STATISTICAL CONTRIBUTIONS TO ORDER RESTRICTED INFERENCE FOR SURVIVAL DATA ANALYSIS

Yunro Chung

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Jason P. Fine

Anastasia Ivanova

Michael G. Hudgens

Shyamal D. Peddada

David B. Richardson

ABSTRACT

Yunro Chung: Statistical Contributions to Order Restricted Inference for Survival Data Analysis
(Under the direction of Jason P. Fine and Anastasia Ivanova)

This dissertation aims to study order restricted inference for survival data analysis where a hazard function is assumed to have a shape restriction with respect to continuous covariates.

In the first chapter, we consider estimation of the semiparametric proportional hazards model with a completely unspecified baseline hazard function where the effect of a continuous covariate is assumed isotonic (or monotone) but otherwise unspecified. The pseudo iterative convex minorant algorithm is proposed to compute the isotonic estimator by optimizing a sequence of pseudo partial likelihood functions. A local consistency is established for a one-step update of the estimator when an initial value is in a shrinking neighborhood of the true value. Analysis of data from a recent HIV prevention study illustrates the practical utility of the methodology in estimating monotonic covariate effects that are nonlinear.

In the second chapter, we consider additive hazards model with a unimodal hazard function in a continuous covariate with unknown mode. A quadratic loss function is defined, which allows efficient computations to estimate the mode and unimodal covariate effects. The methodology is applied to analyze the data from a recent randomized clinical trial of cardiovascular disease in kidney transplant patients.

In the third chapter, we focus on multiple continuous covariates for a shape restricted hazard function. By assuming an additive isotonic structure of the multiple covariates under the proportional hazards model, the hazard function is defined as isotonic with respect to the partial order on the covariates. An efficient computation is proposed by combining the pseudo iterative convex minorant algorithm and the cycling algorithm. We use the proposed method to analyze the data from a recent clinical trial with cardiovascular outcome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Isotonic Hazard Function of a Univariate Continuous Covariate

Isotonic (or monotone) regression (Robertson et al. 1988) is a nonparametric method that can be used to explore an association between a covariate and an outcome variable. An efficient algorithm of the pool-adjacent-violators algorithm (Ayer et al. 1955) is available. The isotonic regression technique has been extended to survival data analysis, where the hazard function is assumed to have an isotonic restriction on covariates. Ancukiewicz et al. (2003) considered the situation where the hazard for an HIV infection increased when a continuous value of CD4 count decreased. They suggested the full-likelihood approach to estimate the isotonic hazard function in CD4 count, but their algorithm was ad hoc, e.g., might not even converge to a local maximum, and appeared computationally prohibitive in large samples. Alternatively, we suggest the isotonic proportional hazard model by incorporating an isotonic function to the semiparametric proportional hazard model (Cox 1972). It allows simple computation by avoiding the estimation of the baseline hazard in the partial likelihood. We further develop the pseudo iterative convex minorant algorithm for a large study, which is computationally stable and efficient than existing methods of iterative quadratic programming and iterative convex minorant algorithm (Groeneboom and Wellner 1992, pp. 69-73). A local consistency is established for a toy estimator, which is an one step estimator using the pseudo iterative convex minorant algorithm where the initial value is in a neighborhood of the true value.

1.2 Unimodal Hazard Function of a Univariate Continuous Covariate

We next focus on a unimodal function, where the hazard function is non-decreasing and non-increasing on $(-\infty, M]$ and $[M, +\infty)$, respectively. The point M is called a mode, which is generally unknown. We consider the unimodal regression approach (Shoung and Zhang 2001) that estimated unimodal functions at each hypothetical mode and estimated the mode to be the value at which the least square function had a minimum value. His profiling algorithm is directly applicable to the popular proportional hazard model, but there may be a computational challenge owing to the complicated structure of the partial likelihood. Alternatively, we consider estimation of the unimodal hazard function under the semiparametric additive hazard model (Lin and Ying 1994). It defines a quadratic loss function having a global Hessian matrix, which does not involve parameters. Thus, once the global Hessian matrix is computed, a standard quadratic programming method can be performed by profiling the mode.

1.3 Isotonic Hazard Function of Multiple Continuous Covariates

We consider a shape restricted hazard function in multiple continuous covariate. By assuming an additive isotonic structure of the multiple covariates in the semiparametric proportional hazard model, we separately added the multiple covariates to the hazard, assuming each covariate has an isotonic effect. Accordingly, the hazard function is defined as isotonic with respect to a partial order on the covariates. The additive isotonic structure have been well-studied for a standard regression setting (Bacchetti 1989). He suggested the cycling algorithm that optimized a univariate isotonic covariate effect with holding other isotonic covariate effects fixed by iterating the cycle. The simple structure of the least square function gave a closed form solution, which could be computed by the pool-adjacent-violators algorithm. In our model, however, the complicated structure of the partial likelihood does not allow the closed form solution, and additional computations are needed. An efficient computation is obtained by implementing the pseudo iterative convex minorant algorithm in conjunction with the cycling algorithm.

CHAPTER 2: LITERATURE REVIEW

2.1 Order Restricted Inference

In this Section, we review literature on order restricted statistical inference based on three types of the likelihood functions: full-likelihood function for bivariate shape restricted hazard functions in Subsection 2.1.1, separable likelihood function for monotone response models in Subsection 2.1.2, and non-separable likelihood function for the panel counting data and case 2 interval censored data 2.1.3.

2.1.1 Constrained Full-likelihood Approach

Ancukiewicz et al. (2003) proposed a full-likelihood approach to estimate hazard function under monotonicity. Their model is defined as

$$\lambda(t,z) = 1 - \{1 - \lambda_0(t)\}^{f(z)},$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, and $f(\cdot)$ is a monotone increasing function. They further assume that $\lambda_0(\cdot)$ has a range in $[0,1)$, and $f(\cdot)$ is non-negative. Let $t_1 < \cdots < t_s$ be the distinct observed failure times, and let $z_1 < \cdots < z_m$ be the distinct covariate values at any of those observed failure time. The full-likelihood function is then defined as

$$l(f,\lambda_0) = \sum_{i=1}^{m}\sum_{j=1}^{s}\big[d_{i,j}log\{1-(1-\lambda_0(t_j))^{f(z_i)}\} + (n_{i,j}-d_{i,j})f(z_i)log\{1-\lambda_0(t_j)\}\big], \quad (2.1)$$

where $d_{i,j}$ and $n_{i,j}$ are the number of patients at risk and the number failed at time t_j with covariate z_i, respectively. They propose an algorithm to maximize the full-likelihood by updating λ_0 given f and updating f given λ_0 iteratively until convergence. During the maximization steps, one additional constraint is imposed to have a unique factorization, where $\sum_{i=1}^{m} f(z)$ is set to the number of observation n. However, their proposed algorithm was ad hoc and did not have a global or even local convergence property.

2.1.2 Monotone Response Model with Constrained and Unconstrained Estimators

Banerjee (2007) suggested the monotone response model. Consider independent and identically distributed data $\{X_i, Z_i\}_{i=1}^{n}$, where $X_i|Z_i = z \sim p(x, \psi(z))$, p is a probability density, and $\psi(\cdot)$ is a monotone increasing (or monotone decreasing) function. One example is the monotone regression model, which is

$$X_i = \psi(Z_i) + \epsilon_i,$$

where ϵ_i is independent of Z_i with mean 0 and variance σ^2. Here, Z_i is a covariate value for the ith subject, and X_i is the response value. By assuming ϵ_i's are Gaussian, this model is expressed as a monotone response, where $X_i|Z_i = z \sim N(\psi(z), \sigma^2)$. Another example is the case 1 interval censored model. Let U_i and Z_i be an event and observation times for the ith subject, respectively, and $X_i = 1$ if $U_i \leq Z_i$, or $X_i = 0$ otherwise. Here, U_i is independent of Z_i. A main goal was to estimate F, a survival function of the U_i's. It is also expressed as a monotone response model, where $X_i|Z_i = z \sim Bernoulli(F(z))$.

Let $l(X_i, \psi(Z_i)) = -log\{p(X_i, \psi(Z_i))\}$. The negative log-likelihood function for the monotone response model is then defined as

$$l(x, \psi(z)) = \sum_{i=1}^{n} l(X_i, \psi(Z_i)) \tag{2.2}$$

Let $\psi_i = \psi(Z_{(i)})$, where $Z_{(i)}$ is the ith smallest value among Z_i's. Let $X_{(i)}$ be the response value corresponded to $Z_{(i)}$. Denote $\hat{\psi}$ as the minimizer over the monotone constraint that minimizes $\sum_{i=1}^{n} l(X_{(i)}, \psi_i)$ subject to $\psi_1 \leq \cdots \leq \psi_n$. Denote $\dot{l}$ and $\ddot{l}$ as first and second derivatives of the negative log-likelihood. It is shown that $\hat{\psi}$ is a minimizer over the monotone constraint if and only if

$$\sum_{j=1}^{n} \dot{l}(X_{(j)}, \hat{\psi}_j) = 0$$

and

$$\sum_{j=i}^{n} \dot{l}(X_{(j)}, \hat{\psi}_j) \geq 0 \text{ (i=1,...,n).}$$

By assuming its Hessian matrix is a diagonal and positive define matrix, the minimizer is characterized by

$$(\hat{\psi}_1, \ldots, \hat{\psi}_n) = slogcm\left[\sum_{j=1}^{i} \ddot{l}(X_{(j)}, \hat{\psi}_j), \sum_{j=1}^{i} \{\hat{\psi}_j \ddot{l}(X_{(j)}, \hat{\psi}_j) - \dot{l}(X_{(j)}, \hat{\psi}_j)\}\right]_{i=0}^{n},$$

where $\sum_{i=0}^{0} = 0$, and $slogcm[x_i, y_i]_{i=0}^{n}$ is the vector of slopes (or left derivatives) of the greatest convex minorant on cumulative sum diagram (x_i, y_i)'s.

