,t)[h] &= n^{-1}\sum_{i=1}^{n} Y_i(t)h(R_i)\exp[g(R_i)],\\
S_n^{(2)}(g,t)[h] &= n^{-1}\sum_{i=1}^{n} Y_i(t)h(R_i)^{\otimes 2}\exp[g(R_i)],\\
G_n(g,t)[h] &= S_n^{(1)}(g,t)[h]/S_n^{(0)}(g,t),\\
V_n(g,t)[h] &= S_n^{(2)}(g,t)[h]/S_n^{(0)}(g,t) - G_n(g,t)[h]G_n^{\top}(g,t)[h],
\end{aligned}$$

where for any vector ξ, $\xi^{\otimes 2}$ simply means $\xi\xi^{\top}$. Let $s^{(0)}(g,t) = E(S_n^{(0)}(g,t))$, $s^{(j)}(g,t)[h] = E(S_n^{(j)}(g,t)[h]), j = 1,2$, $G(g,t)[h] = s^{(1)}(g,t)[h]/s^{(0)}(g,t)$, $V(g,t)[h] = s^{(2)}(g,t)[h]/s^{(0)}(g,t) - G(g,t)[h]G^{\top}(g,t)[h]$. Also, let P_n be the empirical measure of $(T_i,\Delta_i,R_i), 1 \le i \le n$ and let P be the probability measure of (T,Δ,R). Let $P_{\Delta n}$ be the (subprobability) empirical measure of $(T_i,\Delta_i = 1,R_i), 1 \le i \le n$ and let P_{Δ} be one subprobability measure of $(T,\Delta = 1,R)$. It is convenient to use linear functional notation, for example, $P_{\Delta n}f = \int f dP_{\Delta n} = \int \Delta f dP_n = n^{-1}\sum_i \Delta_i f(T_i,\Delta_i,R_i)$ for any f such that this integral is well defined. Let $\|a\|$ denotes the L_2 norm of a column vector a.

The log partial likelihood can be rewritten as

$$l_n(\alpha,\beta) = \sum_{i=1}^{n}\int_0^{\tau}\left\{\omega^{\top}L_i - log(S_n^{(0)}(g,t))\right\}dN_i(t),$$

where $\omega = (\alpha^{\top},\beta^{\top})^{\top}$, $L_i = (Z_i^{\top},X_i^{\top})^{\top}$ and g is the function of R defined by $(\alpha,\beta) : g(R) = \sum_{q=1}^{Q}\alpha_q^{\top}B_q(W_q) + \beta^{\top}X$, $Z_q(W_q) = B_q(W_q) = (B_{q1}(W_q),\ldots,B_{qK}(W_q))^{\top}$. We only consider events over a finite interval $[0,\tau]$. The score function and the observed information are given by

$$U_n(\alpha,\beta) = \sum_{i=1}^{n}\int_0^{\tau}\left\{L_i - G_n(g,t)[L]\right\}dN_i(t),$$

and

$$\Sigma_n(\alpha,\beta) = \sum_{i=1}^{n}\int_0^{\tau} V_n(g,t)[L]dN_i(t),$$

respectively. We first consider the penalized log partial likelihood function with a general penalty function. Let the objective function be

$$Q_n(\alpha, \beta) = \frac{1}{n} l_n(\alpha, \beta) - \sum_{g=1}^{G} p_{\lambda_n}^{(g)}(|\beta_{(g)}|), \tag{4.10}$$

where $p_{\lambda_n}^{(g)}(|\beta_{(g)}|) = p_{\lambda_n}^{(g)}(|\beta_{g1}|, \ldots, |\beta_{gp_g}|)$ is a general p_g-variate penalty function for the linear parameters in the g-th group. We let the penalty functions $p_{\lambda_n}^{(g)}(\cdot)$ $(g = 1, \ldots, G)$ in (4.10) to vary between groups as well as $p_{\lambda_n}^{(g)}(\cdot)$ to depend on the tuning parameter λ_n that varies with n.

Adopting notations of Wang et al. (2009), we write the true parameter vector in the sparse linear part as $\beta^0 = (\beta_{\mathcal{A}}^{0\top}, \beta_{\mathcal{B}}^{0\top}, \beta_{\mathcal{C}}^{0\top})^\top$, where $\mathcal{A} = \{(g, j) : \beta_{g_j}^0 \neq 0\}$, $\mathcal{B} = \{(g, j) : \beta_{g_j}^0 = 0, \beta_{(g)}^0 \neq 0\}$, and $\mathcal{C} = \{(g, j) : \beta_{(g)}^0 = 0\}$. Here $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ contain the indices of nonzero coefficients, indices of zero coefficients that belong to nonzero groups, and indices of zero coefficients that belong to zero groups. Thus, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are disjoint and partition the set of all indices of coefficients. We write $\mathcal{D} = \mathcal{B} \cup \mathcal{C}$, which contains the indices of all zero coefficients. We also define

$$a_n = \max_{(g,j)} \left\{ \frac{\partial p_{\lambda_n}^{(g)}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\partial |\beta_{gj}|} : \beta_{gj}^0 \neq 0 \right\},$$

$$b_n = \max_{(g,j)} \left\{ \frac{\partial^2 p_{\lambda_n}^{(g)}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\partial |\beta_{gj}|^2} : \beta_{gj}^0 \neq 0 \right\}.$$

Lastly, let s be the number of nonzero groups. Without loss of generality, we assume that $\beta_{(g)}^0 \neq 0$ $(g = 1, \ldots, s)$ and $\beta_{(g)}^0 = 0$ $(g = s + 1, \ldots, G)$. Let s_g be the number of nonzero coefficients in group g $(g = 1, \ldots, s)$. Again, without loss of generality, we assume that $\beta_{gj}^0 \neq 0$ $(g = 1, \ldots, s;\ j = 1, \ldots, s_g)$ and $\beta_{gj}^0 = 0$ $(g = 1, \ldots, s;\ j = s_g + 1, \ldots, p_g)$.

The following technical conditions are used in the study of asymptotic.

(B1) The covariate vector $(R^\top = (W^\top, X^\top))$ has a bounded support: without loss of generality the support of W is assumed to be $[0, 1]^Q$, with the marginal density of each covariate in W being continuous and bounded away from zero and infinity, and the covariate vector X is bounded.

(B2) (i) Only observations with censored event times in a finite interval $[0, \tau]$ are used in the partial likelihood. At this point τ, the baseline cumulative hazard function $\Lambda_0(\tau) \equiv \int_0^\tau \lambda_0(s)ds < \infty$. (ii) $P(\Delta = 1|R)$ and $P(T^c > \tau|R)$ are both bounded away from zero with probability one.

(B3) Let $\mathscr{H}_d$ be the collection of all functions on support $[0,\ 1]$ whose m-th order derivative satisfied the Hölder condition of order r with $d \equiv m + r$. That is, for each $h \in \mathscr{H}_d$, there exists a constant $M_0 \in (0, \infty)$ such that $\left|h^{(m)}(s) - h^{(m)}(t)\right| \leq M_0|s - t|^r$, for any $s, t \in [0, 1]$. Then, $\varphi_q^0 \in \mathscr{H}_d$ $(q = 1, \ldots, Q)$, for some $d > 1/2$. The order of the spline satisfies $r > d + 1/2$.

(B4) $E\left\{\sup_{t\in[0,\tau]} Y(t) \left\|L\right\|^2 \exp\left(\omega^\top L\right)\right\} = O(K + p)$.

(B5) Let $\Sigma = \int_0^\tau V(m^0, t)[L]s^{(0)}(m^0, t)\lambda_0(t)dt$, where $m^0 = m^0(W, X)$. The eigenvalues of Σ are bounded away from zero and infinity.

(B6) The p_g-variate penalty function for parameters in the g-th group satisfies the following two conditions:

$$p_{\lambda_n}^{(g)}(|\beta_{(g)}|) \geq 0 \quad (\beta_{(g)} \in \mathbb{R}^{p_g}), \quad p_{\lambda_n}^{(g)}(0) = 0; \tag{4.11}$$

$$p_{\lambda_n}^{(g)}(|\beta_{(g)}|) \geq p_{\lambda_n}^{(g)}(|\beta_{(g)}^*|) \quad (|\beta_{gj}| \geq |\beta_{gj}^*|;\ j = 1, \ldots, p_g). \tag{4.12}$$

Similar conditions to those listed above have been considered in the literature (Hu and Lian, 2013; Wang et al., 2009) and are quite reasonable. Condition (B1) places the boundedness condition on the covariates. It is unpleasant, but not too restrictive because in many practical situations continuous covariates may be typically rescaled to fall between 0 and 1. (B2)(i) avoids the unboundedness of the loss function and pseudo-score functions at the end point of the support of the observed event time. (B2)(ii) ensures that the probability of being right censored at τ and the probability of being observed events are positive and bounded

away from zero regardless of the covariate values. (B3) ensures the uniform continuity of the functions. A condition similar to (B4) was considered by Bradic et al. (2011) for diverging number of parameters following Andersen and Gill (1982). The positive-definiteness of Σ in (B5) is a reasonable assumption by the following discussion. The term $LL^\top$ appears in the definition of Σ. Under mild assumptions, Huang et al. (2010) showed that eigenvalues of $E(ZZ^\top)$ are bounded and bounded away from zero and hence, we can expect that eigenvalues of $E(LL^\top)$ are bounded and bounded away from zero if eigenvalues of $E(XX^\top)$ are, and Z and X are linearly independent. Wang et al. (2009) considered the condition (B6) for hierarchical group variable selection in the PHM.

Theorem 1. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$. *Under the regularity conditions (B1) - (B6), assume that* Q, s *and* s_g *is fixed,* $K \to \infty$, $p \to \infty$, $(K+p)/n \to 0$, $\gamma_n (K+p)^{3/2} = O(1)$, $a_n = O_p(\gamma_n)$ *and* $b_n \to 0$, *then there exists a local maximizer* $(\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ *of* $(\alpha^\top, \beta^\top)^\top$ *in (4.10) and* $\hat{\phi}_q = \sum_{k=1}^K \hat{\alpha}_{qk} B_{qk}$, $\hat{\phi}(w) = \sum_{q=1}^Q \hat{\phi}_q(w_q)$ *such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p\left(\sqrt{(K+p)/n} + K^{-d}\right)$.

Theorem 2. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$ *and* $(\hat{\alpha}^\top, \hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ *be the local maximizer of* $Q_n(\alpha, \beta)$ *in (4.10). For* $(g, j) \in \mathcal{D}$, *i.e.,* $\beta_{gj}^0 = 0$, *under the same conditions as in Theorem 1, if* $\gamma_n^{-1} \partial p_{\lambda_n}^{(g)}(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}|)/\partial |\beta_{gj}| \to \infty$ *as* $n \to \infty$, *then we have* $\hat{\beta}_{gj} = 0$ *with probability approaching 1.*

In the following section, we show how to construct penalty function $p_{\lambda_n}^{(g)}$ such that the conditions in Theorem 2 can be satisfied.

4.3.1 Adaptive Hierarchical Penalty and Further Improvement

The above results are obtained for any general penalty. Following Wang et al. (2009), here we will show the asymptotic results for hierarchically penalized PL-PHM based on criterion (4.9). If we write $\lambda_n = 2\lambda^{1/2}$ in (4.9), then based on Theorems 1 and 2 we have

Corollary 1. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$. *If* $\lambda_n = O_p(\gamma_n)$, *then there exists a local maximizer* $(\hat{\alpha}^\top, \hat{\beta}^\top)^\top = (\hat{\alpha}^\top, \hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ *for the hierarchically penalized PL-PHM in (4.9) such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$; *if further* $p^{-1/2}\gamma_n^{-3/2}\lambda_n \to \infty$ *as* $n \to \infty$, *then* $\hat{\beta}_{\mathcal{C}} = 0$ *with probability tending to 1.*

Comparing Corollary 1 with Theorem 2, we see that although the hierarchical penalty can effectively remove unimportant groups because $\hat{\beta}_{\mathcal{C}} = 0$ with probability approaching to 1, it cannot effectively remove unimportant variables within the important groups as $\hat{\beta}_{\mathcal{D}} = 0$ with probability tending to 1 may not hold. To tackle this limitation, we apply the adaptive idea used in Breiman (1995), Shen and Ye (2002), Zhang and Lu (2007), Zhao and Yu (2006), Zou (2006), Zou (2008), Wang et al. (2009), Liu et al. (2014), and others, which is to penalize different coefficients differently. To do so, we maximize the following objective function

$$Q_n^w(\alpha, \beta) = \frac{1}{n} l_n(\alpha, \beta) - \lambda_n \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_g} w_{n,gj} |\beta_{gj}| \right\}^{1/2}, \tag{4.13}$$

where $w_{n,gj}$'s are pre-specified non-negative weights. The next theorem shows that, by controlling weights properly, the adaptive hierarchically penalized PL-PHM has the selection consistency as stated in Theorem 2.

Theorem 3. *Let us define*

$$w_{n,\max}^{\mathcal{A}} = \max\left\{w_{n,gj} : (g,j) \in \mathcal{A}\right\}, \quad w_{n,\min}^{\mathcal{A}} = \min\left\{w_{n,gj} : (g,j) \in \mathcal{A}\right\};$$

$$w_{n,\max}^{\mathcal{D}} = \max\left\{w_{n,gj} : (g,j) \in \mathcal{D}\right\}, \quad w_{n,\min}^{\mathcal{D}} = \min\left\{w_{n,gj} : (g,j) \in \mathcal{D}\right\}.$$

Let $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$. *Under the same conditions as assumed in Theorem 1, if* $\gamma_n^{-1}\lambda_n w_{n,\max}^{\mathcal{A}} \left(w_{n,\min}^{\mathcal{A}}\right)^{-1/2} \to 0$, $\lambda_n \left(w_{n,\max}^{\mathcal{A}}\right)^2 \left(w_{n,\min}^{\mathcal{A}}\right)^{-3/2} \to 0$, *and* $\gamma_n^{-1}\lambda_n w_{n,\min}^{\mathcal{D}} / (w_{n,\max}^{\mathcal{A}} + w_{n,\max}^{\mathcal{D}})^{1/2} \to \infty$ *as* $n \to \infty$, *there exists a local maximizer* $(\hat{\alpha}^\top, (\hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{D}}^\top))^\top$ *of* $(\alpha^\top, (\beta_{\mathcal{A}}^\top, \beta_{\mathcal{D}}^\top))^\top$ *in (4.13) such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$ *and* $\hat{\beta}_{\mathcal{D}} = 0$ *with probability tending to 1.*

Finally, we specify our λ_n and the weights $w_{n,gj}$ that satisfy conditions in Theorem 3, which are given by the following corollary.

Corollary 2. *Let* $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$ *and* $\tilde{\beta}_n$ *be an estimator such that,* $\left\|\tilde{\beta}_n - \beta^0\right\| = O_p(\gamma_n)$. *If* $\lambda_n = \gamma_n/\log(n)$ *and* $w_{n,gj} = 1/|\tilde{\beta}_{n,gj}|^r$, *where* $r > 0$, *then there exists a local maximizer* $(\hat{\alpha}^\top, (\hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{D}}^\top))^\top$ *of* $(\alpha^\top, (\beta_{\mathcal{A}}^\top, \beta_{\mathcal{D}}^\top))^\top$ *in (4.13) such that* $\left\|\hat{\phi} - \phi^0\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$ *and* $\hat{\beta}_{\mathcal{D}} = 0$ *with probability tending to 1.*

In practice, we choose $(\tilde{\alpha}_n, \tilde{\beta}_n) = \arg\max_{\alpha,\beta} l_n(\alpha, \beta)$, the estimator from the unpenalized score function when p is diverging with n and $p < n$. From Corollary 1 and Corollary 2, we notice that the rates of convergence of the estimators are the same but the selection performance of the adaptive hierarchically penalized method is superior to that of the hierarchically penalized method, because the adaptive method possesses the individual variable selection consistency, while the non-adaptive method holds only group selection consistency.