He further suggested an constrained minimizer $\hat{\psi}^0$ by minimizing the negative log-likelihood function in (2.2) under the monotone constraint and the null hypothesis, $H_0 : \psi(z_0) = \theta_0$. Denote k as the number of Z_i's that are less than or equal to z_0. Since the negative log-likelihood function is separable in terms of $Z_{(i)}$'s, the minimization problem can be separated to two minimization problems: minimize $\sum_{i=1}^{k} l(X_{(i)}, \psi_i)$ subject to $\psi_1 \leq \cdots \leq \psi_k \leq \theta_0$, and minimize $\sum_{i=k+1}^{n} l(X_{(i)}, \psi_i)$ subject to $\theta_0 \leq \psi_{k+1} \leq \cdots \leq \psi_n$. Similar to the unconstrained minimizer, the constrained minimizer is characterized by

$$(\hat{\psi}_1, \ldots, \hat{\psi}_k) = slogcm\left[\sum_{j=1}^{i} \ddot{l}(X_{(j)}, \hat{\psi}_j), \sum_{j=1}^{i} \{\hat{\psi}_j \ddot{l}(X_{(j)}, \hat{\psi}_j) - \dot{l}(X_{(j)}, \hat{\psi}_j)\}\right]_{i=k}^{n} \wedge \theta_0$$

and

$$(\hat{\psi}_{k+1}, \dots, \hat{\psi}_n) = slogcm\left[\sum_{j=1}^{i} \ddot{l}(X_{(j)}, \hat{\psi}_j), \sum_{j=1}^{i}\{\hat{\psi}_j \ddot{l}(X_{(j)}, \hat{\psi}_j) - \dot{l}(X_{(j)}, \hat{\psi}_j)\}\right]_{i=k}^{n} \vee \theta_0.$$

Here, $\wedge$ and $\vee$ are the minimum and maximum operators, respectively. Based on the constrained and unconstrained minimizers, he developed likelihood ratio test.

2.1.3 Non-separable Likelihood Frameworks

The isotonic regression technique has been developed under independent and identically distributed data, where its likelihood function is separable. For example, the full-likelihood function in (2.1) in Subsection 2.1.2 is separable in terms of z_i given λ_0. The likelihood function in (2.2) is also separable in terms of Z_i in Subsection 2.1.2. The separable structure allows a relatively easy computation for an isotonic estimator. On the other hand, the following paragraphs describe two recent works for order restricted inference under non-separable likelihood functions, where the non-separation structure is from dependent data.

A first example is the panel count data (Wellner and Zhang 2000), where each subject is observed multiple time points with respect to the counts of events. Let $N = \{N(t) : t \geq 0\}$ be a counting process with mean function $E(N(t)) = \Lambda_0(t)$, K be an integer-valued random variable, and $T = \{T_{k,j}, j = 1 \dots, k, k = 1, 2, \dots\}$ is potential observation times. Here, N and (K, T) are independent, and $T_{k,j-1} \leq T_{k,j}$ for $j = 1, \dots, k$ and $k = 1, 2, \dots$. Denote $\{N_{K_i}^{(i)}, T_{K_i}^{(i)}, K_i\}_{i=1}^n$ as the independent and identically distributed copies of (N, T, K). By assuming N is a Poisson process that has the independent increment property, the log-likelihood function for Λ is defined as

$$l(\Lambda) = \sum_{i=1}^{n}\left[\sum_{j=1}^{K_i}(N_{K_i,j}^{(i)} - N_{K_i,j-1}^{(i)}) log\{\Lambda(T_{K_i,j}^{(i)}) - \Lambda(T_{K_i,j-1}^{(i)})\} - \Lambda(T_{K_i,K_i}^{(i)})\right],$$

which includes only two adjacent parameters at each j. Thus, it is reformulated as

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{K_i}\{\Delta N_{K_i,j}log(\Delta\Lambda_{K_i,j}) - \Delta\Lambda_{K_i,j}\}\right],$$

where $\Delta N_{K_i,j} = N_{K_i,j}^{(i)} - N_{K_i,j-1}^{(i)}$, and $\Delta\Lambda_{K_i,j} = \Lambda(T_{K_i,j}^{(i)}) - \Lambda(T_{K_i,j-1}^{(i)})$. The reformulated likelihood function was separable in terms of j, so that an isotonic regression method for independent data was used to make an inference for Λ.

A second example is the case 2 interval censored data. Let $\{X_i, T_i, U_i\}_{i=1}^{n}$ be independent sample from $\mathbb{R}_+^3$, where X_i is an event time with distribution function F_0, and T_i and U_i are observation times with a joint distribution H. Here, X_i and (T_i, U_i) are independent with $T_i \le U_i$. The log-likelihood for F is then defined as

$$l(F) = \sum_{i=1}^{n}\left[\delta_i logF(T_i) + \gamma_i log\{F(U_i) - F(T_i)\} + (1-\delta_i-\gamma_i)log\{1-F(U_i)\}\right], \qquad (2.3)$$

where $\delta_i = I(X_i \le T_i)$, and $\gamma_i = I(X_i \in (T_i, U_i])$, and where $I(\cdot)$ is the indicator function. The likelihood function is not separable in terms of T_i. However, it is partially separable in terms of left, right and interval censoring times. This partial separation plays a key role in showing consistency (Groeneboom and Wellner 1992) and asymptotic distributional result (Groeneboom 1996) for the isotonic estimator $\hat{F}$.

2.2 Computational Algorithms for Order Restricted Inference

In this section, we review computational algorithms for order restricted inference: iterative convex minorant algorithm for the case 2 interval censored model in Subsection 2.2.1; the profiling algorithm for the unimodal regression model in Subsection 2.2.2; the cycling algorithm for additive isotonic model in Subsection 2.2.3.

2.2.1 Iterative Convex Minorant Algorithm for the Case 2 Interval Censored Model

The iterative convex minorant algorithm is suggested to solve the case 2 interval censored data (Groeneboom 1996, pp. 69-73). The fundamental idea of the iterative convex minorant algorithm is that a convex optimization problem is reduce to a series of weight isotonic regression problems. The negative log-likelihood of (2.3) is represented as

$$l^n(F) = -\Big\{\sum_{j\in I_1} log\beta_j + \sum_{j\in I_{2a}} log(\beta_{k(j)} - \beta_j) + \sum_{j\in I_3} log(1-\beta_j)\Big\},$$

where $\beta_i = F(\nu_i)$ and $k(j) = \{k : \nu_k = \max(U_i, V_i), \nu_j = \min(U_i, V_i), \gamma_i = 1, i = 1, \ldots, n\}$, and where $I_1 = \{j : \nu_j = U_i \text{ with } \delta_i = 1\}$, $I_{2a} = \{j : \nu_j = \min(U_i, V_i) \text{ with } \gamma_i = 1\}$ and $I_3 = \{j : \nu_j = V_i \text{ with } \delta_i + \gamma_i = 0\}$ for $i = 1, \ldots, n$ and $j = 1 \ldots, l$. Here, $\nu_1 < \ldots < \nu_l$ is a sorted set of time points among U_i's with $\delta_i = 1$ or $\gamma_i = 1$ and V_i's with $\gamma_i = 1$ or $\delta_i + \gamma_i = 0$ and $\gamma_i = 0$ for $i = 1, \ldots, n$. In order to ensure $l(F) > \infty$, it is assumed that $1 \in I_1$ and $l \in I_3$. The goal is to find the maximizer of $l(F)$ over the convex cone $C = \{\beta \in \mathbb{R}^l : \beta_1 \leq \cdots \leq \beta_l\}$. Denote $\dot{l}^n(F)$ as the first derivative of $l^n(F)$. Then, the convex function $l^n(F)$ is approximated locally near $\beta^{(0)}$ by a quadratic function

$$l^n(F) \approx \frac{1}{2}\{\beta - \beta^{(0)} + W(\beta^{(0)})^{-1}\dot{l}^n(\beta^{(0)})\}^T W(\beta^{(0)})\{\beta - \beta^{(0)} + W(\beta^{(0)})^{-1}\dot{l}^n(\beta^{(0)})\},$$

where W is a Hessian matrix. By ignoring off-diagonal elements in W, the approximated quadratic function is reduced to

$$l^n(F) \approx \frac{1}{2}\sum_{i=1}^{n}\{\beta_i - \beta_i^{(0)} + w_i(\beta^{(0)})^{-1}\dot{l}_i^n(\beta^{(0)})\}^2 w_i(\beta^{(0)}),$$

where w_i is the ith diagonal element of W, $i = 1, \ldots, n$. This is an identical problem of estimating the isotonic function β over weight w. Thus, an initial value of $\beta^{(0)}$ is chosen in C, and then, the series of weight isotonic regression functions are solved by using either the greatest convex minorant or pool-adjacent-violators algorithm iteratively until convergence. The convergence criteria is Fenchel's duality conditions

$$\sum_{i=1}^{l} \hat{\beta}_i \dot{l}_i^n(\hat{\beta}_i) = 0$$

and

$$\sum_{i=1}^{l} \beta_i \dot{l}_i^n(\hat{\beta}_i) \geq 0.$$

for or all $(\beta_1, \ldots, \beta_l) \in C$. A distance stopping criteria may be alternatively used but it is a weaker condition than Fenchel's stopping criteria (Wellner and Zhan 1997). An advantage of the iterative convex minorant algorithm is computational speed, since the approximated likelihood function has simpler structure by ignoring the off-diagonal elements in W. At the time when the iterative convex minorant algorithm was suggested, convergence property was not proven. It was conjectured that Fenchel duality conditions did not depend on the Hessian matrix, and the Hessian matrix contained only few nonzero off-diagonal elements. Aragón and Eberly (1992) showed the local convergence under the (unrealistic) assumption where the jump points of the nonparametric maximum likelihood estimation are determined prior to applying the algorithm. Later, Jongbloed (1998) modified the iterative convex minorant algorithm to have a global convergence property by adding a line search algorithm.

2.2.2 Profiling Algorithm for Unimodal Regression

We consider the unimodal regression (Shoung and Zhang 2001) that minimizes

$$LS(f_0) = \sum_{i=1}^{n} \{Y_i - f_0(X_i)\}^2, \tag{2.4}$$

where (X_i, Y_i) are independent and identically distributed sample from (X, Y), $i = 1, \ldots, n$, and f_0 is an unknown unimodal function with an unknown mode m_0. To estimate m_0, they suggested nonparametric least squares estimator, which is

$$\hat{m_0} = X_{\hat{j}}, \quad \hat{j} = \underset{j=1,\ldots,n}{\arg\min}[\min_{f_0 \in F_j} LS(f_0)], \tag{2.5}$$

where $F_j = \{f_0 : f_0 \text{ is a unimodal function with mode } X_j\}$. Let $X_{(i)}$ be the ith largest value among $(X_1, \ldots, X_n)$, and $Y_{(i)}$ be the response value associated with $X_{(i)}$. Then they separated the minimization problem in (2.4) into two minimization problems:

$$\min \sum_{i=1}^{m} \{Y_{(i)} - f_0(X_{(i)})\}^2 \quad \text{subject to} \quad f_0(X_{(1)}) \leq \cdots \leq f_0(X_{(m)}) \tag{2.6}$$

$$\min \sum_{i=m+1}^{n} \{Y_{(i)} - f_0(X_{(i)})\}^2 \quad \text{subject to} \quad f_0(X_{(m+1)}) \geq \cdots \geq f_0(X_{(n)}). \tag{2.7}$$

The isotonic and anti-isotonic regression techniques can be separately performed on (2.6) and (2.7) with the pool-adjacent-violators algorithm. Let $\hat{f}_{0,j}$ be the estimated unimodal function at the mode of $X_{(j)}$. The profiling algorithm is to estimate the unimodal function $\hat{f}_{0,j}$ by profiling every hypothetical mode $X_{(j)}$, $j = 1, \ldots, n$, and estimate mode by (2.5).