We have obtained the rate of convergence and selection consistency of the estimators of ϕ and β with a diverging β. For the sparse finite dimensional parameter vector β, a root-n rate may be obtained using the semiparametric theory such as in Huang (1999). Further, oracle properties of $\hat{\beta}$ are expected to hold. Given the estimator $\hat{\beta}_{\mathcal{A}}$ is obtained by penalizing the log semiparametric partial likelihood, following the lines in Hu and Lian (2013) and Wang et al. (2009), we conjecture that, under certain conditions, $\hat{\beta}_{\mathcal{A}} = \beta_{\mathcal{A}}^0$ with probability 1 and has an asymptotic normal distribution.

4.4 Numerical Results

4.4.1 Simulation Studies

To evaluate the finite-sample performance of the hierarchically penalized (HP) method and its adaptive version (AHP) in the PL-PHM, we conducted two simulation studies. We compared the results with those based on some existing individual (LASSO, Adaptive LASSO SCAD, MCP) and group (Group SCAD or G-SCAD, Group MCP or G-MCP) variable selection methods developed for linear models. LASSO, Adaptive LASSO (A-LASSO) and

SCAD have been used for variable selection in the PLMs (Ma and Du, 2012; Hu and Lian, 2013). The asymptotic theory for G-SCAD and G-MCP under PLMs has not studied in the literature, we used them only for comparison purpose and leave the asymptotic theory for future research. We expect that similar results to those for AHP method will also hold for G-SCAD and G-MCP penalties. In our simulation studies we used `R` packages `ncvreg` and `Coxnet` for individual variable selection penalties and `grpreg` for computing G-SCAD and G-MCP estimates. We used these penalties for variable selection in the linear part after linearizing the nonparametric functions $\phi(\cdot)$ using B-splines where the tuning parameter λ_n is chosen by the built-in five-fold cross validation method. For computation of our AHP group selection method in the PL-PHM, we used the `R` package `penalized` and `R` program written by Wang et al. (2009) for the linear PHM, where the tuning parameter λ_n is chosen to be 10 for HP and 20 for AHP methods based on trial and error method. We have not developed a data-driven tuning parameter selection method for HP and AHP group selection methods. We will investigate this issue in our future work.

Five performance measures are used to compare these methods: number of true groups selected (TG), number of zero group selected (FG), number of true nonzero variables selected as nonzero (TP), number of true zero variables selected as nonzero (FP), and L_2- prediction error (PE) in the excess risk defined as $\left\|\left\{\hat{\beta}^\top Z + \hat{\phi}_1(W_1) + \hat{\phi}_2(W_2)\right\} - \left\{\beta^\top Z + \phi_1(W_1) + \phi_2(W_2)\right\}\right\|$. As a benchmark, we compute the oracle estimates, which are obtained by maximizing (4.6) for model (4.3) which includes only important variables and groups.

Variable selection is a computationally extensive procedure and can take a lot of time if convergence is slow. We used 'WestGrid' (https://www.westgrid.ca) to conduct our simulation studies which benefited us in terms of computational time. WestGrid is helping Compute Canada (https://www.computecanada.ca) to lead the acceleration of research and innovation by bringing together computing facilities, research data management services, and a network of technical experts to meet researchers need. It has multiple computing facilities where

the researchers can send their computing codes and define parameters like computing time, memory, cores to be used based on the computational burden of their jobs. To conduct our simulation studies, we submitted all of our simulations parallelly to the computing facilities at the same time. On average, it took only an hour to conduct 500 simulations in WestGrid.

In Example 1, the number of groups is moderately large, the group sizes are equal and relatively large, and within each group the coefficients are either all nonzero or all zero. In Example 2, the group sizes vary and there are zero coefficients in a nonzero group. In each example, we set sample size n =400 and baseline hazard functions $h_0(t) = 1.0$. The censoring variable is generated from a uniform distribution over $[0, C_o]$, where C_0 is chosen to yield censoring rate = 30%. For each of these settings, we replicate 500 simulations.

Example 1. In this example, there are 7 groups in the linear part, each with 5 covariates, and two nonparametric functions. For the linear covariates, the covariate vector is $X^\top = (X_1^\top, \ldots, X_7^\top)$. The subvector of covariates that belong to the same group is $X_j^\top = (X_{5(j-1)+1}, \ldots, X_{5(j-1)+5});\ j = 1, \ldots, 7$. To generate the covariates $X_1, \ldots, X_{35}$, we first simulate 35 random variables $R_1, \ldots, R_{35}$ independently from the standard normal distribution. Then Z_j $(j = 1, \ldots, 7)$ are simulated from a multivariate normal distribution with mean zero and an AR(1) covariance structure such that $\text{cov}(Z_{j1}, Z_{j2}) = 0.4^{|j_1 - j_2|}$ for $j_1,\ j_2 = 1, \ldots, 7$. The covariates $X_1, \ldots, X_{35}$ are generated as $X_j = (Z_{gj} + R_j)/6$ $(j = 1, \ldots, 35)$, where g_j is the smallest integer greater than $(j-1)/5$ and the X_j's with the same value of g_j belong to the same group. Similar correlation structure was considered in Huang et al. (2009). The nonparametric functions are $\phi_1(W_1) = W_1^2 - (25/12)$ and $\phi_2(W_2) = \exp(-W_2) - 2\sinh(5/2)/5$, where the covariates W's are sampled from $U(-2.5, 2.5)$. Such nonparametric functions were considered in Cui et al. (2013) in a nonparametric additive regression model. The event times in Example 1 are generated from an exponential distribution with a hazard rate given as follows:

$$h(t|X, W) = h_0(t) \exp\left\{\beta^\top X + \phi_1(W_1) + \phi_2(W_2)\right\},$$

where $\beta = (\underbrace{1.2, \ldots, 1.2}_{5}, \underbrace{3.6, \ldots, 3.6}_{5}, \underbrace{2.4, \ldots, 2.4}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0, \ldots, 0}_{5})^\top$.

To estimate nonparametric functions, we use B-splines, see details in Section 4.2 for centering the B-splines in general. Specifically, center $\phi_1(W_1)$ and $\phi_2(W_2)$ such that $E\{\phi_1(W_1)\} = E\{\phi_2(W_2)\} = 0$. We approximated the nonlinear functions using cubic B-spline functions. Lian et al. (2014) used 5 to 8 basis functions in their simulations and found similar results. They reported the results only for 6 basis functions. To ease the computational burden, we also choose $K = 6$ as the number of basis functions in B-splines. This choice of K is small enough to avoid overfitting and big enough to flexibly approximate the smooth functions (Gray, 1992; Cheng and Wang, 2011). In this example, there exists three important groups and all variables within each group are important. This example illustrates that the proposed group selection methods have the ability to identify important groups.

Example 2. In this experiment, the group size differs across groups and some groups have a mixture of important and unimportant variables. There are seven groups: three groups each of size 8 and four groups each of size 4. The covariate vector is $X^\top = (X_1^\top, \ldots, X_7^\top)$, where the seven subvectors of covariates are $X_j^\top = (X_{8(j-1)+1}, \ldots, X_{8(j-1)+8})$, for $j = 1, 2, 3$, and $X_j^\top = (X_{4(j-1)+13}, \ldots, X_{4(j-1)+16})$, for $j = 4, 5, 6, 7$. To generate the covariates $X_1, \ldots, X_{40}$, we first simulate Z_i $(i = 1, \ldots, 7)$ and $R_1, \ldots, R_{40}$ independently from the standard normal distribution. For $j = 1, \ldots, 24$, let g_j be the largest integer less than $j/8 + 1$ and, for $j = 25, \ldots, 40$, let g_j be the largest integer less than $(j - 24)/4 + 1$. The covariates $X_1, \ldots, X_{40}$ are obtained as $X_j = (Z_{g_j} + R_j)/6$ $(j = 1, \ldots, 40)$. The nonparametric functions are generated in the same way as of Example 1. Therefore, the corresponding coefficients in Example 2 are,

$$\beta = (\underbrace{1.2, \ldots, 1.2}_{8}, \underbrace{3.6, 3.4, 3.2, 3.0, 2.8, 0, 0, 0}_{8}, \underbrace{0, \ldots, 0}_{8}, \underbrace{2.4, 0, 0, 0}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{0, \ldots, 0}_{4})^\top.$$

This example considers three important groups in a more complex structure than that in Example 1. These three groups represent three different settings: all variables within the group are important, many variables within the group are important and very few variables

within the group are important, respectively.

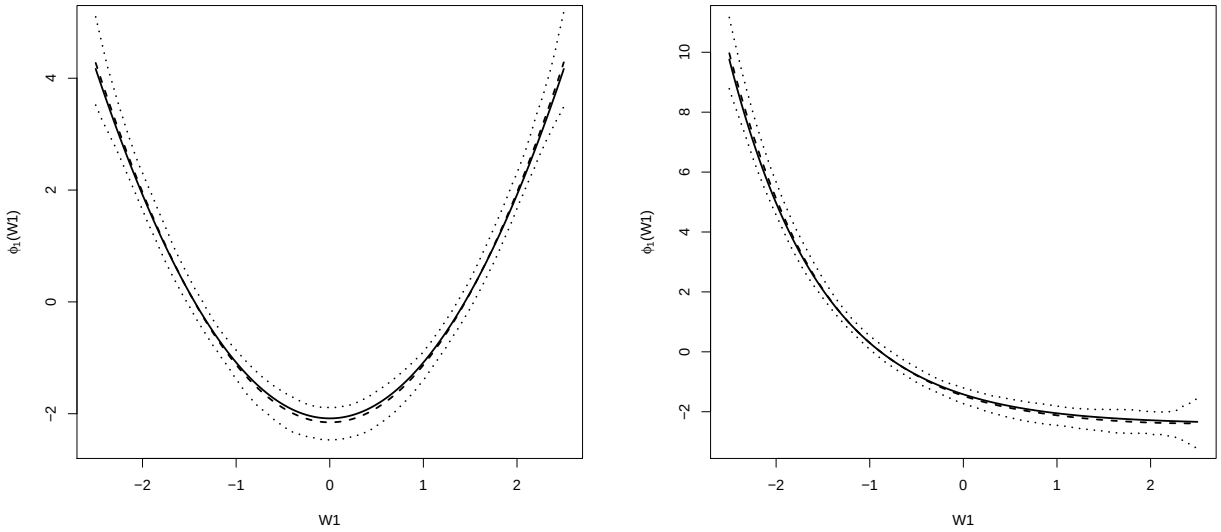
Table 4.1: Simulation results with median and standard deviations (in parentheses) of L_2-PE, TG, FG, TP and FP over 500 simulations for Example 1

	L_2-PE	TG	FG	TP	FP
LASSO	18.79 (5.76)	3 (0.00)	4 (0.73)	15 (0.09)	6 (2.49)
A-LASSO	10.45 (3.95)	3 (0.00)	2 (1.16)	15 (0.37)	3 (2.18)
SCAD	11.18 (4.77)	3 (0.00)	2 (1.17)	15 (0.50)	3 (2.50)
MCP	11.24 (4.75)	3 (0.00)	1 (1.14)	15 (0.71)	1 (1.82)
HP	9.39(3.89)	3 (0.00)	2 (0.99)	15 (0.06)	5 (3.27)
G-SCAD	9.45 (4.53)	3 (0.00)	0 (0.75)	15 (0.00)	0 (3.73)
G-MCP	9.42 (4.45)	3 (0.00)	0 (0.39)	15 (0.00)	0 (1.93)
AHP	8.63 (3.67)	3 (0.60)	0 (0.60)	15 (0.22)	0 (1.00)
Oracle	8.87 (4.26)	3 (0.00)	NA	15 (0.00)	NA

Table 4.2: Simulation results with median and standard deviations (in parentheses) of L_2-PE, TG, FG, TP and FP over 500 simulations for Example 2

	L_2-PE	TG	FG	TP	FP
LASSO	15.74 (4.68)	3 (0.00)	3 (0.92)	14 (0.08)	7 (2.89)
A-LASSO	10.05 (3.33)	3 (0.00)	2 (1.09)	14 (0.49)	4 (2.45)
SCAD	10.84 (4.10)	3 (0.00)	2 (1.24)	14 (0.66)	3 (2.94)
MCP	10.80 (3.84)	3 (0.00)	1 (1.19)	13 (0.95)	2 (2.05)
HP	10.02 (3.82)	3 (0.00)	2 (1.00)	14 (0.06)	10 (3.84)
G-SCAD	9.73 (4.00)	3 (0.00)	0 (0.67)	14 (0.00)	6 (3.39)
G-MCP	9.55 (4.19)	3 (0.06)	0 (0.36)	14 (0.06)	6 (1.70)
AHP	8.55 (3.27)	3 (0.00)	0 (0.64)	14 (0.28)	2 (1.60)
Oracle	8.14 (3.79)	3 (0.00)	NA	14 (0.00)	NA

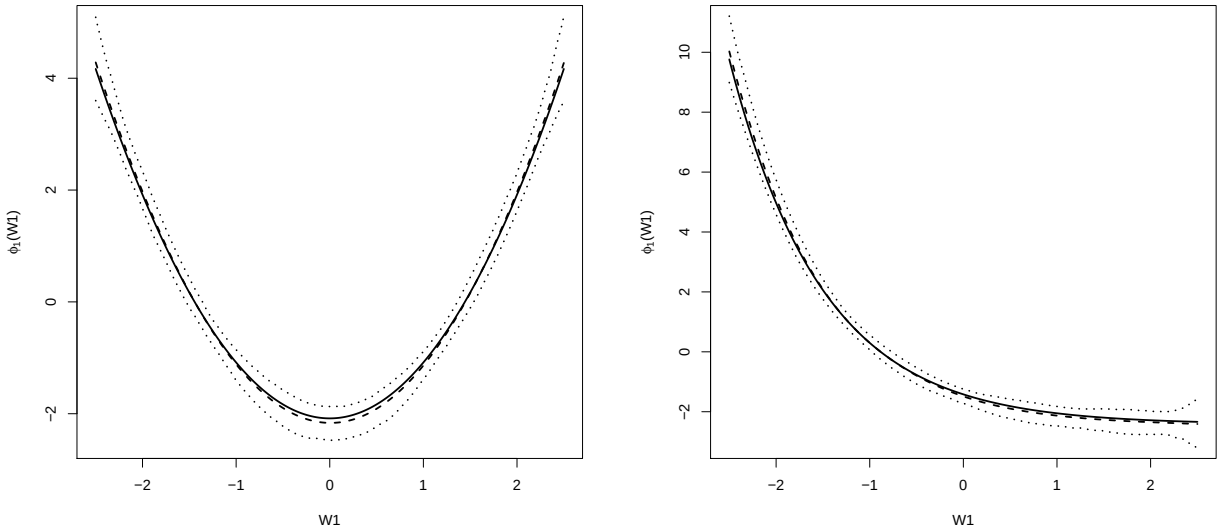
Tables 4.1 and 4.2 summarize variable selection results for Examples 1 and 2 by using the LASSO, A-LASSO, SCAD, MCP, G-SCAD, G-MCP, hierarchical (HP) and adaptive hierarchical (AHP) penalties, respectively. The first four penalties perform individual variable selection, the next three perform group variable selection, and AHP performs adaptive bi-level group selection. From Table 4.1 we see the group variable selection methods perform significantly better than individual variable selection methods with lower L_2-prediction error



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$

Figure 4.1: Estimation of $\phi(\cdot)$'s in Example 1: 95% point-wise confidence bands for $\phi(\cdot)$'s based on 500 replicates. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dot-dashed lines represent the 95% point-wise confidence bands based on 500 estimated values.

and chose more important and less unimportant variables. However, the hierarchical penalty is not performing satisfactorily, it is selecting more groups and more unimportant variables although has the second lowest L_2-prediction error. This performance has been significantly improved in the adaptive version of the penalty, resulting in lowest L_2-prediction error, also, the group and individual variable selection performance is very comparable with the other group selection penalties, G-SCAD and G-MCP. Nonetheless, the superiority of the adaptive hierarchical method stood out with a complex grouping structure among the covariates as shown in Table 4.2. Here, this penalty not only has the smallest L_2-prediction error but also selects significantly lower number of unimportant variables than any other group selection penalties. Hence, if there is known grouping structure available among the covariates, group selection methods are preferable over individual variable selection methods, furthermore, adaptive bi-level group selection should be considered over non-adaptive group selection method especially with a complex grouping structure. The fitted curves and 95% point-wise



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$

Figure 4.2: Estimation of $\phi(\cdot)$'s in Example 2: 95% point-wise confidence bands for $\phi(\cdot)$'s based on 500 replicates. The solid lines stand for the true curves. The dashed lines are the average estimated curves. The dot-dashed lines represent the 95% point-wise confidence bands based on 500 estimated values.

confidence bands for $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are shown in Figures 4.1 and 4.2 for Example 1 and 2, respectively. It is evident that the average estimated curves capture the true curves very well and that the true curves lie in the 95% point-wise confidence bands which is quite narrower.

4.4.2 Application

In this section, we illustrate the application of our proposed method with two real data examples.

4.4.2.1 Primary Biliary Cirrhosis data analysis

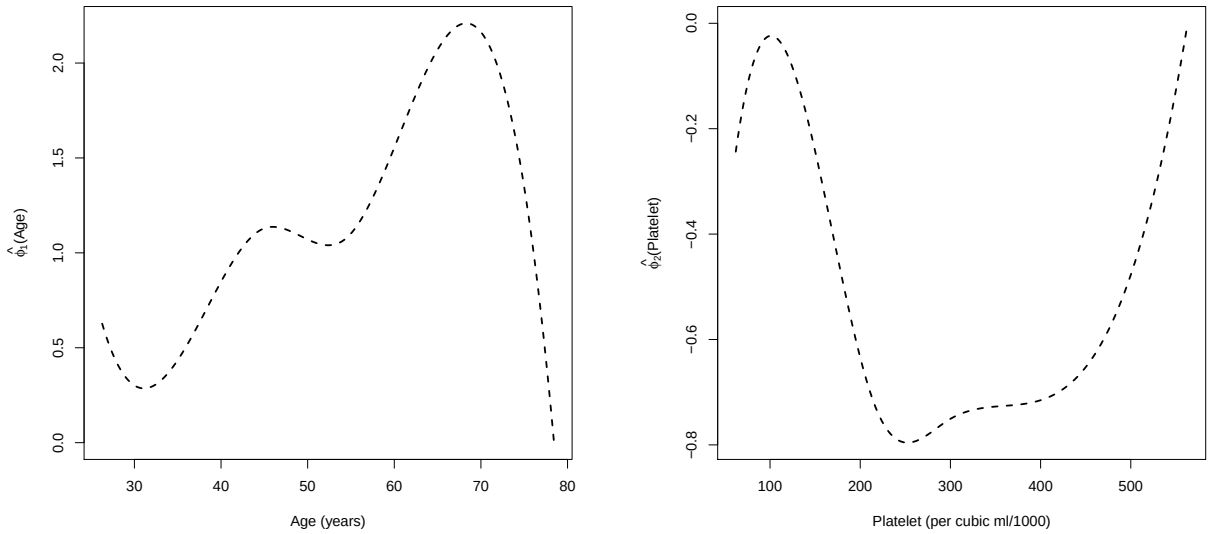
The Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver, a fatal chronic liver disease, was conducted between 1974 and 1984. The data is available in **R** package 'survival'. A total of 424 PBC patients who met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamin were referred to Mayo Clinic during that ten-year interval.

At the end, a total of 312 PBC patients participated in the randomized trial of whom 158 were assigned to the drug D-penicillanmine while the rest were assigned to a control group with placebo drug. During the follow-up 125 patients died due to PBC disease. The primary interest of the study was to investigate the effectiveness of D-penicillanmine in curing PBC disease. Several other covariates such as age, gender, albumin etc. were recorded as baseline covariates at the beginning of the study. Detailed account of the PBC data can be found in Dickson et al. (1989). The PBC data have been analyzed by Huang et al. (2014) for group selection in PHM.

We analyze this data using our proposed adaptive hierarchical penalty (AHP) to identify a smaller set of significant covariates that contribute to the hazards of dying from PBC under a PL-PHM. Our interest is on the main effects of the observed 17 risk factors using 276 complete cases in the full model. Huang et al. (2014) described that these risk factors are clustered into nine categories with 10 continuous and 7 categorical variables (Table 4.3). Since age (Z_1) and platelet (Z_{17}) are the only covariates in groups with a single variable, and are continuous; we consider them as fixed and low dimensional covariates in non-parametric functions, and performed bi-level selection in the rest of the covariates by treating them as linear covariates. We calculated the maximum likelihood estimate (MLE), LASSO, A-LASSO, G-SCAD, G-MCP, HP and AHP estimates. The results are summarized in Table 4.4. All of the methods suggest that gender and treatment should be excluded from the final model which implies treatment (D-penicillanmine) has no effect on curing PBC disease. The performance of group SCAD and group MCP is very similar. The HP selects more variables than AHP. AHP suggest deleting Groups 5 and 8 in addition to gender and treatment.

The two estimated curves for ϕ_1 (Age), ϕ_2 (Platelet) are shown in Figure 4.3, indicating nonlinear effects of age and platelet count on the hazards rate. The age effect $\hat{\phi}_1$(Age) shows that the hazards of death from PBC increases steadily upto about 66 years, and then drops sharply. PBC is a disease of middle aged people, mostly women, with a median age of disease

onset is 50 years (Talwalkar and Lindor, 2003), which explains the increase in risk with aging. The drop in the risk for older population might be due to the fact that as the patients get older, they are more probable of dying from other causes before the incidence of liver failure (Kubota et al., 2009). On the other hand, PBC is associated with low platelet counts (Bassendine et al., 1985). Since all of our participants are PBC patients, it is expected that the hazard of death will decrease as the platelet count increases, as shown in the estimated curve $\hat{\phi}_1(\text{Platelet})$. However, normal platelet count ranges from 150-450 (per cubic microliter/1000) (https://www.hopkinsmedicine.org) and one of the reason of higher hazard beyond the normal limits is inflammatory diseases like liver cirrhosis; which explains the two tails of the curve where the hazard increases with abnormally lower or higher platelet count. Therefore, our analysis provides more detailed nonlinear profiles regarding the effects of age and platelet counts, both of which are of clinical importance.



(a) (b)

Figure 4.3: Estimated curves $\phi_1(Age)$ and $\phi_2(Platelet)$ in the analysis of PBC data.

Table 4.3: PBC data analysis. Dictionary of covariates

Group	Variable	Type	Definition
$G1$: Age	Z_1	C	Age (years)
$G2$: Gender	Z_2	D	Female gender (0 male and 1 female)
$G3$: Phynotype	Z_3	D	Ascites (0 absence and 1 presence)
	Z_4	D	Hepatomegaly (0 absence and 1 presence)
	Z_5	D	Spiders (0 absence and 1 presence)
	Z_6	D	Edemaoed (0 no edema, 0.5 untreated or successfully treated and 1 edema despite diuretic therapy)
$G4$: Liver function damage	Z_7	C	Alkaline phosphatase (units/litre)
	Z_8	C	Sgot (liver enzyme in units/ml)
$G5$: Excretory function of the liver	Z_9	C	Serum bilirubin (mg/dl)
	Z_{10}	C	Serum cholesterol (mg/dl)
	Z_{11}	C	Triglyserides (mg/dl)
$G6$: Liver reserve function	Z_{12}	C	Albumin (g/dl)
	Z_{13}	C	Prothrombin time (seconds)
$G7$: Treatment	Z_{14}	D	D-Penicillamine vs. placebo (1 treatment and 2 control)
$G8$: Reflection	Z_{15}	D	Stage (histological stage of disease, graded 1,2,3 or 4)
	Z_{16}	C	Urine copper (ug/day)
$G9$: Haematology	Z_{17}	C	Platelets (per cubic ml/1000)

Type: type of variable, C: continuous; D: discrete.

Table 4.4: Estimation results of PBC data

Group	Covariates	MLE	LASSO	A-LASSO	G-SCAD	G-MCP	HP	AHP
G_2	gender	-0.4458	0	0	0	0	0	0
G_3	asc	0.6470	0.3665	0	0	0	0.4093	1.1869
	hep	0.0935	0	0	0	0	0.0208	0
	spid	0.2690	0	0	0	0	0.0973	0.5382
	oed	0.2482	0.1931	0.1175	0	0	0.2423	0.4513
G_4	alk	-0.000002	0	0	0	0	0	0
	sgot	0.0038	0	0.0015	0	0	0	0
G_5	bill	0.0963	0.0904	0.1025	0.1289	0.1149	0.0969	0
	chol	0.0004	0	0	0.0006	0.0005	0.0006	0
	trig	-0.0021	0	0	-0.0011	-0.0012	-0.0008	0
G_6	alb	-0.7478	-0.4189	-0.8033	-0.1227	-0.9658	-0.6672	-0.7547
	prot	0.1784	0	0.1004	0.0247	0.2104	0.1126	0
G_7	trt	-0.1476	0	0	0	0	0	0
G_8	stage	0.3517	0.1235	0.3232	0.5835	0.4222	0.3158	0
	cop	0.0039	0.0033	0.0038	0.0052	0.0052	0.0046	0

4.4.2.2 Mantle Cell Lymphoma Data analysis

Mantle cell lymphoma (MCL) is a rare non-Hodgkin B-cell lymphoma which can be at an aggressive form or be more indolent in clinical representation (Rajabi and Sweetenham, 2015). To establish a molecular diagnosis of MCL, clarify its pathogenesis, and to predict the length of survival of these patients, Rosenwald et al. (2003) performed gene expression profiling on the MCL dataset which is available at http://llmpp.nih.gov/MCL. In the dataset, 92 patients were classified as having MCL and the following variables were included:

- Status: patient status at follow up (1 = death, 0 = censored);
- Time: time of follow-up in year;
- INK.ARFdeletion(X_1): deletions of INK4a/ARF (1 = yes, 0 = no);
- ATMdeletion(X_2): deletions of ATM (1 = yes, 0 = no);
- P.53deletion(X_3): deletions of P53 (1 = yes, 0 = no);
- CyclinD.1taqmanresults(X_4): cyclin D1 TaqMan result;
- BMIexpression(X_5): body mass index expression;
- Proliferation.average(X_6): proliferation signature averages.

Ma and Du (2012) performed variable selection in this data set using a partially linear accelerated failure time (AFT) regression model. They selected variables in the linear part without a grouping structure using iterated LASSO and estimated the nonlinear part using a sieve approach. They excluded the covariate Proliferation.average(X_6) from the analysis and included all other covariates X_1 – X_5 in the nonparametric part. In addition, they removed 7 records (patients) with missing covariates; with the rest 85 patients, the censoring rate was 29.4%.

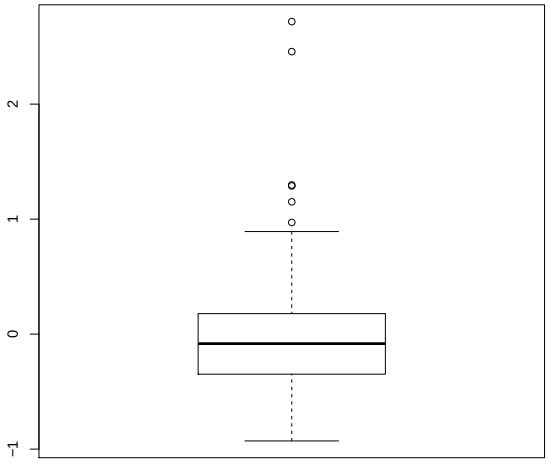
To perform group variable selection in the MCL data using a PL-PHM, we conducted some preliminary diagnosis of the data. Covariates X_1, X_2, and X_3 are binary variables, where

covariates X_4, X_5, and X_6 are continuous. Since variables belonging to the same group usually share some relationships among them, we tested for the significant correlations between the covariates. We tested the correlation between continuous variables by Pearson correlation coefficients; continuous and binary variables by Point-biserial correlation coefficient, and the association between binary variables by Fisher's exact test. The table below illustrates which variables share significant correlations where '✓' indicates significant correlations with the associated sample correlations in the parentheses:

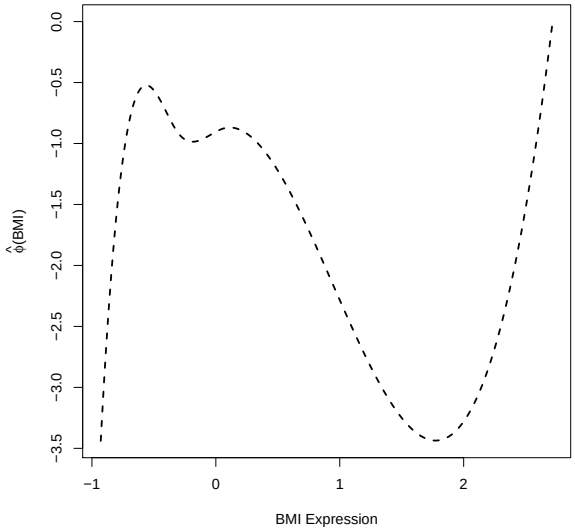
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	✓(1.00)					
X_2	✗ (0.16)	✓(1.00)				
X_3	✗ (0.10)	✗ (0.23)	✓(1.00)			
X_4	✓(0.28)	✗ (-0.10)	✗ (-0.08)	✓(1.00)		
X_5	✗ (-0.17)	✗ (-0.17)	✗ (0.00)	✗ (0.20)	✓(1.00)	
X_6	✓(0.50)	✗ (-0.05)	✓(0.23)	✓(0.41)	✗ (0.17)	✓(1.00)

From the above table we see that X_1, X_4, X_6 shares significant correlation among each other, therefore, we can consider them as a group. X_2 and X_5 do not have significant correlations with any other variables. Note that, X_6 also shares significant correlation with X_3 and they can be considered as a group as well. Thus, X_6 belongs to two overlapping groups; one with (X_1, X_4), another with X_3. However, in this chapter, we assumed covariates can only belong to one group. Therefore, we assign X_6 to the group with (X_1, X_4) based on the strength of the relationship.

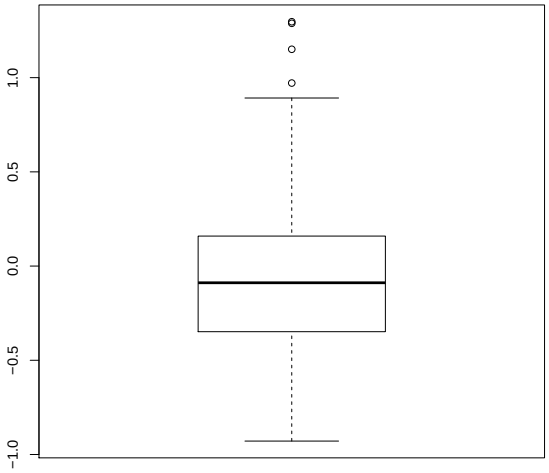
Thus, we have three groups in the linear part of our PL-PHM. Group 1 constitutes of (X_1, X_4, X_6), Group 2 has X_2 and Group 3 has X_3 in it. Similar to the common practice of putting discrete covariates in the linear part and continuous variables in the nonlinear part (Hu and Lian, 2013), we assigned the dichotomous variables in Group 2 and Group 3 in the linear part, and estimated the effect of the continuous variable X_5 on the survival of MCL patients nonparametrically.