2.2.3 Cycling Algorithm for Additive Isotonic Regression

Bacchetti (1989) extended the isotonic regression model to the additive isotonic model to include multiple covariates. Let Y^i and $X^i = (X_1^i, \ldots, X_d^i)$ be response scalar and covariate vector values for the ith subject, respectively, $i = 1, \ldots, n$, Let $X_j^{(i)}$ be the ith largest value among $(X_j^1, \ldots, X_j^n)$. The additive isotonic model minimizes the least square function

$$\sum_{i=1}^{n} \{Y^i - \mu_1(X_1^i) - \cdots - \mu_d(X_d^i)\}^2 \tag{2.8}$$

over the additive isotonic constraint where $\mu_j(X_j^{(1)}) \leq \ldots \leq \mu_j(X_j^{(n)})$ for $j = 1, \ldots, d$. They suggested the cycling algorithm that updated a univariate isotonic function μ_k while holding other isotonic functions $(\mu_1, \ldots, \mu_{k-1}, \mu_{k+1}, \ldots, \mu_n)$ constant. Correspondingly, the least square function in (2.8) is reduced to

$$\sum_{i=1}^{n} \{\tilde{Y}^i - \mu_k(X_k^i)\}^2, \tag{2.9}$$

where $\tilde{Y}^i = Y^i - \sum_{j=1, j \neq k}^{d} \mu_j(X_j^i)$. The reduced least square function in (2.9) has a closed form over the isotonic constraint μ_k, which can be computed by using the pool-adjacent-violators algorithm. By iterating the cycles $k = 1, \ldots, d, 1, \ldots, d, \ldots$, this algorithm is guaranteed to converge to the minimum value of the least square function in (2.8). On the other hands, it does not not guarantee the uniqueness of the isotonic minimizer. In other words, different isotonic estimators might yield the same minimum values of the least square function in (2.8).

CHAPTER 3: PARTIAL LIKELIHOOD ESTIMATION OF ISOTONIC PROPORTIONAL HAZARDS MODELS

3.1 Introduction

In regression analysis, common parametric models, for example, generalized linear models, may employ shape-restrictions on covariate effects, the simplest being that of monotonicity. There is extensive literature on nonparametric isotonic regression models, where the form of a monotone covariate effect is completely unspecified; see Banerjee (2007). Computational and inferential issues have been well studied, particularly for likelihood-based estimation of isotonic generalized linear models, where efficient algorithms are available which exploit the geometric properties of the shape-restricted likelihood and which facilitate a careful theoretical analysis of the large sample properties of the resulting estimators. Unfortunately, these approaches are not easily generalizable to partial likelihood estimation of the semiparametric isotonic proportional hazards model, owing to the lack of an independent and identically distributed structure of the partial likelihood. In survival data settings, constrained nonparametric maximum likelihood estimation was developed by Ancukiewicz et al. (2003) using ad hoc algorithms. Such algorithms may not even converge to a local maximum, and appear computationally prohibitive in large samples. The goal of this paper is theoretically justified computation of isotonic estimators based on partial likelihood in survival data settings.

The closest related work with right censored data is for nonparametric estimation of the hazard function subject to shape constraints in the absence of covariates. Various authors studied maximum likelihood estimation of a hazard function assumed to be monotone in time (Grenander 1956, Marshall and Proschan 1965, Rao 1970, Mukerjee and Wang 1993, Huang and Wellner 1995, Banerjee 2008, Lopuhaä and Nane 2013), including a 2013 Delft University of Technology PhD thesis by G. Nane, where the baseline hazard function is not assumed to be monotone in time. With categorical covariates having regression parameters known to satisfy a monotone ordering, one may post-process unrestricted partial likelihood estimates using the pool-adjacent-violators algorithm (Ayer et al. 1955) to obtain restricted estimators, similar to post-processing of likelihood estimators of parametric regression models with categorical covariates. This approach is not applicable with continuous covariates, owing to the fact that unrestricted estimation is not possible at all values of the covariate. Specialized methods are needed.

Suppose that T is a failure time, C is a censoring time and Z is a scalar continuous covariate, where it is assumed that T and C are independent conditionally on Z. Define $X = \min(T, C)$ and $\Delta = I(T \le C)$, where $I(\cdot)$ is the indicator function. The observed data consist of n independent and identically distributed replicates of (X, Δ, Z), denoted by $\{X_i, \Delta_i, Z_i\}$ $(i = 1, \dots, n)$. The proportional hazards model (Cox 1972) may be specified to incorporate monotone covariate effects, that is, $\lambda(t|Z) = \lambda_0(t)\exp\{\phi(Z)\}$, where $\lambda_0(t)$ is an unspecified baseline hazard function and $\phi(\cdot)$ is a monotone increasing function. In the usual Cox model, the form of $\phi(\cdot)$ is specified parametrically, for example, using low-order polynomials of Z. These parameters may then be estimated by maximizing the partial likelihood without imposing further restrictions on the parameters. When $\phi(\cdot)$ is monotone but otherwise unspecified, care is needed in defining the estimator using the partial likelihood, denoted by

$$pl(\phi) = \prod_{i=1}^{n} \prod_{t \ge 0} \left\{ \frac{e^{\phi(Z_i)}}{\sum_{j=1}^{n} Y_j(t) e^{\phi(Z_j)}} \right\}^{dN_i(t)},$$

where $N_i(t) = I(X_i \le t, \Delta_i = 1)$ is a counting process and $Y_i(t) = I(X_i \ge t)$ is an at-risk process for the ith subject for $i = 1, \dots, n$.

Unlike the usual likelihood based formulation for isotonic linear models (Robertson et al. 1988), the partial likelihood for the isotonic proportional hazards model is a product integral of terms depending on both time and covariate values, where the parameter $\phi(\cdot)$ only enters the partial likelihood at those covariate values in the dataset. To ensure that estimation is well-defined between those values, we restrict the estimator to be piecewise constant, which yields a unique estimator with potential jumps at the observed Z_i's. This assumption is similar to that made in isotonic generalized linear models. For right censored data, we show that the estimator jumps only at those covariate values which are associated with observed failure events with $\Delta_i = 1$; this is made precise in Subsections 3.2.1-3.2.2.

Calculating the constrained partial likelihood estimator is challenging and does not follow directly from earlier likelihood analyses of isotonic generalized linear models. The iterative quadratic programming method for $pl(\phi)$ is applicable to find the constrained estimator, but cannot be efficiently implemented using the pool-adjacent-violators algorithm, may be computationally prohibitive in large samples, and may exhibit poor convergence properties. The iterative convex minorant algorithm is also theoretically justified and has been shown to reduce the computational burden in many isotonic estimation problems, but exhibits similar difficulties in our setting. To overcome these issues, we propose the pseudo iterative convex minorant algorithm which finds the constrained partial likelihood estimator by iteratively minimizing a constrained pseudo partial likelihood. The convergence properties of pseudo iterative convex minorant algorithm can be established similar to those for iterative convex minorant algorithm.

3.2 Constrained partial likelihood estimation

3.2.1 Iterative quadratic programming and iterative convex minorant algorithm without censoring

Define the isotonic estimator of $\phi(\cdot)$ to be the maximizer of the partial likelihood under the monotone constraint that $\phi(Z_{(1)}) \leq \cdots \leq \phi(Z_{(n)})$, where $Z_{(i)}$ is the ith smallest value among $Z_1, \ldots, Z_n$. One must fix one point of the partial likelihood estimator, otherwise there is no unique maximizer because all ordered sets of $\{\phi(Z_{(1)}) + \delta, \ldots, \phi(Z_{(n)}) + \delta\}$ yield the same value of the partial likelihood for any δ. We impose an anchor constraint that $\phi(K) = \delta$ by prespecifying a constant K in the support of Z prior to the analysis of the data. Under the anchor constraint, the model fitted is

$$\lambda(t|Z) = \lambda_0(t)e^{\phi(Z)} = \{\lambda_0(t)e^{\delta}\}e^{\psi(Z)}, \tag{3.1}$$

where $\psi(Z) = \phi(Z) - \delta$ with $\psi(K) = 0$. Since the baseline hazard function absorbs $\exp(\delta)$, what we actually estimate is not $\phi(\cdot)$ but $\psi(\cdot)$. We regard δ as a nuisance parameter, with the only difference between $\psi(\cdot)$ and $\phi(\cdot)$ being the reference group defining the hazard ratio parameters. In other words, $\psi(\cdot)$ is vertically shifted from $\phi(\cdot)$ by δ, where hazard ratios based on $\psi(\cdot)$ and $\phi(\cdot)$ are identical, i.e., $\exp\{\phi(\cdot) - \phi(K)\} = \exp\{\psi(\cdot) - \psi(K)\}$. In practice, since $\psi(\cdot)$ is only estimable at the observed $Z_{(i)}$'s, we set $\psi(Z_{(k)}) = 0$, where $Z_{(k)}$ is the largest $Z_{(i)} \leq K$.

Let $l^N(\psi)$ denote the negative log partial likelihood,

$$l^N(\psi) = \textstyle\sum_{i=1}^n \displaystyle\int_0^\infty \big[-\psi_{(i)} + \log\big\{\textstyle\sum_{j=1}^n Y_{(j)}(u)e^{\psi_{(j)}}\big\}\big]dN_{(i)}(u),$$

where $\psi_{(i)} = \psi(Z_{(i)})$, and $N_{(i)}(u)$ and $Y_{(i)}(u)$ are counting and at-risk processes corresponding to the subject whose covariate is $Z_{(i)}$. In the sequel, as needed, we drop the subparentheses for notational convenience. The score function and Hessian matrix of the negative log partial likelihood are denoted as $U(\psi)$ and $H(\psi)$, respectively, with elements

$$\begin{aligned} u_s(\psi) &= \frac{\partial l^N(\psi)}{\partial \psi_s} = -\int_0^\infty dN_s(u) + \int_0^\infty E_s(\psi,u) d\bar{N}(u), \\ h_{ss}(\psi) &= \frac{\partial^2 l^N(\psi)}{\partial \psi_s^2} = \int_0^\infty \{E_s(\psi,u) - E_s(\psi,u)^2\} d\bar{N}(u), \\ h_{st}(\psi) &= \frac{\partial^2 l^N(\psi)}{\partial \psi_s \partial \psi_t} = -\int_0^\infty E_s(\psi,u) E_t(\psi,u) d\bar{N}(u), \end{aligned}$$

for $s,t = 1,\dots,n$ $(s \neq t)$, where $E_s(\psi,u) = Y_s(u)\exp(\psi_s)/\{\sum_{j=1}^n Y_j(u)\exp(\psi_j)\}$ and $d\bar{N}(u) = \sum_{i=1}^n dN_i(u)$.