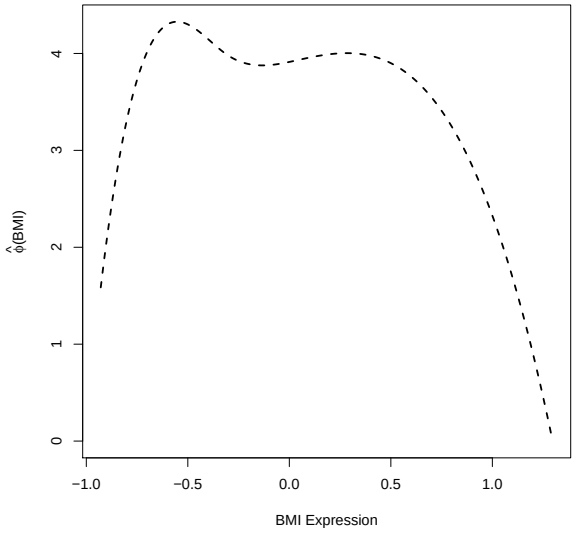
(a) Boxplot of X_5



(b) ϕ(BMI Expression)



(c) Boxplot of X_5 without 2 extreme values



(d) ϕ(BMI Expression) without 2 extreme values

Figure 4.4: Boxplot of BMI Expression and estimated curve of ϕ (BMI Expression) in the analysis of MCL data.

Table 4.5: Estimation results of MCL data

		n = 85				n=83			
Group	Covariates	MLE	LASSO	HP	AHP	MLE	LASSO	HP	AHP
G_1	INK.ARFdeletion (X_1)	-0.33	0	0	0	-0.30	0	0	0
	CyclinD.1taqmanresults (X_4)	1.19	0.89	0.89	0.85	1.16	0.92	0.88	0.82
	Proliferation.average (X_6)	1.82	1.43	1.42	1.48	1.77	1.41	1.39	1.44
G_2	ATMdeletion (X_2)	0.30	0	0	0	0.27	0	0	0
G_5	P.53deletion (X_3)	-0.25	0	0	0	-0.30	0	0	0

Table 4.5 shows the estimation and variable selection performance by four methods. The MLE is the partial maximum likelihood estimates of the linear covariates by maximizing (4.6), where we approximated the nonlinear function using B-splines. For the full data set (n=85), we see that the hierarchical penalty (HP), adaptive hierarchical penalty (AHP) and LASSO discard all the dichotomous variables, $X_1 - X_3$. Similar results were obtained in Ma and Du (2012). Figures 4.4 (a) and (b) present the box-plot and nonlinear profile of BMIexpression(X_5), respectively. We found two extreme outliers which fall outside of upper inner fence ($Q_1 - 3 * IQR$) or upper outer fence ($Q_3 + 3 * IQR$) where Q_1, Q_3 and IQR are first quartile, third quartile and inter-quartile range, respectively. Figures 4.4 (c) and (d) show the boxplot and nonlinear profile of BMIexpression (X_5) after discarding the outliers. In addition, Table 4.5 also shows the variable selection performance when these two extreme values are omitted (n=83). From this comparison analysis, we see that the performance of variable selection is almost the same, but the estimated nonparametric function of X_5 is quite different in the right tail when the two large X_5 values are included. BMI significantly impacts on the overall survival in indolent non-Hodgkins lymphoma and mantle cell lymphoma (Weiss et al., 2017) and obesity is a well-known risk factor for the development of lymphomas (Patel et al., 2013). This may tell the investigators that large BMIexpression(X_5) values could increase the risk of death of MCL patients.

In our model, we included Proliferation.average(X_6) for variable selection which was not incorporated by Ma and Du (2012), and found out it has nonzero effect in estimating the

survival of MCL patients. Both LASSO and AHP showed that X_6 has a strong effect, further investigation can provide additional knowledge of this effect.

4.5 Concluding Remarks

In this chapter, we proposed a hierarchically penalized method for variable selection in the PL-PHM with diverging number of parameters. Our model allows high dimensional linear covariates and fix low dimensional nonparametric covariates to be included in the same model to predict the hazard of failure time, which is more appealing and useful than models with only a linear term or with a large number of nonlinear functions of covariates. We approximated the nonparametric functions using B-splines and performed adaptive bi-level group variable selection in the linear covariates. Our proposed method can effectively remove unimportant groups and select important variables within a group in the linear part, and estimate both parametric and nonparametric components simultaneously. We use the theory of counting processes and martingales to establish the asymptotic convergence and selection consistency of the proposed estimators. We developed computational algorithm for our proposed estimators and presented simulation studies along with two real data analyses. Numerical studies indicate that the adaptive hierarchically penalized method performs better than existing individual variable selection methods (LASSO, Adaptive LASSO SCAD, MCP) as well as non-adaptive group variable selection methods (group SCAD, group MCP and HP penalties), especially in the cases with a complex grouping structure among the covariates. Our computation cost was somewhat high since the computation algorithm takes a while to converge; however, our estimators were precise in terms of estimation accuracy and selection consistency at the cost of high computational time.

We did not get the asymptotic normality of $\hat{\beta}_{\mathcal{A}}$, this remains an open question for our future research to explore. But intuitively, given that our objective function in Q_n is log partial likelihood, following the lines in Wang et al. (2009) and Hu and Lian (2013), we

conjecture that, the estimators for the nonzero coefficients $\hat{\beta}_{\mathcal{A}}$ have the same asymptotic distribution as they would have if the zero coefficients were known in advance, therefore, it possesses the oracle property of Fan and Li (2001).

In applications, it is important to have the goodness of fit procedures available for assessing the model fit. Lin et al. (1993) proposed martingale-based residuals to graphically and numerically check the adequacy of the proportional hazards model with right censored data. Kim and Lee (1998) adopted two methods for model checking of the additive hazards model with right censored data by dividing the data into two groups and testing for the proportional hazards assumption to the additive hazards model to test the monotone departure from the additivity. One method is based on the martingale residuals and the other is based on the difference between weighted estimators of the excess risk. These model-checking techniques were developed for the linear models. In our case, we can consider each B-spline basis function as a covariate in the model, then the proposed model becomes a linear model, and their methods can be applied to choose either the PL-AHM or PL-PHM in practice.

In a Bayesian framework, the variable selection problem can be viewed as the identification of nonzero regression parameters based on the posterior distributions. Bayesian models attempt to avoid the over-fitting problems of frequentist methods by basing predictions on modes of posterior distributions rather than estimators. For uncensored data, bi-level group variable selection using Bayesian selection method has been investigated by Zhang et al. (2014), Xu and Ghosh (2015) and Mallick and Yi (2017). Faraggi and Simon (1998) is one of the first to consider Bayesian variable selection method for censored survival data where they performed individual variable selection in the Cox PH model. Later, Sha et al. (2006) conducted individual variable selection for analyzing microarray data with the AFT model and Lee et al. (2011) performed individual variable selection in the Cox PH model where the shrinkage prior is obtained through a scale mixture representation of normal and gamma distributions and the cumulative baseline hazard function is modeled as a priori by a gamma

process. Recently, Lee et al. (2015) performed group variable selection in the Cox PH model. As it appears in the literature, no variable selection has been investigated on the PL-PHM using Bayesian methods and can be worthy of future research.

4.6 Appendix

The proofs of Lemmas 1-2 follow those of Wang et al. (2009) closely.

Proof of Lemma 1. Let $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$ denote the criterion that we would like to maximize in equation (4.7), let $Q^\dagger(\lambda, \alpha, \gamma, \theta)$ denote the corresponding criterion in equation (4.8), and let $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ denote a local maximizer of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$. We will prove that $(\hat{\alpha}^\dagger = \hat{\alpha}^*, \hat{\gamma}_g^\dagger = \lambda_\gamma \hat{\gamma}_g^*, \hat{\theta}_{(g)}^\dagger = \hat{\theta}_{(g)}^*/\lambda_\gamma)$ is a local maximizer of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$.

Replacing $\gamma^* = \gamma^\dagger/\lambda_\gamma$ and $\theta^* = \theta^\dagger \lambda_\gamma$ in (4.7), we immediately have $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta) = Q^\dagger(\lambda, \alpha, \lambda_\gamma \gamma, \theta/\lambda_\gamma)$. Since $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ is a local maximizer of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$, therefore, by the definition of local maximizer there exists $\delta > 0$ such that if $(\alpha', \gamma', \theta')$ satisfies $|\alpha' - \hat{\alpha}^*| + |\gamma' - \hat{\gamma}^*| + |\theta' - \hat{\theta}^*| < \delta$, then $Q^*(\lambda_\gamma, \lambda_\theta, \alpha', \gamma', \theta') \leq Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$. We choose δ' such that $\delta'/\min(\lambda_\gamma, 1/\lambda_\gamma) \leq \delta/2$. Then, $\min(\lambda_\gamma, 1/\lambda_\gamma) \leq 1$ and $\delta' \leq \min(\lambda_\gamma, 1/\lambda_\gamma)\delta/2 \leq \delta/2$. Thus, for any $(\alpha'', \gamma'', \theta'')$ satisfying $|\alpha'' - \hat{\alpha}^\dagger| + |\gamma'' - \hat{\gamma}^\dagger| + |\theta'' - \hat{\theta}^\dagger| < \delta' \leq \delta/2$, we have, $|\alpha'' - \hat{\alpha}^\dagger| = |\alpha'' - \hat{\alpha}^*| \leq \delta/2$. Also,

$$\left|\frac{\gamma''}{\lambda_\gamma} - \hat{\gamma}^*\right| + \left|\lambda_\gamma \theta'' - \hat{\theta}^*\right| \leq \frac{\lambda_\gamma \left|\frac{\gamma''}{\lambda_\gamma} - \hat{\gamma}^*\right| + \frac{1}{\lambda_\gamma}\left|\lambda_\gamma \theta'' - \hat{\theta}^*\right|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} = \frac{\left|\gamma'' - \lambda_\gamma \hat{\gamma}^*\right| + \left|\theta'' - \frac{\hat{\theta}^*}{\lambda_\gamma}\right|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})}$$
$$= \frac{|\gamma'' - \hat{\gamma}^\dagger| + |\theta'' - \hat{\theta}^\dagger|}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} < \frac{\delta'}{\min(\lambda_\gamma, \frac{1}{\lambda_\gamma})} \leq \frac{\delta}{2}.$$

Therefore, $\left|\alpha'' - \hat{\alpha}^*\right| + \left|\gamma''/\lambda_\gamma - \hat{\gamma}^*\right| + \left|\lambda_\gamma \theta'' - \hat{\theta}^*\right| < \delta/2 + \delta/2 = \delta$. Hence,

$$Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}'', \hat{\gamma}''/\lambda_\gamma, \lambda_\gamma \hat{\theta}'') \leq Q^*(\lambda_\gamma, \lambda_\theta, \hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*),$$

which gives us

$$Q^\dagger(\lambda, \hat{\alpha}'', \hat{\gamma}'', \hat{\theta}'') \leq Q^\dagger(\lambda, \hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger).$$

So, $(\hat{\alpha}^\dagger = \hat{\alpha}^*, \hat{\gamma}^\dagger = \lambda_\gamma \hat{\gamma}^*, \hat{\theta}^\dagger = \hat{\theta}^*/\lambda_\gamma)$ is a local maximizer of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$.

Similarly, we can prove that for any local maximizer $(\hat{\alpha}^\dagger, \hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$, there is a corresponding local maximizer $(\hat{\alpha}^*, \hat{\gamma}^*, \hat{\theta}^*)$ of $Q^*(\lambda_\gamma, \lambda_\theta, \alpha, \gamma, \theta)$ such that $\hat{\alpha}^* = \hat{\alpha}^\dagger$ and $\hat{\gamma}_g^* \hat{\theta}_{gj}^* = \hat{\gamma}_g^\dagger \hat{\theta}_{gj}^\dagger$. □

Proof of Lemma 2. Suppose $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local maximizer of (4.8). Let $\hat{\beta}$ satisfy $\hat{\beta}_{gj} = \hat{\gamma}_g \hat{\theta}_{gj}$. It is trivial that $\hat{\gamma}_g = 0$ if and only if $\hat{\theta}_{(g)} = 0$. Hence, if $\hat{\gamma}_g \neq 0$, then $|\hat{\beta}_{(g)}| \neq 0$.

Let (α, β) be fixed at $(\hat{\alpha}, \hat{\beta})$. Then maximizing $Q^\dagger(\lambda, \alpha, \gamma, \theta)$ in (4.8) only depends on the penalty. For some g with $|\hat{\beta}_{(g)}| \neq 0$, the corresponding penalty term is $-\gamma_g - \lambda \sum_{j=1}^{p_g} |\hat{\beta}_{gj}|/\gamma_g$, which is maximized at $\hat{\gamma}_g = (\lambda |\hat{\beta}_{(g)}|)^{1/2}$, and $\hat{\theta}_{(g)} = \hat{\beta}_{(g)}/\hat{\gamma}_g$.

Let $Q(\lambda, \alpha, \beta)$ be the corresponding criterion to be maximized in equation (4.9). By Lemma 1, the local maximizer $\hat{\alpha}$ of α in (4.7) and (4.8) are the same, so we only need to consider other parameters, e.g., β, and fix α at $\hat{\alpha}$ in both (4.7) and (4.8). We first show that $(\hat{\alpha}, \hat{\beta})$ is a local maximizer of $Q(\lambda, \alpha, \beta)$, i.e., there exists a $\delta' > 0$ such that if $|\Delta\alpha| + |\Delta\beta| < \delta'$, then $Q(\lambda, \hat{\alpha} + \Delta\alpha, \hat{\beta} + \Delta\beta) \leq Q(\lambda, \hat{\alpha}, \hat{\beta})$. Particularly, taking $\Delta\alpha = 0$, it becomes $|\Delta\beta| < \delta'$, then $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta) \leq Q(\lambda, \hat{\alpha}, \hat{\beta})$. Denote $\Delta\beta = \Delta\beta^{(1)} + \Delta\beta^{(2)}$, where $\Delta\beta_{(g)}^{(1)} = 0$ if $|\hat{\beta}_{(g)}| = 0$ and $\Delta\beta_{(g)}^{(2)} = 0$ if $|\hat{\beta}_{(g)}| \neq 0$. We thus, have $|\Delta\beta| = |\Delta\beta^{(1)} + \Delta\beta^{(2)}| = |\Delta\beta^{(1)}| + |\Delta\beta^{(2)}|$.