Theorem 3.1. *Suppose that there is no censoring. The negative log partial likelihood* $l^N(\psi)$ *is convex. It is strictly convex when an anchor constraint is imposed that* $\psi_k = \psi(Z_{(k)}) = 0$.

Let Ψ^k be $\{\psi \in \mathbb{R}^n : \psi_1 \leq \cdots \leq \psi_n, \psi_k = 0\}$. The problem of maximizing the partial likelihood over the monotone and anchor constraints is equivalent to minimizing the strictly convex function $l^N(\psi)$ over the convex cone Ψ^k. We denote the minimizer of $l^N(\psi)$ over Ψ^k by $\hat{\psi} = (\hat{\psi}_1,\dots,\hat{\psi}_n)$, which we refer to as the isotonic partial likelihood estimator.

To uniquely estimate ψ at covariate values other than those in $Z_{(1)},\dots,Z_{(n)}$, we assume, similar to previous work on isotonic regression, that the estimator is a right-continuous step function with jumps at the order statistics of the Z_i's. Under this assumption, the strict convexity in Theorem 3.1 coupled with the following theorem give a unique characterization of the isotonic partial likelihood estimator:

Theorem 3.2. *Suppose that there is no censoring. The isotonic partial likelihood estimator* $\hat{\psi}$ *minimizes* $l^N(\psi)$ *over the convex cone* Ψ^k *if and only if Fenchel's duality condition* *holds that $\hat{\psi} \in \Psi^k$ satisfies*

$$\textstyle\sum_{j=1}^{i} u_j(\hat{\psi}) \leq 0 \quad (i = 1, \ldots, k-1), \;\; \sum_{j=i}^{n} u_j(\hat{\psi}) \geq 0 \quad (i = k+1, \ldots, n), \tag{3.2}$$

$$\textstyle\sum_{i=1, i \neq k}^{n} \hat{\psi}_i u_i(\hat{\psi}) = 0. \tag{3.3}$$

Moreover, $\hat{\psi}$ is uniquely determined by (3.2) and (3.3).

Iterative quadratic programming can be applied to find the isotonic partial likelihood estimator. It is designed to approximate a convex function by a quadratic function and find a solution by minimizing the quadratic function. A second order Taylor series approximation of $l^N(\psi)$ about ψ^0 is

$$\begin{aligned} l^N(\psi) &\approx l^N(\psi^0) + (\psi - \psi^0)U(\psi^0) + (\psi - \psi^0)H(\psi^0)(\psi - \psi^0)/2 \\ &= \frac{1}{2}\{\psi - \xi(\psi^0)\}H(\psi^0)\{\psi - \xi(\psi^0)\} + g(\psi^0), \end{aligned} \tag{3.4}$$

where $\xi(\psi^0) = \psi^0 - H(\psi^0)^{-1}U(\psi^0)$, and $g(\psi^0) = l^N(\psi^0) - U(\psi^0)H(\psi^0)^{-1}U(\psi^0)/2$ which does not depend on ψ. The procedure of the iterative quadratic programming method is that we set an initial value $\psi^{(0)} \in \Psi^k$, and update $\psi^{(m)} \in \Psi^k$ by minimizing the first term in (3.4), $\{\psi^{(m)} - \xi(\psi^{(m-1)})\}H(\psi^{(m-1)})\{\psi^{(m)} - \xi(\psi^{(m-1)})\}$, until convergence. The solution can be found by using a quadratic programming method with equality and inequality constraints. In the simulations reported in Section 3.4, we find that the procedure may be numerically unstable, with convergence dependent on the anchor constraint.

A challenge of the iterative quadratic programming method is to compute $H(\psi)^{-1}$, whose dimension is the same order as the sample size. This calculation may be computationally expensive or even infeasible. To simplify the computations, one may apply the iterative convex minorant algorithm (Groeneboom and Wellner 1992, pp. 69-73) that replaces $H(\psi)$ with $\text{diag}\{H(\psi)\}$ in the approximated partial likelihood in (3.4), where $\text{diag}\{H(\psi)\}$ is a diagonal matrix having the same diagonal elements as $H(\psi)$. Then the approximated partial likelihood in (3.4) reduces to

$$\frac{1}{2}\sum_{i=1}^{n}\left\{\psi_i - \xi_i(\psi^0)\right\}^2 h_{ii}(\psi^0) + g(\psi^0), \tag{3.5}$$

where $\xi_i(\psi^0) = \psi_i^0 - u_i(\psi^0)/h_{ii}(\psi^0)$. This is identical to finding a monotone increasing function that minimizes $\sum_{i=1}^{n}\{\psi_i - \xi_i(\psi^0)\}^2 h_{ii}(\psi^0)$ over the class of monotone increasing functions Ψ^k with weight h. One may use the pool-adjacent-violators algorithm to find the minimizer (Ayer et al. 1955). The procedure of iterative convex minorant algorithm is to set an initial value of $\psi^{(0)} \in \Psi^k$, and apply the pool-adjacent-violators algorithm to update $\psi^{(m)}$ until convergence. The convergence criteria is based on Fenchel's duality condition in Theorem 3.2. It characterizes isotonic estimator $\hat{\psi}$, and in practice, one will check this condition and application of the iterative convex minorant algorithm. To incorporate the anchor constraint, we impose a constraint on iterative convex minorant algorithm (Banerjee 2007), where at each mth step after applying the pool-adjacent-violators algorithm, we set $\psi_k^{(m)} = 0$; $\psi_i^{(m)} = 0$ if $\psi_i^{(m)} > 0$ for $i = 1, \ldots, k-1$; $\psi_i^{(m)} = 0$ if $\psi_i^{(m)} < 0$ for $i = k+1, \ldots, n$. The iterative convex minorant algorithm with the anchor constraint may be unstable, which strongly depends on the choice of the anchor point, as shown in the simulation studies in Section 3.4.

3.2.2 Pseudo iterative convex minorant algorithms with no censoring

While in theory the anchor constraint has no effect on estimation, in practice the convergence of both iterative quadratic programming method and iterative convex minorant algorithm may be impacted: different anchor constraints may yield different estimates. To address this issue, and to reduce the computational burden of the algorithms, we propose the pseudo iterative convex minorant algorithm via iteratively minimizing the constrained pseudo partial likelihood,

$$l^P(\psi|\nu) = \textstyle\sum_{s=1}^n \displaystyle\int_0^\infty \big\{-\psi_s dN_s(u) + E_s^P(\psi_s, u|\nu) d\bar{N}(u)\big\}, \tag{3.6}$$

where $E_s^P(\psi_s, u \mid \nu) = Y_s(u)e^{\psi_s}/\{\sum_{j=1}^n Y_j(u)e^{\nu_j}\}$ for constants $\nu_1, \dots, \nu_n$. The pseudo partial likelihood score function and Hessian matrix are defined as

$$\begin{aligned} u_s^P(\psi_s|\nu) &= -\int_0^\infty dN_s(u) + \int_0^\infty E_s^P(\psi_s, u|\nu) d\bar{N}(u), \\ h_{ss}^P(\psi_s|\nu) &= \int_0^\infty E_s^P(\psi_s, u|\nu) d\bar{N}(u) > 0, && (3.7) \\ h_{st}^P(\psi_s|\nu) &= 0,\ \ s, t = 1, \dots, n\ (s \neq t). && (3.8) \end{aligned}$$

The anchor constraint is not needed for $l^P(\psi|\nu)$ because it is a strictly convex function by (3.7) and (3.8). Let $\Psi = \{\psi \in \mathbb{R}^n : \psi_1 \leq \cdots \leq \psi_n\}$ be the convex cone obtained by removing the anchor constraint from Ψ^k. The procedure of pseudo iterative convex minorant algorithm is

Step 3.1: Set an initial value of $\dot{\psi}^{(0)} \in \Psi^k$ (or $\dot{\psi}^{(0)} \in \Psi$).

Step 3.2: Update $\dot{\psi}^{(m)}$ such that $\dot{\psi}^{(m)} = \arg\min_{\psi \in \Psi} l^P(\psi|\nu = \dot{\psi}^{(m-1)})$.

Step 3.3: Repeat Step 3.2 until convergence under the distance stopping criteria $d_e(\dot{\psi}^{(m)}, \dot{\psi}^{(m-1)}) < \dot{\epsilon}$ for small $\dot{\epsilon} > 0$, where $d_e(x, y) = \sum_{i=1}^n |\exp(x_i) - \exp(y_i)|$.

Step 3.4: Let $\ddot{\psi}_i = \dot{\psi}_i^{(m)} - \dot{\psi}_k^{(m)}$ $(i = 1, \dots, n)$ such that $\ddot{\psi} = (\ddot{\psi}_1, \dots, \ddot{\psi}_n) \in \Psi^k$.

Theorem 3.3. *The minimizer* $\dot{\psi}$ *in Step 3.2 minimizes* $l^P(\psi|\nu)$ *over the convex cone* Ψ *if and only if Fenchel's duality condition holds that*

$$\textstyle\sum_{j=i}^n u_j^P(\dot{\psi}_j|\nu) \geq 0 \quad (i = 1, \dots, n) \tag{3.9}$$

with equality holding if $i = 1$*, and*

$$\textstyle\sum_{i=1}^{n} \dot{\psi}_i u_i^P(\dot{\psi}_i|\nu) = 0. \qquad (3.10)$$

Moreover, $\dot{\psi}$ *is uniquely determined by (3.9) and (3.10).*

Theorem 3.4. *Suppose that* $\hat{\psi}^+$ *minimizes* $\sum_{i=1}^{n}(\psi_i^+ - w_i^{-1}\Delta_i)^2 w_i$ *over the class of isotonic functions in* Ψ*, where* $\Delta_i = \int_0^\infty dN_i(u)$ *and* $w_i = \int_0^\infty [\{Y_i(u)d\bar{N}(u)\}/\{\sum_{j=1}^{n} Y_j(u)\exp(\nu_j)\}]$*. Then,* $\dot{\psi} = \{\log(\hat{\psi}_1^+), \dots, \log(\hat{\psi}_n^+)\}$ *is the unique minimizer of* $l^P(\psi|\nu)$ *over* Ψ*.*

In Step 3.2, we are not guaranteed to eventually satisfy the convergence criteria in Step 3.3, i.e., it is possible to construct data sets and choose starting values such that the algorithm will not converge. However, in practice we have found this to be unlikely (see Section 3.4). Moreover, the following theorem indicates that if the algorithm does converge for any $\dot{\epsilon}$, then the estimate $\ddot{\psi}$ converges to the unique minimizer of the constrained partial likelihood in Theorems 3.1 and 3.2 as $\dot{\epsilon} \to 0$. This provides theoretical justification for the pseudo iterative convex minorant algorithm.