We first show $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta})$ for some δ'. We already have $\hat{\gamma}_g = (\lambda |\hat{\beta}_{(g)}|)^{1/2}$ and $\hat{\theta}_{(g)} = \hat{\beta}_{(g)}/\hat{\gamma}_g$ if $|\hat{\gamma}_g| \neq 0$, and $\hat{\theta}_{(g)} = 0$ if $|\hat{\gamma}_g| = 0$. Let $\hat{\gamma}_g' = (\lambda |\hat{\beta}_{(g)} + \Delta\beta_{(g)}^{(1)}|)^{1/2}$ and $\hat{\theta}_{(g)}' = (\hat{\beta}_{(g)} + \Delta\beta_{(g)}^{(1)})/\hat{\gamma}_g'$ if $|\hat{\gamma}_g| \neq 0$, and let $\hat{\gamma}_g' = 0$ and $\hat{\theta}_{(g)}' = 0$ if $|\hat{\gamma}_g| = 0$. Then we have $Q^\dagger(\lambda, \hat{\alpha}, \hat{\gamma}', \hat{\theta}') = Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$ and $Q^\dagger(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta}) = Q(\lambda, \hat{\alpha}, \hat{\beta})$. Hence, we only need to show $Q^\dagger(\lambda, \hat{\alpha}, \hat{\gamma}', \hat{\theta}') \leq Q^\dagger(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta})$. As $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local maximizer of $Q^\dagger(\lambda, \alpha, \gamma, \theta)$, for fixed $\hat{\alpha}$, there exists a δ such that for any (γ', θ') satisfying $|\gamma' - \hat{\gamma}| + |\theta' - \hat{\theta}| < \delta$, we have $Q^\dagger(\lambda, \hat{\alpha}, \gamma', \theta') \leq Q^\dagger(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta})$. Let $a = \min\left\{|\hat{\beta}_{(g)}| : |\beta_{(g)}| \neq 0,\ g = 1, \ldots, G\right\}, b =$

$\max\left\{|\hat{\beta}_{(g)}| : |\beta_{(g)}| \neq 0,\ g = 1, \ldots, G\right\}$ and $\delta^{'} < a/2$. It is seen that,

$$\begin{aligned}
\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| - |\hat{\beta}_{(g)}|\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left|(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2})^2 - (|\hat{\beta}_{(g)}|^{1/2})^2\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left|(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2})(|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2})\right| &\leq \left|\Delta\beta^{(1)}_{(g)}\right|, \\
\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2}\right| &\leq \frac{\left|\Delta\beta^{(1)}_{(g)}\right|}{|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2}}.
\end{aligned}$$

Since when $\min_g \left\{|\hat{\beta}_{(g)}|\right\} = a \neq 0$, and when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'} < a/2$, we have

$$|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| \geq |\hat{\beta}_{(g)}| - |\Delta\beta^{(1)}_{(g)}| \geq a - \frac{a}{2} = \frac{a}{2} > 0,$$

and

$$|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} + |\hat{\beta}_{(g)}|^{1/2} \geq \left(\frac{a}{2}\right)^{1/2} + a^{1/2} = (2^{-1/2} + 1)a^{1/2} \geq 2^{1/2}a^{1/2} = (2a)^{1/2}.$$

Therefore,

$$\left||\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|^{1/2} - |\hat{\beta}_{(g)}|^{1/2}\right| \leq \frac{|\Delta\beta^{(1)}_{(g)}|}{(2a)^{1/2}}.$$

Hence,

$$|\hat{\gamma}_g^{'} - \hat{\gamma}_g| = \left|(\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2} - (\lambda|\hat{\beta}_{(g)}|)^{1/2}\right| \leq \frac{\lambda|\Delta\beta^{(1)}_{(g)}|}{(2\lambda a)^{1/2}}.$$

Next, if $|\hat{\gamma}_g| = 0$, then $\hat{\theta}^{'}_{(g)} = \hat{\theta}_{(g)} = 0$, and $|\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)}| = 0$. If $|\hat{\gamma}_g| \neq 0$, then

$$\begin{aligned}
\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)} &= \frac{(\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)})}{\hat{\gamma}^{'}_g} - \frac{\hat{\beta}_{(g)}}{\hat{\gamma}_g} \\
&= \frac{\hat{\beta}_{(g)}\hat{\gamma}_g + \Delta\beta^{(1)}_{(g)}\hat{\gamma}_g - \hat{\beta}_{(g)}\hat{\gamma}^{'}_g}{\hat{\gamma}^{'}_g\hat{\gamma}_g} \\
&= \frac{\hat{\beta}_{(g)}[\hat{\gamma}_g - \hat{\gamma}^{'}_g] + \Delta\beta^{(1)}_{(g)}\hat{\gamma}_g}{\hat{\gamma}^{'}_g\hat{\gamma}_g}. \qquad (4.14)
\end{aligned}$$

We already have $|\hat{\beta}_{(g)}| \leq b$ and $|\hat{\gamma}^{'}_g - \hat{\gamma}_g| \leq \lambda|\Delta\beta^{(1)}_{(g)}|/(2\lambda a)^{1/2}$. Consider

$$\hat{\gamma}^{'}_g\hat{\gamma}_g = (\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2}(\lambda|\hat{\beta}_{(g)}|)^{1/2}.$$

Since $|\hat{\gamma}_g| = (\lambda|\hat{\beta}_{(g)}|)^{1/2} \geq \lambda^{1/2}a^{1/2}$, when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'}$ and $\delta^{'} < a/2$, if $\hat{\gamma}_g \neq 0$, then $|\hat{\beta}_{(g)}| \neq 0$, $\Delta\beta^{(2)}_{(g)} = 0$, it implies, $|\Delta\beta^{(1)}_{(g)}| \leq |\Delta\beta^{(1)}| < \delta \Rightarrow |\Delta\beta^{(1)}_{(g)}| < \delta^{'} < a/2$ and $|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}| \geq |\hat{\beta}_{(g)}| - |\Delta\beta^{(1)}_{(g)}| \geq a - a/2 = a/2 > 0$. Therefore, $|\hat{\gamma}^{'}_g| = (\lambda|\hat{\beta}_{(g)} + \Delta\beta^{(1)}_{(g)}|)^{1/2} \geq \lambda^{1/2}(a/2)^{1/2}$ and $|\hat{\gamma}^{'}_g\hat{\gamma}_g| \geq \lambda^{1/2}a^{1/2}\lambda^{1/2}(a/2)^{1/2} = \lambda a 2^{-1/2}$. From (4.14) we have,

$$\begin{aligned}
|\hat{\theta}^{'}_{(g)} - \hat{\theta}_{(g)}| &\leq \frac{|\hat{\beta}_{(g)}|}{|\hat{\gamma}^{'}_g\hat{\gamma}_g|}|\hat{\gamma}^{'}_g - \hat{\gamma}_g| + |\Delta\beta^{(1)}_{(g)}|\frac{|\hat{\gamma}_g|}{|\hat{\gamma}_g||\hat{\gamma}^{'}_g|} \\
&\leq \frac{b\lambda|\Delta\beta^{(1)}_{(g)}|}{(2\lambda a)^{1/2}(\lambda a 2^{-1/2})} + |\Delta\beta^{(1)}_{(g)}|\frac{1}{\lambda^{1/2}(a/2)^{1/2}} \\
&\leq \left[\frac{b\lambda}{(2\lambda a)^{1/2}(\lambda a)2^{-1/2}} + \frac{1}{(\lambda a/2)^{1/2}}\right]|\Delta\beta^{(1)}_{(g)}| \\
&= |\Delta\beta^{(1)}_{(g)}|\left[\frac{1}{(\lambda a/2)^{1/2}} + \frac{b}{a(\lambda a)^{1/2}}\right].
\end{aligned}$$

Therefore, we are able to choose a $\delta^{'} > 0$ satisfying $\delta^{'} < a/2$ such that $|\hat{\gamma}^{'}_g - \hat{\gamma}_g| + |\hat{\theta}^{'}_g - \hat{\theta}_g| < \delta$ when $|\Delta\beta^{(1)}_{(g)}| < \delta^{'}$. Hence we have $Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}^{'}, \hat{\theta}^{'}) \leq Q^{\dagger}(\lambda, \hat{\alpha}, \hat{\gamma}, \hat{\theta})$ due to the local maximality, that is, $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta})$.

Next we show $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$. This is trivial when $\Delta\beta^{(2)} = 0$. If $\Delta\beta^{(2)} \neq 0$, then $\Delta\beta^{(1)} = 0$ and we have

$$Q(\lambda, \hat{\alpha}, \hat{\beta}+\Delta\beta^{(1)}+\Delta\beta^{(2)}) - Q(\lambda, \hat{\alpha}, \hat{\beta}+\Delta\beta^{(1)}) = (\Delta\beta^{(2)})^{\top}n^{-1}\frac{\partial l_n(\hat{\alpha}, \beta^*)}{\partial\beta} - 2\sum_{g=1}^{G}(\lambda|\Delta\beta^{(2)}_{(g)}|)^{1/2},$$

where β^* is a vector between $\hat{\beta}+\Delta\beta^{(1)}+\Delta\beta^{(2)}$ and $\hat{\beta}+\Delta\beta^{(1)}$. Since $|\Delta\beta^{(2)}| < \delta^{'}$, for a small enough $\delta^{'}$, the second term in the above equality dominates the first term, hence we have $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)} + \Delta\beta^{(2)}) \leq Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta^{(1)})$. Thus we have shown that there exists a $\delta^{'} > 0$ such that if $|\Delta\beta| < \delta^{'}$, then $Q(\lambda, \hat{\alpha}, \hat{\beta} + \Delta\beta) \leq Q(\lambda, \hat{\alpha}, \hat{\beta})$, which implies that $\hat{\beta}$ is a local maximizer of $Q(\lambda, \hat{\alpha}, \beta)$.

Similarly, we can prove that if $(\hat{\alpha}, \hat{\beta})$ is a local maximizer of $Q(\lambda, \alpha, \beta)$, then $(\hat{\alpha}, \hat{\gamma}, \hat{\theta})$ is a local maximizer of $Q^{\dagger}(\lambda, \alpha, \gamma, \theta)$, where $\hat{\gamma}_g = (\lambda|\hat{\beta}_{(g)}|)^{1/2}$ and $\hat{\theta}_{(g)} = \hat{\beta}_{(g)}/\hat{\gamma}_g$ if $|\hat{\beta}_{(g)}| \neq 0$, and $\hat{\gamma}_g = 0$ and $\hat{\theta}_{(g)} = 0$ if $|\hat{\beta}_{(g)}| = 0$. □

Proof of Theorem 1. Let $\alpha^0 = (\alpha_1^{0^{\top}}, \ldots, \alpha_Q^{0^{\top}})^{\top}$ be a QK dimensional vector that satisfies $\left\|\phi_j^0 - \alpha_j^{0^{\top}}B_j\right\|_{\infty} = O(K^{-d}), 1 \leq j \leq Q$. Then, $\left\|\phi^0 - \alpha^{0^{\top}}B\right\|_{\infty} = O(K^{-d})$ and $\left\|\phi^0 - \alpha^{0^{\top}}B\right\| =$

$O(K^{-d})$ since Q is fixed. Such approximation rates are possible due to our smoothness assumption (B2) and well known approximation properties of B-spline (De Boor, 1978).

Let $\gamma_n = \sqrt{(K+p)/n} + K^{-d}$ and $u \in \mathbb{R}^{QK+p}$ with $\|u\| = D$, where $u = (u_1, u_2)$, u_1 is a QK-vector, and u_2 is a p-vector. To prove Theorem 1, we first show that $\left\|\hat{\phi} - \alpha^{0^\top} B\right\| = O_p(\gamma_n)$, and $\left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$ where $\hat{\phi} = \hat{\alpha}^{0^\top} B$. Then it is sufficient to show that for any $\epsilon > 0$, there exists a constant D such that

$$P\left\{\sup_{\|u\|=D} Q_n((\alpha^0, \beta^0) + \gamma_n u) < Q_n(\alpha^0, \beta^0)\right\} \geq 1 - \epsilon, \tag{4.15}$$

when n is big enough. This implies that with probability of at least $1-\epsilon$, there exists a local maximum in the ball $\{(\alpha^0, \beta^0) + \gamma_n u : \|u\| \leq D\}$. Hence, there exists a local maximizer such that $\left\|\hat{\phi} - \alpha^{0^\top} B\right\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$.

Since p_{λ_n} satisfies conditions (4.11) and (4.12), we have,

$$\begin{aligned}
&Q_n((\alpha^0, \beta^0) + \gamma_n u) - Q_n(\alpha^0, \beta^0) \\
&= n^{-1}\left\{l_n((\alpha^0, \beta^0) + \gamma_n u) - l_n(\alpha^0, \beta^0)\right\} \\
&- \sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0 + \gamma_n u_{2,g1}\right|, \ldots, \left|\beta_{gs_g}^0 + \gamma_n u_{2,gs_g}\right|, \left|\beta_{g(s_g+1)}^0 + \gamma_n u_{2,g(s_g+1)}\right|, \ldots, \left|\beta_{gp_g}^0 + \gamma_n u_{2,gp_g}\right|\right)\right. \\
&\left. - p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|, \ldots, \left|\beta_{gs_g}^0\right|, \left|\beta_{g(s_g+1)}^0\right|, \ldots, \left|\beta_{gp_g}^0\right|\right)\right\} \\
&- \sum_{g=s+1}^{G}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0 + \gamma_n u_{2,g1}\right|, \ldots, \left|\beta_{gp_g}^0 + \gamma_n u_{2,gp_g}\right|\right) - p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|, \ldots, \left|\beta_{gp_g}^0\right|\right)\right\} \\
&\leq n^{-1}\left\{l_n((\alpha^0, \beta^0) + \gamma_n u) - l_n(\alpha^0, \beta^0)\right\} \\
&- \sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0 + \gamma_n u_{2,g1}\right|, \ldots, \left|\beta_{gs_g}^0 + \gamma_n u_{2,gs_g}\right|, \left|\beta_{g(s_g+1)}^0 + \gamma_n u_{2,g(s_g+1)}\right|, \ldots, \left|\beta_{gp_g}^0 + \gamma_n u_{2,gp_g}\right|\right)\right. \\
&\left. - p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|, \ldots, \left|\beta_{gs_g}^0\right|, \left|\beta_{g(s_g+1)}^0\right|, \ldots, \left|\beta_{gp_g}^0\right|\right)\right\} \\
&\leq n^{-1}\left\{l_n((\alpha^0, \beta^0) + \gamma_n u) - l_n(\alpha^0, \beta^0)\right\} \\
&- \sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0 + \gamma_n u_{2,g1}\right|, \ldots, \left|\beta_{gs_g}^0 + \gamma_n u_{2,gs_g}\right|, 0\right) - p_{\lambda_n}^{(g)}\left(\left|\beta_{g1}^0\right|, \ldots, \left|\beta_{gs_g}^0\right|, 0\right)\right\} \\
&= A - B.
\end{aligned}$$

For A, denote $\omega^0 = (\alpha^0, \beta^0)$. By Taylor expansion at $\gamma_n = 0$, we have

$$
\begin{aligned}
n^{-1}\, l_n(\omega^0 + \gamma_n u) &= n^{-1}\, l_n(\omega^0) + n^{-1}\, \gamma_n U(\omega^0)^\top u + (2n)^{-1}\gamma_n^2 u^\top \frac{\partial U(\omega^0)}{\partial \omega^0} u + A_n \\
A &= n^{-1}\left\{ l(\omega^0 + \gamma_n u) - l_n(\omega^0) \right\} \\
&= n^{-1}\, \gamma_n U(\omega^0)^\top u + (2n)^{-1}\gamma_n^2 u^\top \frac{\partial U(\omega^0)}{\partial \omega^0} u + A_n \\
&\triangleq A_1 + A_2 + A_n, \qquad (4.16)
\end{aligned}
$$

where $A_n = (6n)^{-1}\sum_{j,k,l}(\omega_j - \omega_j^0)(\omega_k - \omega_k^0)(\omega_l - \omega_l^0)(\partial^2 U_l(\tilde{\omega})/\partial\omega_j\partial\omega_k)$, U_l is the l-th component of U, and $\tilde{\omega}$ is a value between ω^0 and $\omega = \omega^0 + \gamma_n u$. We first consider

$$
U(\omega^0) = \sum_i \int_0^\tau \left\{ L_i - \frac{S_n^{(1)}(m_n^0, t)[L]}{S_n^{(0)}(m_n^0, t)} \right\} dN_i(t), \quad \text{where } \; m_n^0(R) = {\alpha^0}^\top Z + {\beta^0}^\top X.
$$

Observe,

$$
\begin{aligned}
&\sum_i \left\{ L_i - \frac{S_n^{(1)}(m_n^0, t)[L]}{S_n^{(0)}(m_n^0, t)} \right\} Y_i(t) \exp\left\{\omega^\top L_i\right\} h_0(t) \\
&= \sum_i \left\{ L_i - \frac{\sum_i L_i Y_i(t) \exp\left\{\omega^\top L_i\right\}}{\sum_i Y_i(t) \exp\left\{\omega^\top L_i\right\}} \right\} Y_i(t) \exp\left\{\omega^\top L_i\right\} h_0(t) \\
&= 0.
\end{aligned}
$$