Theorem 3.5. *Suppose that for any* $\dot{\epsilon} > 0$*, there exists* $r(\dot{\epsilon})$ *such that the pseudo iterative convex minorant algorithm converges at* $r(\dot{\epsilon})$*th iteration under the distance stopping criteria* $d_e(\dot{\psi}^{(r(\dot{\epsilon}))}, \dot{\psi}^{(r(\dot{\epsilon})-1)}) < \dot{\epsilon}$*. Then, as* $\dot{\epsilon} \to 0$*,* $\ddot{\psi} = (\dot{\psi}_1^{(r(\dot{\epsilon}))} - \dot{\psi}_k^{(r(\dot{\epsilon}))}, \dots, \dot{\psi}_1^{(r(\dot{\epsilon}))} - \dot{\psi}_k^{(r(\dot{\epsilon}))})$ *converges to the unique minimizer of* $l^N(\psi)$ *over* Ψ^k*.*

3.2.3 Censoring

Suppose that some failure times are censored. The fact that censored subjects contribute limited information to the partial likelihood restricts the form of the isotonic partial likelihood estimator. As stated in Proposition 3.6, the isotonic partial likelihood estimator has jumps only at the covariate values associated with uncensored subjects. Thus we focus on uncensored subjects and estimate $\psi(\cdot)$ at covariate values associated with these uncensored subjects. One may view the form of this estimator in line with traditional survival analysis where the estimated survival function jumps only at the observed failure times (Kaplan and Meier 1958). To estimate $\psi(\cdot)$ computationally, we suggest replacing a parameter for a censored subject with the parameter for an uncensored subject having covariate value which is closest to that for the censored subject amongst all uncensored subjects having smaller covariate values than the censored subject.

Let $n^\star$ be the number of subjects with observed failure time of the total n subjects, and $Z_i^\star$ be their covariate values, $i = 1, \ldots, n^\star$. Define $n^\star$ disjoint intervals of $I_1^\star = (-\infty, Z_{(1)}^\star) \cup [Z_{(1)}^\star, Z_{(2)}^\star), I_2^\star = [Z_{(2)}^\star, Z_{(3)}^\star), \ldots, I_{n^\star}^\star = [Z_{(n^\star)}^\star, +\infty)$, where $Z_{(i)}^\star$ is the ith order statistic amongst the $Z_i^\star$'s. We can then construct the replacement parameters algorithm where $\psi(Z_h)$ is replaced with $\psi(Z_i^\star)$ if $Z_h \in I_i^\star$ for $h = 1, \ldots, n$; $i = 1, \ldots, n^\star$. Accordingly, at-risk processes for censored subjects are added to corresponding at-risk processes for observed subjects such that $Y_i^\star(t) = \sum_{h \in R_i} Y_h(t)$, where $R_i = \{h : Z_h \in I_i^\star, h = 1, \ldots, n\}$. The partial likelihood for censored data is then defined by

$$pl^C(\psi^\star) = \prod_{i=1}^{n^\star} \prod_{t \geq 0} \left\{ \frac{e^{\psi_i^\star}}{\sum_{j=1}^{n^\star} Y_j^\star(t) e^{\psi_j^\star}} \right\}^{dN_i^\star(t)},$$

where $\psi_i^\star = \psi(Z_{(i)}^\star)$ and $N_i^\star(t)$ is the counting process corresponding to $Z_{(i)}^\star$. We assume that $\psi_j = \psi_1^\star$ if $Z_j < Z_{(1)}^\star$ for $j = 1, \ldots, n$, otherwise ψ_j is not included in $pl^C(\psi)$. This enables estimation of $\psi(\cdot)$ at all values of Z including the left side of $Z_{(1)}^\star$. As stated in Proposition 3.6, the replacement parameters algorithm with $pl^C(\psi)$ is justified.

Proposition 3.6. *Assume that $\psi_j = \psi_1^\star$ if $Z_j < Z_{(1)}^\star$ for $j = 1, \ldots, n$. Then the isotonic partial likelihood estimator has jumps only at $Z_i^\star$'s, and thus, the unique maximizer of $pl^C(\psi)$ is also the unique maximizer of $pl(\psi)$.*

Since $pl^C(\psi)$ has the same form as $pl(\psi)$, Theorems 3.1 to 3.5 are all valid under censored data, so that iterative quadratic programming method, iterative convex minorant algorithm, and pseudo iterative convex minorant algorithm are applicable to find the unique maximizer of $pl^C(\psi)$.

3.2.4 Time-dependent covariate

Consider this model $\lambda(t|Z(t)) = \lambda_0(t)\exp[\psi\{Z(t)\}]$, where $Z(t)$ is a time-dependent covariate. It is assumed that the monotone increasing function $\psi(\cdot)$ does not change over time. Similar to censored data, the fact that the values of the time-dependent covariates prior to the first observed failure time do not contribute to the partial likelihood restricts the form of the isotonic partial likelihood estimator. As stated in Proposition 3.7, the isotonic partial likelihood estimator has jumps at the time-dependent covariate values associated with uncensored subjects only at their failure times. One may use replacement parameters algorithm where the parameters for subjects having their failure times observed are substituted for other parameters in the partial likelihood.

Formally, let $n^\star$ be the number of subjects with observed failure time of the total n subjects, and $Z_i^\star(t)$ be their covariates for $i = 1, \ldots, n^\star$. Let $Z_i^* = Z_i^\star(X_i^\star)$, the ith subject's covariate at time of failure. Define $n^\star$ disjoint intervals by $I_1^* = (-\infty, Z_{(1)}^*) \cup [Z_{(1)}^*, Z_{(2)}^*), I_2^* = [Z_{(2)}^*, Z_{(3)}^*), \ldots, I_{n^\star}^* = [Z_{(n^\star)}^*, +\infty)$, where $Z_{(i)}^*$ is the ith order statistic amongst the $Z_i^\star$'s. We can then construct the replacement parameters algorithm where $\psi(Z_h(X_j))$ is replaced with $\psi(Z_{(i)}^*)$ if $Z_h(X_j) \in I_i$ for $h, j = 1, \ldots, n$; $i = 1, \ldots, n^\star$. Accordingly, we express an at-risk process as $Y_i^\star(t) = \sum_{h \in R_i(t)} Y_h(t)$, where $R_i(t) = \{h : Z_h(t) \in I_i^\star, h = 1, \ldots, n\}$. The partial likelihood is then defined as

$$pl^D(\psi^\star) = \prod_{i=1}^{n^\star} \prod_{t \geq 0} \left\{ \frac{e^{\psi_i^*}}{\sum_{j=1}^{n^\star} Y_j^\star(t) e^{\psi_j^*}} \right\}^{dN_i^\star(t)},$$

where $\psi_i^* = \psi(Z_{(i)}^*)$ and $N_i^\star(t)$ is a process corresponding to $Z_{(i)}^*$. Since $Z_i^\star$'s are only defined for subjects with observed failure times, $pl^D(\psi)$ is applicable for both complete and censored data with the time-dependent covariate. As stated in Proposition 3.7, $pl^D(\psi)$ with the replacement parameters algorithm is justified.

Proposition 3.7. *Assume that* $\psi\{Z_i(X_j)\} = \psi^*_1$ *if* $Z_i(X_j) < Z^*_{(1)}$ *for* $i, j = 1, \ldots, n$. *Then the isotonic partial likelihood estimator has jumps only at* Z^*_i*, and thus, the unique maximizer of* $pl^D(\psi)$ *is also the unique maximizer of* $pl(\psi)$.

Unlike the censoring case with a time independent covariate where parameters for censored subjects are replaced, with time-dependent covariates, both censored and uncensored subjects may have parameters replaced. This may prevent some parameters from being estimated, when all parameters are replaced by the same parameter at an observed failure time. Nevertheless, one may still estimate $\psi(\cdot)$ by assuming that the isotonic estimator does not have jumps at a covariate value for the excluded parameters. Since $pl^D(\psi)$ has the same form as $pl(\psi)$, iterative quadratic programming method, iterative convex minorant algorithm, and pseudo iterative convex minorant algorithm are applicable to find the unique maximizer of $pl^D(\psi)$.

3.2.5 Local consistency of the pseudo partial likelihood estimator

We prove the local consistency of the pseudo partial likelihood estimator for a time independent covariate when an initial guess is sufficiently close to the true value, i.e, $\nu_{n,i} = \psi_0(Z_i) + \epsilon_{n,i}$ where $\psi_0(\cdot)$ is the true monotone increasing function, $\epsilon_{n,i}$ are small positive numbers converging to zero as n go to infinity and $\nu_{n,i}$, $i = 1, \ldots, n$, satisfy the monotonicty constraint for each n. Let $\mathbb{P}_n$ denote the empirical measure on $\{X_i, \Delta_i, Z_i\}_{i=1}^n$, and let P denote the true probability measure corresponding to the distribution of $\{X, \Delta, Z\}$. Denote $\dot{\psi}^{(1)}$ as the minimizer of n^{-1} times the pseudo partial likelihood in (3.6) with initial values $\nu_{n,i}$, $i = 1, \ldots, n$, i.e. $\dot{\psi}^{(1)} = \arg\min_{\psi \in \Psi} \mathbb{P}_n\{l_n(\psi(Z)\,|\,\underline{\psi}_0, \underline{\epsilon}_n)\}$, where

$$l_n(\psi(Z)\,|\,\underline{\psi}_0, \underline{\epsilon}_n) = \int_0^\tau \left[-\psi(Z) dN(t) + Y(t) e^{\psi(Z)} \frac{\mathbb{P}_n\{dN(t)\}}{n^{-1} \sum_{j=1}^n Y_j(t) e^{\psi_0(Z_j) + \epsilon_{n,j}}} \right],$$

which is strictly convex, where $\underline{\psi}_0 = \{\psi_0(Z_1), \dots, \psi_0(Z_n)\}$ and $\underline{\epsilon}_n = \{\epsilon_{n,i}, \dots, \epsilon_{n,n}\}$. Let