Since

$$
M_i(t) = N_i(t) - \int_0^\tau Y_i(t) \exp\left\{\omega^\top L_i\right\} h_0(t) dt,
$$

this implies that,

$$
U(\omega^0) = \sum_i \int_0^\tau \left\{ L_i - \frac{S_n^{(1)}(m_n^0, t)[L]}{S_n^{(0)}(m_n^0, t)} \right\} dM_i(t). \qquad (4.17)
$$

Similar to Lemma 5.3 of Huang (1999), we have

$$
P_{\Delta n}\left[\frac{S_n^{(1)}(m_n^0, t)[L]}{S_n^{(0)}(m_n^0, t)} - \frac{S_n^{(1)}(m^0, t)[L]}{S_n^{(0)}(m^0, t)} \right] = P_\Delta \left[\frac{s^{(1)}(m_n^0, t)[L]}{s^{(0)}(m_n^0, t)} - \frac{s^{(1)}(m^0, t)[L]}{s^{(0)}(m^0, t)} \right] + o_p(n^{-1/2}), \qquad (4.18)
$$

where $m^0(R) = \phi^0(W) + \beta^{0^\top} X$. Let $s(m_n^0, t) = s^{(1)}(m_{n,}^0 t)[L]/s^{(0)}(m_n^0, t)$ and $s(m^0, t) = s^{(1)}(m^0, t)[L]/s^{(0)}(m^0, t)$. By Taylor series expansion, for some ξ between m^0 and m_n^0 we have

$$s(m_n^0, t) - s(m^0, t) = \frac{\partial s(m^0, t)}{\partial m^0}(m_n^0 - m^0) + \frac{1}{2}\frac{\partial^2 s(\xi, t)}{\partial \xi^2}(m_n^0 - m^0)^2$$

$$\left|s(m_n^0, t) - s(m^0, t)\right| \leq \left|\frac{\partial s(m^0, t)}{\partial m^0}\ d\right| + \left|\frac{1}{2}\frac{\partial^2 s(\xi, t)}{\partial \xi^2}\ d^2\right|,$$

where $d = m_n^0 - m_0$. Let $W(t) = Y(t) \exp\left(m^0, t\right) / s^{(0)}(m^0, t)$. Then, by Lemma A.4 of Huang (1999), we have

$$\begin{aligned}
\left|\frac{\partial s(m^0, t)}{\partial m^0}\ d\right|^2 &\leq \left|E\left\{W(t)h(R)d(R)\right\} - E\left\{W(t)h(R)\right\} E\left\{W(t)d(R)\right\}\right|^2 \\
&= \left|E\left[W(t)\left\{h(R) - E(W(t)h(R))\right\}\left\{d(R) - E(W(t)d(R))\right\}\right]\right|^2 \\
&= \left[E\left\{K_1 d(R) - K_1 E(K_2 d(R))\right\}\right]^2 \\
&\leq 2E\left\{K_1 d(R)\right\}^2 + 2E\left\{K_1 E(K_2 d(R))\right\}^2 \\
&\leq 2E\left\{K_1^2\right\} E\left\{d^2(R)\right\} + 2E\left\{K_3^2\right\} E\left\{d^2(R)\right\} \\
&= K_4 E\left\{d^2(R)\right\} \\
&= K_4 \left\|d\right\|^2 \\
&= K_4 \left\|m_n^0 - m_0\right\|^2 \\
&= O_p\left(\left\|m_n^0 - m_0\right\|^2\right).
\end{aligned}$$

Therefore, from the approximation rate given in the beginning of the proof of Theorem 1, we have,

$$\left|\frac{\partial s(m^0, t)}{\partial m^0}\ d\right| \leq O_p\left(\left\|m_n^0 - m_0\right\|\right) = O_p(K^{-d}).$$

Similarly, using Lemma A.4 of Huang (1999) gives us

$$\left|\frac{1}{2}\frac{\partial^2 s(\xi, t)}{\partial \xi^2}\ d^2\right| = O_p(K^{-2d}).$$

Therefore, from (4.18) we have,

$$P_{\Delta n}\left[\frac{S_n^{(1)}(m_n^0, t)[L]}{S_n^{(0)}(m_n^0, t)} - \frac{S_n^{(1)}(m^0, t)[L]}{S_n^{(0)}(m^0, t)}\right] = O(K^{-d}) + o_p(n^{-1/2}) = O_p(K^{-d}).$$

Consequent, from (4.17), we obtain

$$\begin{aligned} U(\omega^0) &= \sum_i \int_0^\tau \left\{ L_i - \frac{S_n^{(1)}(m^0,t)[L]}{S_n^{(0)}(m^0,t)} \right\} dM_i(t) + O_p\left(nK^{-d}\right) \\ &= \sum_i \int_0^\tau \left\{ L_i - G_n(m^0,t)[L] \right\} dM_i(t) + O_p\left(nK^{-d}\right) \\ &= \xi_n + O_p(nK^{-d}), \end{aligned}$$

where $\xi_n = \sum_i \int_0^\tau \{L_i - G_n(m^0,t)[L]\} dM_i(t)$. Direct algebraic calculations show that, $E\left\{\|\xi_n\|^2\right\} = E\left\{tr\left(\xi_n^\top \xi_n\right)\right\} = tr\left\{E\left(\xi_n^\top \xi_n\right)\right\} = tr\left\{E\left(\|\xi_n\|^2\right)\right\}$. Let,

$$\xi_n = \sum_i \int_0^\tau \left\{ L_i - G_n(m^0,t)[L] \right\} dM_i(t) = \sum_i \int_0^\tau H_i(t) dM_i(t),$$

where $H_i(t) = \{L_i - G_n(m^0,t)[L]\}$. Since ξ_n is a martingale integral, we have $E(\xi_n|\mathcal{F}_t^-) = 0$ where $\mathcal{F}_t^-$ denotes the past up to the beginning of the small time interval $[t, t+dt)$, and

$$\begin{aligned} V(\xi_n|\mathcal{F}_t^-) &= E(\xi_n^{\otimes 2}|\mathcal{F}_t^-) \\ &= E(\xi_n \xi_n^\top|\mathcal{F}_t^-) \\ &= E\sum_i \int_0^\tau \mathrm{Var}\left\{H_i(t) dM_i(t)|\mathcal{F}_t^-\right\} \\ &= E\sum_i \int_0^\tau H_i(t)^{\otimes 2} d\left\langle M \right\rangle(t) \\ &= E\int_0^\tau \sum_i \left\{L_i - G_n(m^0,t)[L]\right\}^{\otimes 2} \Lambda_i(t) dt, \end{aligned}$$

where $\Lambda_i(t) = h_0(t) Y_i(t) \exp\left\{\phi^0(W) + X^\top \beta^0\right\}$. We can show that

$$\begin{aligned} &\sum_i \left\{L_i - G_n(m^0,t)[L]\right\}^{\otimes 2} Y_i(t) \exp\left\{m^0(R_i)\right\} \\ &= \sum_i L_i^{\otimes 2} Y_i(t) \exp\left\{m^0(R_i)\right\} - \sum_i G_n(m^0,t)[L]^{\otimes 2} Y_i(t) \exp\left\{m^0(R_i)\right\} \\ &\le \sum_i L_i^{\otimes 2} Y_i(t) \exp\left\{m^0(R_i)\right\}. \end{aligned}$$

Then, $V(\xi_n|\mathcal{F}_t^-) \le E\int_0^\tau \sum_i L_i^{\otimes 2} \Lambda_i(t) dt$. Assume $\sup_{t,W,X} |h_0(t) \exp\left\{\phi^0(W) + \beta^{0\top} X\right\}| \le \tilde{M}$.

Therefore,

$$\begin{aligned} E\left\{\|\xi_n\|^2\right\} &= \mathrm{tr}\left[E\left\{\int_0^\tau \sum_i \left(L_i - G_n(m^0,t)[L]\right)^{\otimes 2} \Lambda_i(t)dt\right\}\right] \\ &\leq \tilde{M}\left[E\left\{\int_0^\tau \sum_i \left(L_i - G_n(m^0,t)[L]\right)^{\otimes 2} Y_i(t)dt\right\}\right] \\ &\leq nE\left\{\mathrm{tr}L_i^{\otimes 2} Y_i(t)\right\}. \end{aligned}$$

By condition (B4), we have,

$$\|\xi_n\| = O_p(\sqrt{n(K+p)}),$$

and

$$\left\|U(\omega^0)\right\| = O_p(\sqrt{n(K+p)} + nK^{-d}). \tag{4.19}$$

Consequently, from (4.16),

$$A_1 = \gamma_n O_p(\gamma_n)\|u\| = O_p(\gamma_n^2)\|u\|.$$

Next, for A_2, we already have,

$$\begin{aligned} U(\omega^0) &= \sum_i \int_0^\tau \left\{L_i - \frac{S_n^{(1)}(m^0,t)[L]}{S_n^{(0)}(m^0,t)}\right\} dN_i(t) + O_p(nK^{-d}), \\ \frac{\partial U(\omega^0)}{\partial \omega^0} &= -\sum_i \int_0^\tau \left[\frac{S_n^{(0)}(m^0,t)S_n^{(2)}(m^0,t)[L] - \left\{S_n^{(1)}(m^0,t)[L]\right\}^{\otimes 2}}{\left\{S_n^{(0)}(m^0,t)\right\}^2}\right] dN_i(t) + O_p(nK^{-d}) \\ &= -\sum_i \int_0^\tau V(m^0,t)[L]\left\{dM_i(t) + \lambda(t|L)dt\right\} + O_p(nK^{-d}) \\ &= -\left\{\sum_i \int_0^\tau V_n(m^0,t)[L]dM_i(t) + \sum_i \int_0^\tau V_n(m^0,t)[L]S_n^{(0)}(m^0,t)\lambda_0(t)dt\right\} + O_p(nK^{-d}) \\ &= -n\left(\vartheta_{\omega^0} + \Sigma_n\right) + O_p(nK^{-d}), \end{aligned}$$

where $\vartheta_{\omega^0} = n^{-1}\sum_i \int_0^\tau V_n(m^0,t)[L]dM_i(t)$ and $\Sigma_n = n^{-1}\sum_i \int_0^\tau V_n(m^0,t)[L]S^{(0)}(m^0,t)\lambda_0(t)dt$.

Thus,

$$A_2 = -(1/2)\gamma_n^2\left\{u^\top\left(n^{-1}\frac{\partial U(\omega^0)}{\partial \omega^0}\right)u\right\} = -(1/2)\gamma_n^2\left[u^\top\Sigma u + u^\top\left\{(\Sigma_n - \Sigma) + \vartheta_{\omega^0}\right\}u + u^\top O(K^{-d})u\right].$$

By Lemmas 2.3 and 4.1 of Bradic et al. (2011),

$$\|(\Sigma_n - \Sigma) + \vartheta_{\omega^0}\| \leq \|\Sigma_n - \Sigma\| + \|\vartheta_{\omega^0}\| = o_p(1).$$

Since Σ is positive definite and its eigen values are bounded away from zero and infinity, we have,

$$A_2 = -(1/2)\gamma_n^2(1 + o_p(1) + O(K^{-d}))\,\|u\|^2\,.$$

Finally, since $\|\omega - \omega^0\|_2 \leq \gamma_n$ and the average of i.i.d. terms, $n^{-1}\frac{\partial^2 U_l(\tilde{\omega})}{\partial\omega_j\partial\omega_k}$, is of order $O_p(1)$, by the Cauchy-Schwarz inequality and condition $\gamma_n(K+p)^{3/2} = O(1)$ from Theorem 1, we have $A_n = (K+p)^{3/2}O_p(\gamma_n^3) = O_p(\gamma_n^2)$.

For the penalty part, by Taylor expansion of the penalty function we have,

$$\begin{aligned}
B &= \sum_{g=1}^{s}\left\{p_{\lambda_n}^{(g)}\left(|\beta_{g1}^0 + \gamma_n u_{2,g1}|,\ldots,|\beta_{gp_g}^0 + \gamma_n u_{2,gs_g},0|\right) - p_{\lambda_n}^{(g)}\left(|\beta_{g1}^0|,\ldots,|\beta_{gp_g}^0,0|\right)\right\}\\
&= \sum_{g=1}^{s}\left\{\sum_{j=1}^{s_g}\frac{\partial p_{\lambda_n}^{(g)}\left(|\beta_{g1}^0|,\ldots,|\beta_{gp_g}^0|\right)}{\partial|\beta_{g_j}|}\mathrm{sgn}(\beta_{gj}^0)\gamma_n u_{2,gj}\right.\\
&\left.+\frac{1}{2}\sum_{i=1}^{s_g}\sum_{j=1}^{s_g}\frac{\partial^2 p_{\lambda_n}^{(g)}\left(|\beta_{g1}^0|,\ldots,|\beta_{gp_k}^0|\right)}{\partial|\beta_{gi}|\partial|\beta_{gj}|}\gamma_n^2 u_{2,gi}u_{2,gj}\right\} + o_p\left\{\gamma_n^2(u_{2,g1}^2+\cdots+u_{2,gs_g}^2)\right\}\\
&\leq q_1^{1/2}a_n\gamma_n\|u_2\| + \frac{1}{2}\gamma_n^2 b_n\|u_2\|^2 + o_p(\gamma_n^2\|u_2\|^2)\\
&= q_1^{1/2}O_p(\gamma_n)\gamma_n\|u_2\| + o_p(\gamma_n^2\|u_2\|^2)\ \text{ as }\quad b_n\to 0\\
&= q_1^{1/2}O_p(\gamma_n^2)\|u_2\| + o_p(\gamma_n^2\|u_2\|^2)\\
&\triangleq B_1 + B_2,
\end{aligned}$$

where $q_1 = \sum_{g=1}^{s} s_g$. We see that, by choosing a sufficiently large D, A_2 dominates A_1, A_n, B_1, B_2 uniformly in $\|u\| = D$. Thus, we have shown that $\|\hat{\alpha} - \alpha^0\| + \left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$. Then, $\left\|\hat{\phi} - \alpha^{0^\top}B\right\| = O_p(\gamma_n)$ and $\left\|\hat{\beta} - \beta^0\right\| = O_p(\gamma_n)$. By $\left\|\phi^0 - \alpha^{0^\top}B\right\|_\infty = O(K^{-d})$ and the triangle inequality, we have

$$\begin{aligned}
\left\|\hat{\phi} - \phi^0\right\| &\leq \left\|\hat{\phi} - \alpha^{0^\top}B\right\| + \left\|\alpha^{0^\top}B - \phi^0\right\|\\
&= O_p(\gamma_n) + O(K^{-d})\\
&= O_p(\gamma_n).
\end{aligned}$$

Hence, $\left\|\hat{\phi}-\phi^0\right\|+\left\|\hat{\beta}-\beta^0\right\|=O_p(\gamma_n)$. □