$$u_n(\psi(Z) \,|\, \underline{\psi}_0, \underline{\epsilon}_n) = \int_0^\tau \left[-dN(t) + Y(t) e^{\psi(Z)} \frac{\mathbb{P}_n\{dN(t)\}}{n^{-1} \sum_{j=1}^n Y_j(t) e^{\psi_0(Z_j) + \epsilon_{n,j}}} \right],$$
$$u_n(\psi(Z) \,|\, \underline{\psi}_0) = \int_0^\tau \left[-dN(t) + Y(t) e^{\psi(Z)} \frac{\mathbb{P}_n\{dN(t)\}}{\mathbb{P}_n\{Y(t) e^{\psi_0(Z)}\}} \right],$$
$$u(\psi(Z) \,|\, \psi_0(Z)) = \int_0^\tau \left[-dN(t) + Y(t) e^{\psi(Z)} \frac{P\{dN(t)\}}{P\{Y(t) e^{\psi_0(Z)}\}} \right],$$

where $u_n(\psi(Z) \,|\, \underline{\psi}_0, \underline{\epsilon}_n)$ is the first derivative of $l_n(\psi(Z) \,|\, \underline{\psi}_0, \underline{\epsilon}_n)$, $N(t) = I(X \le t, \Delta = 1)$ and $Y(t) = I(X \le t)$. Let $X_i, \Delta_i | Z_i = z \sim p(X, \Delta | \psi(z))$ and $Z_i \sim p_z$, $i = 1, \dots, n$, where p is the product of Lebesgue measure on $\mathbb{R}^+$ and counting measure on $\{0, 1\}$ and p_z is a Lebesgue density on I_z where I_z is the domain of Z. Assume $\{X_i, \Delta_i, Z_i\}_{i=1}^n$ are independent and identical distributed data. Let z_0 be an interior point of I_z. Let Θ denote a parameter space, which is an open subset of $\mathbb{R}$. Assume:

(A1) $P\{Y(t)\} > 0$ and $E\{N(t)\} < \infty$ for $t \in (0, \tau]$.

(A2) p_z is positive and continuous in a neighborhood of z_0.

(A3) $\psi(\cdot)$ is continuous and differentiable in a neighborhood of z_0 with $|\psi'(z_0)| > 0$.

(A4) Let $L = \inf\{\psi(z) : z \in I_z\}$ and $U = \sup\{\psi(z) : z \in I_z\}$. Then, $L, U \in \Theta$ with $-\infty < L < U < \infty$.

(A5) $E_{\theta_0}\{u(\theta_1 | \theta_0)^2\}$ is uniformly bounded in a compact rectangle containing $[L, U] \times [L, U]$ for $\theta_0, \theta_1 \in \Theta$.

(A6) For $\theta_0, \theta_1, \theta_2 \in \Theta$, $E_{\theta_0}\{u(\theta_1 | \theta_0)\} \ne E_{\theta_0}\{u(\theta_2 | \theta_0)\}$ whenever $\theta_1 \ne \theta_2$.

Assumptions (A1)–(A3) are standard in survival and isotonic regression models. By Assumption (A4), $\psi_0(\cdot)$ is bounded between L and U on I_z. Assumptions (A5) and (A6) are mild in that $u(\cdot)$ is a uniformly bounded and strictly increasing function of θ. The uniform strong consistency of the isotonic estimator is proven under Assumptions (A1)–(A6), with the result stated in the following theorem. It holds with censored data.

Theorem 3.8. *Let $\dot{\psi}^{(1)}$ be the first step estimate from the pseudo iterative convex minorant algorithm with $\nu_{n,i} = \psi_0(Z_i) + \epsilon_{n,i}$ for $i = 1, \dots, n$. Let z_0 be an interior point of I_z. Take $\epsilon_{n,i} = c_{n,i}/n$, $i = 1, \dots, n$, such that v_i satisfies a monotonicity constraint, where $l \le c_{n,i} \le u$ for fixed $-\infty < l \le u < \infty$. Then, there exists σ_1 and σ_2 where z_0 falls in the interior of $[\sigma_1, \sigma_2]$ such that $\sup_{z \in [\sigma_1, \sigma_2]} |\dot{\psi}^{(1)}(z) - \psi_0(z)| \to 0$ almost surely.*

This result states that the estimator based on a one-step update is consistent if the initial value is in an n^{-1} neighborhood of the true value. The proof, which is given in Section 3.7, follows from Lemma 2.1 in the detailed version of Banerjee (2007), which can be found on his webpage (M. Banerjee, University of Michigan) with one modification, namely establishing that

$$
\begin{aligned}
&\sup\left|\mathbb{P}_n[u_n\{\psi(Z)\,|\,\underline{\psi}_0, \underline{\epsilon}_n\}] - P[u\{\psi(Z)\,|\,\psi_0(Z)\}]\right| \\
&\le \sup\left|\mathbb{P}_n[u_n\{\psi(Z)\,|\,\underline{\psi}_0, \underline{\epsilon}_n\}] - \mathbb{P}_n[u_n\{\psi(Z)\,|\,\underline{\psi}_0\}]\right| \\
&+ \sup\left|\mathbb{P}_n[u_n\{\psi(Z)\,|\,\underline{\psi}_0\}] - P[u\{\psi(Z)\,|\,\psi_0(Z)\}]\right| \qquad (3.11)
\end{aligned}
$$

converges to zero almost surely over all bounded monotone increasing functions. We show that the first term after the $\le$ sign in (3.11) converges to zero as n goes to infinity if $\epsilon_{n,i}$ goes to zero sufficiently fast. We then show the convergence of the second term by using empirical process theory if $u_n\{\psi(Z)\}$ is a P-Glivenko-Cantelli function and $P[Y(t)\exp\{\psi_0(Z)\}]$ is bounded away from zero, which is similar to a situation which has been studied for counting process regression (Kosorok 2007, p.56). It suggests that if one starts the pseudo iterative convex minorant algorithm sufficiently close to the true parameters, then the resulting one step estimator is consistent. This local consistency is similar to what was shown in Chapters 5·1 and 5·2 of Groeneboom and Wellner (1992), where they defined their toy estimator as one step of the iterative convex minorant algorithm starting the iteration at the true parameters. Our toy estimator is valid under weaker conditions, in the sense that the initial value is not necessarily equal to the true value but rather in an n^{-1} neighborhood of the truth.

3.3 Extensions

3.3.1 Baseline hazard function

As discussed previously, the baseline hazard function $\lambda_0(t)$ and shift parameter δ are not identifiable, because $\{\lambda_0(t), \phi(Z)\}$ and $\{\lambda_0(t)\exp(\delta), \phi(Z) - \delta\}$ give the same model in (3.1). Ancukiewicz et al. (2003) deal with this problem by imposing a constraint $\sum_{i=1}^n \phi(Z_i) = n$, but this may be unstable and the interpretation of the parameter estimates is complicated, owing to the dependency on n. In Subsection 3.2.1, we imposed the anchor constraint of $\psi(Z) = \phi(Z) - \delta$ that allows to estimate $\lambda_0^\star(t)$, where $\lambda_0^\star(t) = \lambda_0(t)\exp(\delta)$ is a baseline hazard function including an anchor effect. In fact, it is the same approach of the standard proportional hazard model that defines a baseline hazard function at a reference group, Z_R. Let $\Lambda_0^\star(t) = \int_0^t \lambda_0^\star(t)$ be a cumulative baseline hazard function including an anchor effect. Then, the profile estimator of the cumulative baseline hazard function is available,

$$\hat{\Lambda}_0^\star(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n Y_j(u) e^{\hat{\psi}\{Z_j(u)\}}},$$

where $\hat{\psi}(\cdot)$ is the isotonic estimator from the partial likelihood.

3.3.2 Additional covariates

Suppose there are an additional p covariates in the model $\lambda(t|Z(t), W(t)) = \lambda_0(t)\exp[\psi\{Z(t)\} + \beta W(t)]$, where $W(\cdot)$ is a $p \times 1$ dimensional covariate process and β is a $p \times 1$ vector of regression parameters. The partial likelihood is then defined as

$$pl(\psi, \beta) = \prod_{i=1}^n \prod_{t\geq 0} \left[\frac{e^{\psi\{Z_i(t)\} + \beta W_i(t)}}{\sum_{j=1}^n Y_j(t) e^{\psi\{Z_j(t)\} + \beta W_j(t)}} \right]^{dN_i(t)}. \tag{3.12}$$

The partial likelihood can be maximized by the following procedure. We set initial values of $(\psi^{(0)}, \beta^{(0)}) \in \Psi^k \times \mathbb{R}^p$. We then update $\psi^{(m)}$ given $\beta = \beta^{(m-1)}$ using iterative quadratic programming method, iterative convex minorant algorithm, or pseudo iterative convex minorant algorithm, and update $\beta^{(m)}$ given $\psi = \psi^{(m)}$ using the Newton-Raphson algorithm, where

$$
\begin{aligned}
\beta^{(m)} =& \beta^{(m-1)} - H(\psi^{(m)}, \beta^{(m-1)})^{-1} U(\psi^{(m)}, \beta^{(m-1)}), \\
U(\psi, \beta) =& \textstyle\sum_{i=1}^{n} \displaystyle\int_0^\infty \left\{ W_i(t) - \frac{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)} W_j(t)}{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)}} \right\} dN_i(t), \\
H(\psi, \beta) =& \textstyle\sum_{i=1}^{n} \displaystyle\int_0^\infty \left[- \frac{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)} W_j(t)^{\otimes 2}}{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)}} \right. \\
& \left. + \frac{\left\{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)} W_j(t)\right\}^{\otimes 2}}{\left\{\sum_{j=1}^{n} Y_j(t) e^{\psi\{Z_j(t)\}+\beta W_j(t)}\right\}^2} \right] dN_i(t),
\end{aligned}
$$

where in general $W^{\otimes 2} = WW$. These two steps are iteratively repeated until convergence. The convergence criteria is $d(\psi^{(m)}, \psi^{(m-1)}) + d(\beta^{(m)}, \beta^{(m-1)}) < \epsilon$, where $d(\cdot, \cdot)$ is Euclidean distance and ϵ is a small positive number. The same statement in Proposition 3.7 can be made, so that the replacement parameters algorithm is justified for $pl(\psi, \beta)$. Thus, during the step to update ψ given $\beta = \hat{\beta}$, the iterative quadratic programming method, iterative convex minorant algorithm or pseudo iterative convex minorant algorithm are available to optimize the reduced partial likelihood,

$$
pl^D(\psi, \hat{\beta}) = \prod_{i=1}^{n^\star} \prod_{t \geq 0} \left\{ \frac{e^{\psi_i^* + \hat{\beta} W_i^*(t)}}{\sum_{j=1}^{n^\star} Y_j^\circ(t, \hat{\beta}) e^{\psi_j^*}} \right\}^{dN_i^*(t)},
$$

where $Y_i^\circ(t,\hat\beta) = \sum_{h \in R_i(t)} Y_h(t) \exp\{\hat\beta W_h(t)\}$, and $W_i^*(t)$ is the covariate vector process corresponding to $Z_{(i)}^*$.