Proof of Theorem 2. Here we will prove the sparsity: $\mathrm{pr}(\hat{\beta}_{\mathcal{D}}=0)\to 1$ as $n\to\infty$. By Taylor expansion, we have

$$\begin{aligned}\frac{\partial Q_n(\hat{\alpha},\hat{\beta})}{\partial\beta_{gj}} &= n^{-1}\frac{\partial l_n(\hat{\alpha},\beta^0)}{\partial\beta_{gj}}+\sum_{g',j'}n^{-1}\frac{\partial^2 l_n(\alpha^0,\beta^*)}{\partial\beta_{g'j'}\partial\beta_{gj}}(\hat{\beta}_{g'j'}-\beta^0_{g'j'})\\ &\quad-\frac{\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|}\mathrm{sgn}(\hat{\beta}_{gj})\\ &= C_1+C_2+C_3,\end{aligned}\tag{4.20}$$

where β^* lies between $\hat{\beta}$ and β^0. Using the result from (4.19), we have $|C_1|=O_p(\gamma_n)$. By the convergence rate in Theorem 1 and $n^{-1}\sum_{g',j'}\partial^2 l_n(\hat{\alpha},\beta^*)/\partial\beta_{g'j'}\partial\beta_{gj}=O_p(1)$, $|\hat{\beta}_{g'j'}-\beta^0_{g'j'}|=O_p(\gamma_n)$. Thus, $|C_2|=O_p(\gamma_n)$. It follows from the definition of $\hat{\beta}_{gj}$ that, if $\hat{\beta}_{gj}\neq 0$,

$$\begin{aligned}\frac{\partial Q_n(\hat{\alpha},\hat{\beta})}{\partial\beta_{gj}} &= O_p(\gamma_n)+O_p(\gamma_n)-\frac{\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|}\mathrm{sgn}(\hat{\beta}_{gj})\\ &= \gamma_n\left\{O_p(1)-\gamma_n^{-1}\frac{\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|}\mathrm{sgn}(\hat{\beta}_{gj})\right\}.\end{aligned}\tag{4.21}$$

Next, we show that there is a contradiction in (4.21) if $pr\left\{\beta^0_{\mathcal{D}}=0\right\}$ does not tend to 1 when $n\to\infty$, then there exist $(g,j)\in\mathcal{D}$, such that $\hat{\beta}_{gj}\neq 0$. By the condition given in Theorem 2, that is, $\gamma_n^{-1}\partial p^{(g)}_{\lambda_n}\left(|\hat{\beta}_{g1}|,\ldots,|\hat{\beta}_{gp_g}|\right)/\partial|\beta_{gj}|\to\infty$ with probability tending to 1 as $n\to\infty$, for an arbitrary $\epsilon>0$, when n is large we have

$$\frac{\partial Q_n(\hat{\alpha},\hat{\beta})}{\partial\beta_{gj}}<0,\ 0<\hat{\beta}_{gj}<\epsilon,\quad \frac{\partial Q_n(\hat{\alpha},\hat{\beta})}{\partial\beta_{gj}}>0,\ -\epsilon<\hat{\beta}_{gj}<0.$$

This is in conflict with $\partial Q_n(\hat{\alpha},\hat{\beta})/\partial\beta_{gj}=0$ and results in a contradiction when $\hat{\beta}_{gj}\neq 0$. Therefore, $\mathrm{pr}(\hat{\beta}_{gj}=0)\to 1$ as $n\to\infty$. □

Proof of Corollary 1. We only need to check that the conditions in Theorem 1 hold for the penalty function $p^{(g)}_{\lambda_n}(|\beta_{(g)}|)=\lambda_n(|\beta_{g1}|+\cdots+|\beta_{gp_g}|)^{1/2},\ g=1,...,G$.

For $\beta_{gj} \in \mathcal{A}$, i.e., $\beta_{gj}^0 \neq 0$, we have,

$$\begin{aligned} a_n &= \max_{(g,j)\in\mathcal{A}} \frac{\delta p_{\lambda_n}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\delta|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{\delta\lambda_n(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)^{1/2}}{\delta|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{2}\lambda_n(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)^{-1/2} \\ &\leq \frac{1}{2}\lambda_n M^{-1/2} = O_p(\gamma_n), \end{aligned}$$

and

$$\begin{aligned} b_n &= \max_{(g,j)\in\mathcal{A}} \left|\frac{\delta^2 p_{\lambda_n}(|\beta_{g1}^0|, \ldots, |\beta_{gp_g}^0|)}{\delta|\beta_{gj}|^2}\right| \\ &= \max_{(g,j)\in\mathcal{A}} \left|\frac{\delta^2\lambda_n(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)^{1/2}}{\delta|\beta_{gj}|^2}\right| \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{4}\lambda_n(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)^{-3/2} \\ &\leq \frac{1}{4}\lambda_n M^{-3/2} \to 0, \end{aligned}$$

where $M = \min_g(|\beta_{g1}^0| + \cdots + |\beta_{gp_g}^0|)$. Therefore, the rate of convergence follows from Theorem 1.

For sparsity, suppose there exists $(g, j) \in \mathcal{C}$ for which $\hat{\beta}_{gj} \neq 0$. Since for all $(g, j) \in \mathcal{C}$, $\beta_{gj}^0 = 0;\ j = 1, \ldots, p_g$, we have

$$\begin{aligned} \gamma_n^{-1}\frac{\partial p_{\lambda_n}\left(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}|\right)}{\partial|\beta_{gj}|} &= \gamma_n^{-1}\frac{\delta\lambda_n(|\hat{\beta}_{g1}| + \cdots + |\hat{\beta}_{gp_g}|)^{1/2}}{\delta|\beta_{gj}|} \\ &= \frac{\gamma_n^{-1}\lambda_n}{2(|\hat{\beta}_{g1}| + \cdots + |\hat{\beta}_{gp_g}|)^{1/2}}. \end{aligned}$$

According to the first conclusion of Corollary 1, there exists a γ_n^{-1} consistent local maximizer $\hat{\beta} = (\hat{\beta}_{\mathcal{A}}^\top, \hat{\beta}_{\mathcal{B}}^\top, \hat{\beta}_{\mathcal{C}}^\top)^\top$ for the non-adaptive hierarchically penalized likelihood (4.9), which implies $\left\|\hat{\beta}_{\mathcal{C}} - \beta_{\mathcal{C}}^0\right\| \leq M^*\gamma_n$ or for $\hat{\beta}_{gj} \neq 0$, we have $|\hat{\beta}_{gj} - \beta_{gj}^0| = |\hat{\beta}_{gj}| \leq M^*\gamma_n$ for some constant M^*.

Thus,

$$\begin{aligned}\frac{\gamma_n^{-1}\lambda_n}{2(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}} &\geq \frac{\gamma_n^{-1}\lambda_n}{2(M^*\gamma_n+\cdots+M^*\gamma_n)^{1/2}} \\ &= \frac{1}{2M^{*1/2}} \times \frac{\gamma_n^{-1}\lambda_n\gamma_n^{-1/2}}{p_g^{1/2}} \\ &\geq \frac{\gamma_n^{-3/2}\lambda_n p^{-1/2}}{2M^{*1/2}} \quad (\text{since } p \geq p_g).\end{aligned}$$

Therefore, for $\gamma_n^{-3/2}\lambda_n p^{-1/2} \to \infty$ when $n \to \infty$, we have, $\gamma_n^{-1}\delta\lambda_n(|\hat{\beta}_{g1}|+\cdots+|\hat{\beta}_{gp_g}|)^{1/2}/\delta|\beta_{gj}| \to \infty$, which results in a contradiction when $\hat{\beta}_{gj} \neq 0$. So, for all $(g,j) \in \mathcal{C}$, $\hat{\beta}_{gj} = 0$. □

Proof of Theorem 3. We only need to check that the conditions in Theorem 1 hold for the penalty function $p_{\lambda_n}^{(g)}(|\beta_{(g)}|) = \lambda_n(w_{n,g1}|\beta_{g1}|+\cdots+w_{n,gp_g}|\beta_{gp_g}|)^{1/2}$.

For $\beta_{gj} \in \mathcal{A}$, i.e., $\beta_{gj}^0 \neq 0$, we have,

$$\begin{aligned}a_n &= \max_{(g,j)\in\mathcal{A}} \frac{\delta p_{\lambda_n}(|\beta_{g1}^0|,\ldots,|\beta_{gp_g}^0|)}{\delta|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{\delta\lambda_n(w_{n,g1}|\beta_{g1}^0|+\cdots+w_{n,gp_g}|\beta_{gp_g}^0|)^{1/2}}{\delta|\beta_{gj}|} \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{2}\lambda_n w_{n,gj}(w_{n,g1}|\beta_{g1}^0|+\cdots+w_{n,gp_g}|\beta_{gp_g}^0|)^{-1/2} \\ &\leq \frac{1}{2}\lambda_n w_{n,\max}^{\mathcal{A}}\left(w_{n,\min}^{\mathcal{A}}\right)^{-1/2} M^{-1/2} = O_p(\gamma_n),\end{aligned}$$

and

$$\begin{aligned}b_n &= \max_{(g,j)\in\mathcal{A}} \left|\frac{\delta^2 p_{\lambda_n}(|\beta_{g1}^0|,\ldots,|\beta_{gp_g}^0|)}{\delta|\beta_{gj}|^2}\right| \\ &= \max_{(g,j)\in\mathcal{A}} \left|\frac{\delta^2\lambda_n(w_{n,g1}|\beta_{g1}^0|+\cdots+w_{n,gp_g}|\beta_{gp_g}^0|)^{1/2}}{\delta|\beta_{gj}|^2}\right| \\ &= \max_{(g,j)\in\mathcal{A}} \frac{1}{4}\lambda_n(w_{n,gj})^2(w_{n,g1}|\beta_{g1}^0|+\cdots+w_{n,gp_g}|\beta_{gp_g}^0|)^{-3/2} \\ &\leq \frac{1}{4}\lambda_n\left(w_{n,\max}^{\mathcal{A}}\right)^2\left(w_{n,\min}^{\mathcal{A}}\right)^{-3/2} M^{-3/2} \to 0,\end{aligned}$$

where $M = \min_g(|\beta_{g1}^0|+\cdots+|\beta_{gp_g}^0|)$. Thus, the consistency follows from Theorem 1.

Next, we prove the sparsity. Assume $\hat{\beta}_{gj}$ is a local maximizer of $Q_n^w(\alpha,\beta)$ in (4.13) with $\left\|\hat{\beta}_{gj}-\beta_{gj}^0\right\| = O_p(\gamma_n)$. We can find a constant M^*, such that $|\hat{\beta}_{gj}| \leq M^*$ for all (g,j) with

probability tending to 1. Then for $(g, j) \in \mathcal{D}$, i.e., $\beta^0_{gj} = 0$, we have

$$\begin{aligned}\gamma_n^{-1} \frac{\partial p_{\lambda_n}\left(|\hat{\beta}_{g1}|, \ldots, |\hat{\beta}_{gp_g}|\right)}{\partial |\beta_{gj}|} &= \frac{\delta \lambda_n (w_{n,g1}|\hat{\beta}_{g1}| + \cdots + w_{n,gp_g}|\hat{\beta}_{gp_g}|)^{1/2}}{\delta |\beta_{gj}|} \\ &= \frac{\gamma_n^{-1} \lambda_n w_{n,gj}}{2(w_{n,g1}|\hat{\beta}_{g1}| + \cdots + w_{n,gp_g}|\hat{\beta}_{gp_g}|)^{1/2}} \\ &\geq \frac{\gamma_n^{-1} \lambda_n w^{\mathcal{D}}_{n,\min}}{2M^{*1/2}(w^{\mathcal{A}}_{n,\max} + w^{\mathcal{D}}_{n,\max})^{1/2}}.\end{aligned}$$

Therefore, when $\gamma_n^{-1} \lambda_n w^{\mathcal{D}}_{n,\min} / (w^{\mathcal{A}}_{n,\max} + w^{\mathcal{D}}_{n,\max})^{1/2} \to \infty$ as $n \to \infty$, then $\hat{\beta}_{gj} = 0$ with probability approaching to 1, and by Theorem 2, we have $\text{pr}(\hat{\beta}_{\mathcal{D}} = 0) \to 1$. □

Proof of Corollary 2. We only need to verify that $w_{n,gj} = |\tilde{\beta}_{n,gj}|^{-r}$ satisfy the conditions in Theorem 3. Let $A = \max_{g,j} \left\{\beta^0_{g_j}\right\}$ and $B = \min_{g,j} \left\{\beta^0_{g_j} : \beta^0_{g_j} \neq 0\right\}$. Then by the consistency of $\tilde{\beta}_n$, $w^{\mathcal{A}}_{n,\max} \to B^{-r}$ and $w^{\mathcal{A}}_{n,\min} \to A^{-r}$. Thus, if $\lambda_n = \gamma_n / \log(n)$, we have $\gamma_n^{-1} \lambda_n w^{\mathcal{A}}_{n,\max} \left(w^{\mathcal{A}}_{n,\min}\right)^{-1/2} \to 0$ and $\lambda_n \left(w^{\mathcal{A}}_{n,\max}\right)^2 \left(w^{\mathcal{A}}_{n,\min}\right)^{-3/2} \to 0$, as $n \to \infty$.

For each (g, j) with $\beta^0_{n,gj} = 0$, we have $\tilde{\beta}_{gj} = O_p(\gamma_n)$. Therefore, $w^{\mathcal{D}}_{n,\min} / (w^{\mathcal{A}}_{n,\max} + w^{\mathcal{D}}_{n,\max})^{1/2} = O_p(\gamma_n^{-1/2})$. Thus, for $\lambda_n = \gamma_n / \log(n)$, we have $\gamma_n^{-1} \lambda_n w^{\mathcal{D}}_{n,\min} / (w^{\mathcal{A}}_{n,\max} + w^{\mathcal{D}}_{n,\max})^{1/2} \to \infty$. □

Chapter 5

Discussion and Future Research

Partially linear models (PLMs) are important generalizations of linear models. Compared to linear models, the PLMs possess desirable flexibility of non-parametric regression models because they have both linear and non-linear components. Together with the flexibility of a nonparametric regression model and desirable asymptotic properties of the linear estimators with simple interpretability, PLMs are very useful models for analyzing high-dimensional data where variable selection plays an important role. Since grouping structures arise naturally in many statistical modeling problems, we have studied the bi-level selection and estimation in PL survival models using right censored data. We studied the partially linear Cox proportional hazards model (PL-PHM) and the partially linear additive hazards model (PL-AHM) as an alternative to the earlier model. For bi-level variable selection, we investigated the performance of the adaptive hierarchical penalty, which is a special case of the group bridge penalty. For comparison purpose, we also studied several existing penalties and compared the results with the adaptive hierarchical penalty.

In Chapter 2, we estimated the PL-AHM with right-censored and left-truncated data. We approximated the nonparametric components using computationally favorable B-splines. We extended the pseudoscore method (Lin and Ying, 1994) in our model for estimating coefficients. In Chapter 3 with a high-dimensional data, we performed group variable selection in the PL-AHM model using adaptive hierarchical penalty where the parameters can diverge with the sample size. The proposed method can select significant groups and important variables within selected groups simultaneously. In the same vein as of Chapter 3, in Chapter 4, we investigated the group variable selection in the PL-PHM with a high-dimensional data where parameters can be naturally grouped.

Theoretically, in Chapter 2, we established the asymptotic normality of the parameters under the assumption that the true nonlinear functions are B-spline functions whose knot locations and the number of knots are held fixed. In Chapter 3, we established the asymptotic convergence rate and selection consistency of the penalized estimators of the PL-AHM. In Chapter 4, when the dimension of the nonparametric functions are fixed and low, we established the asymptotic convergence rate and selection consistency of the PL-PHM. Numerically, we conducted extensive simulation studies to explore the performances of all the proposed models and methods, and demonstrated the comparability and superiority of our methods to the existing approaches. Four real data examples are provided to illustrate the use of the proposed methods.

5.1 Future Research

Our proposed methods can be extended in several directions for future research.

(I). In this thesis, we considered the partially linear structure of our models is already available. In literature, popular approaches of constructing a PLM are separating the covariates based on the shape of the estimated nonparametric functions from univariate analysis, or, putting discrete covariates in the linear part and continuous ones in the nonlinear part of the model. However, it is an interesting problem which is largely overlooked. Zhang et al. (2011) proposed a new approach to select structure in a PL model, and called it LAND (Linear And Nonlinear Discoverer). They showed that the LAND estimator is able to identify the underlying true model structure correctly, and simultaneously estimate the multivariate regression function consistently. This approach may be extended to our PL survival models for structure selection.