3.4 Simulations

We conducted simulation studies to examine the performance of iterative quadratic programming method, iterative convex minorant algorithm, and pseudo iterative convex minorant algorithm. As a gold standard, we also evaluated the pseudo partial likelihood by setting ν to the true value in $l^P(\psi|\nu)$. For the first part of the simulation studies, we considered a time independent covariate Z that was generated from a uniform distribution on $(0,1)$. Three forms of monotone increasing functions on the interval (0,1) were considered: $\phi(Z) = Z$, $\phi(Z) = Z^{1/2}$ and $\phi(Z) = Z^2$. The failure time was then generated from a proportional hazards model with baseline hazard function being exponential with scale parameter $\alpha = 1$. The same scenarios were used for the second part of the simulation study with a time-dependent covariate $Z(t)$. The time-dependent covariate was piecewise constant. To construct $Z(t)$, we generated independent uniform (0,1) random variables on disjoint time intervals $(x_{j-1}, x_j]$, where $x_0 = 0, x_1 = 0{\cdot}22, x_2 = 0{\cdot}44, \dots, x_9 = 2, x_{10} = +\infty$. The censoring times were independently generated from a uniform distribution giving 30% censoring. We repeated the simulations 500 times with sample sizes 100, 500 and 1000. We set stopping values of ϵ and $\dot\epsilon$ to 10^{-3} and 10^{-5}, respectively. Two anchor points were considered, $K = 0{\cdot}5$ and $K = 0$. For each data set an initial value of $\psi_i^{(0)}$ for $i = 1, \dots, n$, was set to $|\hat\gamma|\bar Z_i$, where $\bar Z_i = Z_{(i)} - Z_{(k)}$, and $\hat\gamma$ was the estimated coefficient of $\bar Z_i$ from the standard Cox model.

The anchor effect was evaluated for iterative quadratic programming method and iterative convex minorant algorithm by comparing the two isotonic estimates. We define the percentage of matches as $MC = \sum_{r=1}^{500} MC_r/500$, where $MC_r = 1$ if $\max_{i \in \{1,\dots,n\}} |\hat\psi_r^1(Z_i) - \{\hat\psi_r^2(Z_i) - \hat\psi_r^2(0{\cdot}5)\}| < 0{\cdot}001$, and $MC_r = 0$ otherwise. Here, $\hat\psi_r^k(\cdot)$ is an estimated monotone increasing function for the rth data set for $r = 1, \ldots, 500$, where $k = 1$ for $K = 0{\cdot}5$ and $k = 2$ for $K = 0$. Note that for pseudo iterative convex minorant algorithm, by construction, the estimates are the same for all anchor constraints. To evaluate the performance of the different algorithms, we computed the integrated mean squared error $\int_0^1 E\{\psi^k(Z) - \hat{\psi}^k(Z)\}^2 dZ$ for $k = 1, 2$, where $\psi^k(Z) = \phi(Z) - \phi(K)$. Based on equally spaced grid points of z_g's between 0·001 and 0·999, the integrated mean squared error was approximated by $\sum_{r=1}^{R} \sum_{g=1}^{G} \{\psi^k(z_g) - \hat{\psi}_r^k(z_g)\}^2/(GR)$, with $G = 1000$ grid points and $R = 500$ simulation runs. We also computed the percentage of convergence based on Fenchel's duality condition in Theorem 3.2, $\max_{i \in \{1,\ldots,k-1\}} \sum_{j=i}^{n} u_j(\hat{\psi}) < \epsilon$, $\min_{i \in \{k+1,\ldots,n\}} \sum_{j=i}^{n} u_j(\hat{\psi}) > -\epsilon$ and $|\sum_{j=1, j \neq k}^{n} \hat{\psi}_j u_j(\hat{\psi})| < \epsilon$. To demonstrate Theorem 3.5, after pseudo iterative convex minorant algorithm converged under the distance stopping criteria, we additionally check Fenchel's duality condition to report the percentage of convergence.

Tables 3.1 and 3.2, reported in the Web Supplement, show simulations results for time independent and time-dependent covariates. Non-convergent cases are excluded for calculating the integrated mean squared error, the matched case percentage, and the computing time. The pseudo iterative convex minorant algorithm has good convergence results in agreement with Theorem 3.5. In addition, the pseudo iterative convex minorant algorithm dramatically improves the computational speed, especially with large sample sizes. As the iterative quadratic programming method needs to calculate the inverse of the full Hessian matrix, computational time increases cubically with the number of observed failure events. Both iterative quadratic programming method and iterative convex minorant algorithm fail to converge using Fenchel's duality condition in roughly 10% of datasets. The results for iterative convex minorant algorithm depends heavily on the anchor constraint. In particular, the iterative convex minorant algorithm is extremely slow when the anchor is set to 0. The results for pseudo iterative convex minorant algorithm, as well as for iterative quadratic programming method, do not depend on the anchor constraint. For small sample sizes, iterative quadratic programming method and iterative convex minorant algorithm have larger integrated mean squared errors than pseudo iterative convex minorant algorithm, but that the differences vanish as the sample size increases. As expected, the pseudo partial likelihood with known ν has the smallest integrated mean squared error.

3.5 HIV data

The Breastfeeding, Antiretroviral and Nutrition study was a randomized trial conducted between April 21, 2004 and Jan 28, 2010 in Liongwe, Malawi (Jamieson et al. 2012). 2369 pairs of HIV-infected breastfeeding mothers and their uninfected infants were randomized to one of the three groups: a maternal antiretroviral regimen ($n = 849$), daily infant nevirapine ($n = 852$), or standard of care as control ($n = 668$). A primary endpoint of the trial was HIV transmission to the infant. Infants were scheduled to be tested for HIV every few weeks up to 48 weeks after the birth. By 48 weeks there were 76, 62, and 74 infants observed to be HIV infected in the maternal antiretroviral, nevirapine, and control arms, respectively. The Breastfeeding, Antiretroviral and Nutrition study measured mothers' CD4 count (cells per mm^3) at the baseline, which has been shown to be an important predictor of mother to child transmission. Lower CD4 counts are indicative of a weakened immune system and typically are associated with higher levels of virus in HIV infected individuals. Therefore it is reasonable to assume the hazard of transmission of HIV from mother-to-infant decreases monotonically as a function of CD4 count. A standard Cox model with CD4 included in the linear predictor showed a decreasing hazard in the CD4 count (estimated hazard ratio=0·864 for a 100 unit increase in CD4; $P < 0{\cdot}01$), adjusted for the group effect (estimated hazard ratio=0·769 for antiretroviral versus control; estimated hazard ratio=0·620 for nevirapine versus control).

Table 3.1: Simulation results for time independent covariates: IMSE multiplied by 10^3 (median CPU time in seconds), convergence percentage and matched case percentage. The first and second lines are for anchor points of $K = 0{\cdot}5$ and $K = 0$ respectively.

Type	$\phi(Z)$	n	IQM			ICM			PICM			PPL
			IMSE	Conv	MC	IMSE	Conv	MC	IMSE	Conv	MC	IMSE
Comp	Z	100	281(2)	88%	100%	1359(1)	91%	75%	156(0)	95%	100%	81(0)
			281(1)	88%		118(3)	86%		156(0)	95%		81(0)
		500	33(134)	89%	100%	79(15)	89%	81%	32(0)	99%	100%	20(0)
			33(43)	89%		28(140)	87%		32(0)	99%		20(0)
		1000	17(827)	89%	100%	26(65)	90%	87%	18(2)	99%	100%	17(1)
			17(505)	89%		17(1279)	87%		18(1)	99%		17(1)
	$Z^{1/2}$	100	367(1)	89%	100%	1803(0)	91%	82%	133(0)	93%	100%	80(0)
			367(1)	89%		104(3)	84%		133(0)	93%		80(0)
		500	35(145)	89%	100%	138(11)	89%	90%	28(0)	98%	100%	19(0)
			35(44)	89%		25(105)	87%		28(0)	98%		19(0)
		1000	16(1078)	90%	100%	39(54)	90%	95%	16(1)	99%	100%	14(1)
			16(684)	90%		15(1032)	89%		16(1)	99%		14(1)
	Z^2	100	205(1)	87%	100%	632(0)	91%	61%	166(0)	97%	100%	64(0)
			205(1)	87%		117(3)	87%		166(0)	97%		64(0)
		500	32(145)	88%	100%	66(14)	89%	63%	32(0)	99%	100%	19(0)
			32(45)	88%		29(315)	84%		32(0)	99%		19(0)
		1000	16(1190)	88%	100%	23(61)	89%	66%	18(1)	100%	100%	17(1)
			16(573)	88%		17(1932)	78%		18(1)	100%		17(1)
Cens	Z	100	316(1)	89%	100%	841(0)	91%	75%	184(0)	99%	100%	40(0)
			316(0)	89%		145(1)	90%		184(0)	99%		40(0)
		500	39(48)	89%	100%	39(8)	89%	83%	42(0)	100%	100%	31(0)
			39(18)	89%		39(118)	88%		42(0)	100%		31(0)
		1000	22(290)	90%	100%	23(39)	91%	87%	23(1)	100%	100%	10(1)
			22(172)	90%		23(561)	89%		23(1)	100%		10(1)
	$Z^{1/2}$	100	268(1)	89%	100%	722(0)	91%	82%	161(0)	98%	100%	40(0)
			316(0)	89%		133(2)	90%		161(0)	98%		40(0)
		500	36(39)	89%	100%	38(6)	90%	92%	38(0)	100%	100%	29(0)
			36(19)	89%		35(74)	89%		38(0)	100%		29(0)
		1000	20(328)	90%	100%	21(31)	92%	94%	21(1)	100%	100%	10(1)
			20(238)	90%		21(727)	92%		21(1)	100%		10(1)
	Z^2	100	280(1)	88%	100%	650(0)	91%	63%	194(0)	99%	100%	53(0)
			280(0)	88%		145(2)	89%		194(0)	99%		53(0)
		500	38(39)	88%	100%	38(6)	89%	66%	41(0)	100%	100%	34(0)
			38(20)	88%		39(118)	87%		41(0)	100%		34(0)
		1000	21(302)	88%	100%	22(32)	90%	69%	23(1)	100%	100%	10(1)
			21(237)	88%		22(905)	85%		23(1)	100%		10(1)

IQM: iterative quadratic programming; ICM: iterative convex minorant algorithm; PICM: pseudo iterative convex minorant algorithm; PPL: pseudo partial likelihood; Comp: complete case; Cens: censoring case (about 30%); IMSE: integrated mean squared error; Conv: convergence percentage; MC: matched case percentage.