(II). For the purpose of variable selection, we considered only right censored data in our survival models. Lu and Song (2015) and Lu et al. (2016) estimated the PL-AHM and PL-PHM with current status data. Recently, Afzal et al. (2017) estimated the PL-AHM with

right-censored and left-truncated data. Our variable selection methods can be extended to such different censoring and truncated data.

(III). In our thesis, we focused on variable selection in the linear part of the models, where we estimated the fix and low dimensional nonparametric functions. Our method can be extended to select nonparametric functions simultaneously. The basis functions to approximate a nonparametric function can be viewed as a group and group variable selection can be used to select the important nonparametric functions.

(IV). In addition to the PL-AHM and PL-PHM models, our method can be extended to other survival models for variable selection, such as the accelerated failure time model, transformation model, etc. We can also perform component selection in nonparametric proportional hazards model and nonparametric additive hazard model, in a similar fashion to Cui et al. (2013) where they performed component selection in nonparametric additive regression model.

(V). In our thesis, we considered the covariates do not overlap and each variable belongs to only one group. However, in real data, some covariates can belong to several groups. For example, one gene can be shared by many different pathways. Wang et al. (2009) considered overlapping groups in group variable selection in the PHM. Our proposed group selection methods can be extended to problems with overlapping groups.

(VI). For variable selection, we considered time independent covariates. Our proposed method can be extended to incorporate time-dependent covariates.

(VII). We investigated the variable selection problem where the covariates can diverge with the sample size, however, we restricted our attention to $p < n$. It will be interesting to see how our proposed methods perform in an ultra-high dimensional setting ($p \gg n$).

(VIII). In a Bayesian framework, bi-level group variable selection has not been investigated in censored survival data. This could be our future research direction as well.

Bibliography

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer.

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925.

Afzal, A. R., Dong, C., and Lu, X. (2017). Estimation of partly linear additive hazards model with left-truncated and right-censored data. *Statistical Modelling*, 17(6):423–448.

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

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 62(1):43–72.

Aneiros, G., Ling, N., and Vieu, P. (2015). Error variance estimation in semi-functional partially linear regression models. *Journal of Nonparametric Statistics*, 27(3):316–330.

Bassendine, M., Collins, J., Stephenson, J., Saunders, P., and James, O. (1985). Platelet associated immunoglobulins in primary biliary cirrhosis: a cause of thrombocytopenia? *Gut*, 26(10):1074–1079.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):3092–3120.

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369–380.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Breslow, N. E. and Day, N. E. (1987). *Statistical methods in cancer research*, volume 2. International Agency for Research on Cancer Lyon.

Buckley, J. (1984). Additive and multiplicative models for relative survival rates. *Biometrics*, 40(1):51–62.

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC Press.

Cui, X., Peng, H., Wen, S., and Zhu, L. (2013). Component selection in the additive regression model. *Scandinavian Journal of Statistics*, 40(3):491–510.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag, New York.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, 10(1):1–7.

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.

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320.

Fan, J., Gijbels, I., and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25(4):1661–1690.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *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(1):74–99.

Fang, K., Wang, X., Zhang, S., Zhu, J., and Ma, S. (2015). Bi-level variable selection via adaptive sparse group lasso. *Journal of Statistical Computation and Simulation*, 85(13):2750–2760.

Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, 54(4):1475–1485.

Feng, Y., Ma, L., and Sun, J. (2015). Regression analysis of current status data under the additive hazards model with auxiliary covariates. *Scandinavian Journal of Statistics*, 42(1):118–136.

Fleming, T. R. and Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons, Hoboken, New Jersey.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gorst-Rasmussen, A. and Scheike, T. (2011). Coordinate descent methods for the penalized semiparametric additive hazards model. Technical report, Department of Mathematical Sciences, Aalborg University.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

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

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*, volume 41. Academic Press, New York.

Hall, P. and Müller, H.-G. (2003). Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *Journal of the American Statistical Association*, 98(463):598–608.

Hao, M., Song, X., and Sun, L. (2014). Reweighting estimators for the additive hazards model with missing covariates. *Canadian Journal of Statistics*, 42(2):285–307.

Härdle, W., Liang, H., and Gao, J. (2012). *Partially linear models*. Springer Science & Business Media.

Hu, J., Liu, F., and You, J. (2014). Panel data partially linear model with fixed effects, spatial autoregressive error components and unspecified intertemporal correlation. *Journal of Multivariate Analysis*, 130:64–89.

Hu, Y. and Lian, H. (2013). Variable selection in a partially linear proportional hazards model with a diverging dimensionality. *Statistics & Probability Letters*, 83(1):61–69.

Huang, C.-Y. and Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100(4):877–888.

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(4):481–499.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313.

Huang, J., Liu, L., Liu, Y., and Zhao, X. (2014). Group selection in the Cox model with a diverging number of covariates. *Statistica Sinica*, 24(4):1787–1810.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Huang, J. Z. and Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, 62(3):793–802.

Huffer, F. W. and McKeague, I. W. (1991). Weighted least squares estimation for aalen's additive risk model. *Journal of the American Statistical Association*, 86(413):114–129.

Jicai, L., Zhang, R., Zhao, W., and Lv, Y. (2016). Variable selection in partially linear hazard regression for multivariate failure time data. *Journal of Nonparametric Statistics*, 28(2):375–394.

Johnson, B. A. (2009). Rank-based estimation in the l1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*, 10(4):659–666.

Kai, B., Li, R., and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, 39(1):305–332.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kim, J. and Lee, S.-Y. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika*, 85(3):593–603.

Kim, J., Sohn, I., Jung, S.-H., Kim, S., and Park, C. (2012). Analysis of survival data with group lasso. *Communications in Statistics-Simulation and Computation*, 41(9):1593–1605.

Kim, K. H. (2016). Inference of the trend in a partially linear model with locally stationary regressors. *Econometric Reviews*, 35(7):1194–1220.

Kleinbaum, D. G. (1998). Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science & Business Media, New York City, New York.

Kubota, J., Ikeda, F., Terada, R., Kobashi, H., Fujioka, S.-i., Okamoto, R., Baba, S., Morimoto, Y., Ando, M., Makino, Y., Taniguchi, H., Yasunaka, T., Miyake, Y., Iwasaki, Y.,

and Yamamoto, K. (2009). Mortality rate of patients with asymptomatic primary biliary cirrhosis diagnosed at age 55 years or older is similar to that of the general population. *Journal of Gastroenterology*, 44(9):1000–1006.

Lai, T. L. and Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics*, 19(2):531–556.

Lee, E. T. and Wang, J. (2003). *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons.

Lee, K. H., Chakraborty, S., and Sun, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *The International Journal of Biostatistics*, 7(1):1–32.

Lee, K. H., Chakraborty, S., and Sun, J. (2015). Survival prediction and variable selection with simultaneous shrinkage and grouping priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(2):114–127.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284.

Leng, C. and Ma, S. (2007). Path consistent model selection in additive risk model via lasso. *Statistics in Medicine*, 26(20):3753–3770.

Li, J., Fine, J., and Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics*, 71(1):122–130.

Li, J., Wang, C., and Sun, J. (2012). Regression analysis of clustered interval-censored failure time data with the additive hazards model. *Journal of Nonparametric Statistics*, 24(4):1041–1050.

Li, W. and Xue, L. (2015). Efficient inference in a generalized partially linear model with random effect for longitudinal data. *Communications in Statistics-Theory and Methods*, 44(2):241–260.

Lian, H., Li, J., and Tang, X. (2014). Scad-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part. *Journal of Multivariate Analysis*, 125:50–64.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248.

Liang, H., Wang, S., Robins, J. M., and Carroll, R. J. (2011). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, 99(466):357–367.

Lin, D., Oakes, D., and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, 85(2):289–298.

Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71.

Lin, D. Y., Wei, L.-J., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264.

Liu, H., Yang, H., and Xia, X. (2017). Robust estimation and variable selection in censored partially linear additive models. *Journal of the Korean Statistical Society*, 46(1):88–103.

Liu, J., Zhang, R., and Zhao, W. (2014). Hierarchically penalized additive hazards model with diverging number of parameters. *Science China Mathematics*, 57(4):873–886.

Long, Q., Chung, M., Moreno, C. S., and Johnson, B. A. (2011). Risk prediction for prostate cancer recurrence through regularized estimation with simultaneous adjustment for nonlinear clinical effects. *The Annals of Applied Statistics*, 5(3):2003–2023.

Lu, X., Pordeli, P., Burke, M. D., and Song, P. X.-K. (2016). Partially linear single-index proportional hazards model with current status data. *Journal of Multivariate Analysis*, 151:14–36.

Lu, X. and Song, P. X.-K. (2012). On efficient estimation in additive hazards regression with current status data. *Computational Statistics & Data Analysis*, 56(6):2051–2058.

Lu, X. and Song, P. X.-K. (2015). Efficient estimation of the partly linear additive hazards model with current status data. *Scandinavian Journal of Statistics*, 42(1):306–328.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528.

Lv, J., Yang, H., and Guo, C. (2016). Variable selection in partially linear additive models for modal regression. `http://dx.doi.org/10.1080/03610918.2016.1171346`.

Ma, S. (2011). Additive risk model for current status data with a cured subgroup. *Annals of the Institute of Statistical Mathematics*, 63(1):117–134.

Ma, S. and Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statistica Sinica*, 22(3):1003–1020.

Ma, S. and Huang, J. (2007). Combining clinical and genomic covariates via Cov-TGDR. *Cancer informatics*, 3:371–378.

Ma, S. and Kosorok, M. R. (2005). Penalized log-likelihood estimation for partly linear transformation models with current status data. *The Annals of Statistics*, 33(5):2256–2290.

Ma, S., Song, X., and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8(60):1–17.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference*, 141(1):204–219.

Ma, Y.-b., You, J.-h., and Zhou, Y. (2013). Generalized profile LSE in varying-coefficient partially linear models with measurement errors. *Acta Mathematicae Applicatae Sinica, English Series*, 29(3):477–490.

Mallick, H. and Yi, N. (2017). Bayesian group bridge for bi-level variable selection. *Computational Statistics & Data Analysis*, 110:115–133.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, 89(3):649–658.

Martinussen, T. and Scheike, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4):602–619.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884.

Ni, X., Zhang, H. H., and Zhang, D. (2009). Automatic model selection for partially linear models. *Journal of Multivariate Analysis*, 100(9):2100–2111.

O'neill, T. J. (1986). Inconsistency of the misspecified proportional hazards model. *Statistics & probability letters*, 4(5):219–222.

O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *The Annals of Statistics*, 21(1):124–145.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Patel, A. V., Diver, W. R., Teras, L. R., Birmann, B. M., and Gapstur, S. M. (2013). Body mass index, height and risk of lymphoid neoplasms in a large united states cohort. *Leukemia & lymphoma*, 54(6):1221–1227.

Pierce, D. A. and Preston, D. L. (1984). Hazard function modelling for dose-response analysis of cancer incidence in the a-bomb survivor data. Technical Report 5, Oregon State University.

Pittman, J., Huang, E., Dressman, H., Horng, C.-F., Cheng, S. H., Tsou, M.-H., Chen, C.-M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., and West, M. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431–8436.

Pocock, S. J., Gore, S. M., and Kerr, G. R. (1982). Long term survival analysis: the curability of breast cancer. *Statistics in medicine*, 1(2):93–104.

Qiu, Z., Chen, X., and Zhou, Y. (2015). A kernel-assisted imputation estimating method for the additive hazards model with missing censoring indicator. *Statistics & Probability Letters*, 98:89–97.

Rajabi, B. and Sweetenham, J. W. (2015). Mantle cell lymphoma: observation to transplantation. *Therapeutic advances in hematology*, 6(1):37–48.

Ramsay, J. and Silverman, B. (2006). *Functional Data Analysis, 2nd Ed.* Springer, New York.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, 346(25):1937–1947.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B., Chiorazze, M., Giltnane, J., Hurt, E., Zhao, H., Averett, L., Henrickson, S., Yang, L., Powell, J., Wilson, W., Jaffe, E., Simon, R., Klausner, R., Montserrat, E., and Bosch, F. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2):185–197.

Rothman, K. J. (2012). *Epidemiology: an Introduction*, volume 8. Oxford University Press, New York.

Sankaran, P. and Anisha, P. (2012). Additive hazards models for gap time data with multiple causes. *Statistics & Probability Letters*, 82(7):1454–1462.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*, volume 14. Wiley, New York.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22(18):2262–2268.

Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, 97(457):210–221.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):413–436.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470.

Talwalkar, J. A. and Lindor, K. D. (2003). Primary biliary cirrhosis. *The Lancet*, 362(9377):53–61.

Tang, J. and Dickinson, J. (1998). Smoking and risk of myocardial infarction. studying relative risk is not enough. *BMJ (Clinical research ed.)*, 317(7164):1018.

Thomas, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. *Modern statistical methods in chronic disease epidemiology*, pages 197–210.

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

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

Tong, X., Hu, T., and Sun, J. (2012). Efficient estimation for additive hazards regression with bivariate current status data. *Science China Mathematics*, 55(4):763–774.

Tsiatis, A. A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93–108.

van der Vaart, A. and Wellner, J. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press, Cambridge, UK.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics*, 39(4):1827–1851.

Wang, P., Tong, X., Zhao, S., and Sun, J. (2015). Regression analysis of left-truncated and case i interval-censored data with the additive hazards model. *Communications in Statistics-Theory and Methods*, 44(8):1537–1551.

Wang, Q. and Sun, Z. (2007). Estimation in partially linear models with missing responses at random. *Journal of Multivariate Analysis*, 98(7):1470–1493.

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika*, 96(2):307–322.

Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369–1384.

Weiss, L., Melchardt, T., Egle, A., Hopfinger, G., Hackl, H., Greil, R., Barth, J., and Rummel, M. (2017). Influence of body mass index on survival in indolent and mantle cell lymphomas: analysis of the stil nhl1 trial. *Annals of Hematology*, 96(7):1155–1162.

Xia, X. and Yang, H. (2016). Variable selection for partially time-varying coefficient error-in-variables models. *Statistics*, 50(2):278–297.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Yan, Y. and Yi, G. Y. (2016). A class of functional methods for error-contaminated survival data under additive hazards models with replicate measurements. *Journal of the American Statistical Association*, 111(514):684–695.

Yang, J., Lu, F., and Yang, H. (2017). Quantile regression for robust estimation and variable selection in partially linear varying-coefficient models. `http://dx.doi.org/10.1080/02331888.2017.1314482`.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters*, 57(2):135–143.

Yin, G., Li, H., and Zeng, D. (2008). Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, 103(483):1200–1213.

You, J. and Chen, G. (2006). Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *Journal of Multivariate Analysis*, 97(2):324–341.

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.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zhang, C., Han, Y., and Jia, S. (2016). Accounting for time series errors in partially linear model with single-or multiple-run. *Journal of Computational and Graphical Statistics*, 25(1):123–143.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495):1099–1112.

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

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620.

Zhang, Z. (2016). Semiparametric estimation of partially linear transformation models under conditional quantile restriction. *Econometric Theory*, 32:458–497.

Zhao, P. and Xue, L. (2010). Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. *Journal of Multivariate Analysis*, 101(8):1872–1883.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhao, S., Hu, T., Ma, L., Wang, P., and Sun, J. (2015). Regression analysis of informative current status data with the additive hazards model. *Lifetime Data Analysis*, 21(2):241–258.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. `https://arxiv.org/abs/1006.2871`.

Zhou, Y. and Liang, H. (2009). Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates. *The Annals of Statistics*, 37(1):427–458.

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

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, 95(1):241–247.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.