Table 3.2: Simulation results for time-dependent covariates: IMSE multiplied by 10^3 (median CPU time in seconds), convergence percentage and matched case percentage. The first and second lines are for anchor points of $K = 0{\cdot}5$ and $K = 0$ respectively.

Type	$\phi\{Z(t)\}$	n	IQM			ICM			PICM			PPL
			IMSE	Conv	MC	IMSE	Conv	MC	IMSE	Conv	MC	IMSE
Comp	$Z(t)$	100	223(2)	83%	100%	860(1)	84%	80%	127(0)	99%	100%	67(0)
			223(1)	83%		127(3)	82%		127(0)	99%		67(0)
		500	38(169)	88%	100%	87(24)	89%	87%	32(5)	99%	100%	35(4)
			38(71)	88%		32(257)	88%		32(5)	99%		35(4)
		1000	21(998)	92%	100%	40(124)	92%	89%	19(31)	100%	100%	19(31)
			21(800)	92%		19(1849)	90%		19(31)	100%		19(30)
	$Z(t)^{1/2}$	100	289(2)	74%	100%	838(0)	77%	85%	135(0)	98%	100%	38(0)
			289(1)	74%		98(3)	75%		135(0)	98%		38(0)
		500	38(179)	90%	100%	104(20)	90%	91%	31(5)	99%	100%	31(4)
			38(61)	90%		30(279)	90%		31(4)	99%		31(4)
		1000	18(928)	93%	100%	34(89)	93%	95%	17(31)	100%	100%	9(30)
			18(917)	93%		17(1551)	93%		17(30)	100%		9(30)
	$Z(t)^2$	100	207(2)	89%	100%	820(1)	86%	65%	117(0)	98%	100%	41(0)
			207(1)	89%		103(4)	85%		117(0)	98%		41(0)
		500	34(164)	89%	100%	60(24)	89%	69%	32(5)	99%	100%	22(5)
			34(61)	89%		31(260)	87%		32(5)	99%		22(5)
		1000	18(1110)	88%	100%	18(134)	89%	71%	18(31)	100%	100%	9(31)
			18(732)	88%		18(1782)	87%		18(31)	100%		9(31)
Cens	$Z(t)$	100	283(1)	84%	100%	266(0)	80%	75%	207(0)	100%	100%	159(0)
			283(1)	84%		159(2)	79%		207(0)	100%		159(0)
		500	48(59)	88%	100%	81(13)	89%	83%	43(5)	100%	100%	46(4)
			48(25)	88%		41(137)	89%		43(4)	100%		46(4)
		1000	26(360)	87%	100%	26(64)	86%	89%	27(31)	100%	100%	38(30)
			26(180)	87%		26(792)	86%		27(31)	100%		38(30)
	$Z(t)^{1/2}$	100	189(1)	84%	100%	442(0)	84%	83%	151(0)	99%	100%	126(0)
			189(1)	84%		131(2)	83%		151(0)	99%		126(0)
		500	40(57)	87%	100%	40(12)	87%	89%	41(4)	100%	100%	51(4)
			40(30)	87%		40(128)	87%		41(4)	100%		51(4)
		1000	24(368)	87%	100%	24(65)	87%	91%	25(30)	100%	100%	32(30)
			24(243)	87%		24(884)	87%		25(30)	100%		32(30)
	$Z(t)^2$	100	238(1)	84%	100%	342(0)	82%	68%	165(0)	99%	100%	104(0)
			238(1)	84%		130(2)	81%		165(0)	99%		104(0)
		500	39(58)	88%	100%	39(12)	89%	67%	39(5)	100%	100%	20(4)
			39(26)	88%		39(159)	87%		39(5)	100%		20(4)
		1000	22(360)	87%	100%	22(72)	88%	82%	23(31)	100%	100%	64(30)
			22(275)	87%		23(816)	85%		23(31)	100%		64(30)

IQM: iterative quadratic programming; ICM: iterative convex minorant algorithm; PICM: pseudo iterative convex minorant algorithm; PPL: pseudo partial likelihood; Comp: complete case; Cens: censoring case (about 30%); IMSE: integrated mean squared error; Conv: convergence percentage; MC: matched case case percentage.

The extent to which the effect of CD4 is adequately captured by a simple proportional hazards model is unclear.

We fit the isotonic model assuming only a monotone relationship between the hazard and CD4 count. The model was fit using the pseudo iterative convex minorant algorithm in Subsection 3.3.2 with anchor constraint $K = 300$. Choosing $K = 200$, $K = 500$, or $K = 1000$ yielded the same results, except that the iterative convex minorant algorithm did not converge when $K = 200$. Standard Cox models were also fit using polynomials of order one, two, and three for CD4. Treatment group indicators were included in all models using a linear term as in (3.12).

Figure 3.1 displays the hazard ratios based on the estimated isotonic and polynomial functions. The isotonic partial likelihood estimator does not have jumps between the CD4 counts of 1100 and 2000 because the corresponding 23 infants are all censored. Historically, individuals with HIV have been started on antiretroviral when their CD4 dipped below some cut-off between 200 and 500 cells per mm^3 because this range would correspond to higher viral load and greater infectiousness. The isotonic estimator provides a clear picture of this phenomenon, with a rapid decrease in risk occurring up to CD4 count 500 followed by a gradual levelling for larger counts. After adjusting for the monotone effect on CD4 count, the estimated hazard ratio is 0·762 for antiretroviral versus control, and 0·621 for nevirapine versus control, similar to the standard proportional hazards analysis.

The polynomial models do not provide particularly good fit compared to the isotonic estimator over this range except the cubic model. This is further supported by goodness-of-fit statistics calculated by stratifying individuals by CD4 quantiles (Parzen and Lipsitz 1999), where the goodness-of-fit statistics has an asymptotic chi-square distribution with three degrees of freedom. The cubic polynomial and isotonic models have smaller goodness-of-fit statistics (cubic, 0·1; isotonic, 0·2) than the simpler models do

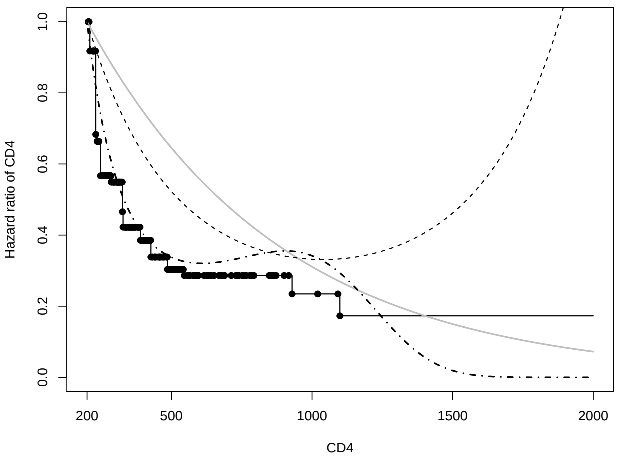

Figure 3.1: The Breastfeeding, Antiretroviral and Nutrition study. Estimated hazard ratios based on the isotonic partial likelihood estimator (black solid) and standard proportional hazards models with polynomials of order one (grey solid), two (dashed) and three (dot-dashed). The reference group is CD4 count equal to 200. The circles indicate HIV infections.

(linear, 4·5; quadratic, 2·3), where smaller goodness-of-fit statistics indicates a better model. However, the cubic term is not significant ($P = 0{\cdot}11$), and the estimated hazard ratio inexplicably increases between 500 and 1000, e.g., an infant whose mother has CD4 count of 900 has 1·108 times higher risk of HIV infection than an infant whose mother has CD4 count of 600. This demonstrates that simple parametric models may not adequately capture the nonlinear effect of CD4 count on mother to child transmission of HIV. An alternative approach to using low-dimensional polynomials could entail higher-dimensional parametric models, e.g., using splines. However, such an approach would require adding a monotonicity constraint to preclude results similar to those of the cubic polynomial model here.

3.6 Discussion

It remains to formally establish the global consistency and asymptotic distribution of the estimator. It is somewhat unclear how to apply earlier theoretical developments for

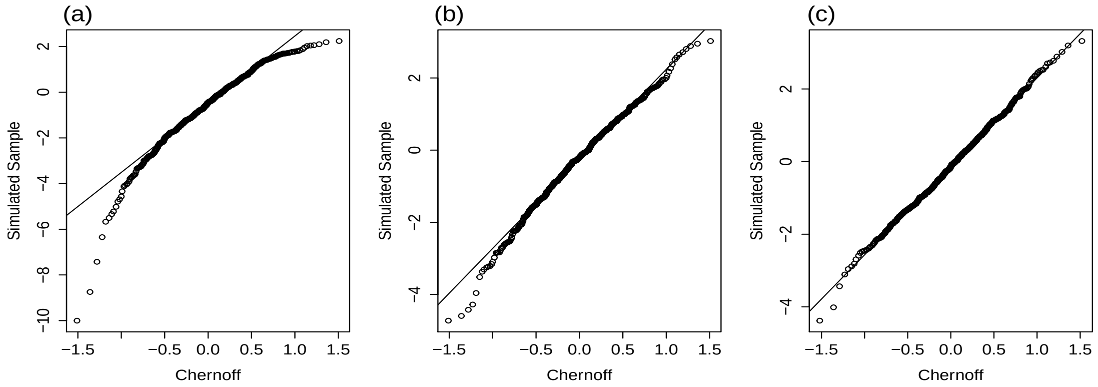

Figure 3.2: Quantile-quantile plot of simulated sample versus Chernoff's distribution at $z_0 = 0{\cdot}1$ with no censoring and $\psi(z) = z$. (a) $n = 100$; (b) $n = 500$; (c) $n = 1000$.

likelihood based analyses of isotonic regression models, as the log partial likelihood is not a sum of independent terms, because the partial likelihood has a non-separable structure at each failure time in terms of $T_{(i)}$'s and $Z_{(i)}$'s. This problem is made more challenging by the fact that each term in the partial likelihood involves multiple parameters, with the number of parameters increasing as the sample size increases. This differs from the usual set-up, where each term involves a single parameter, or possibly a small fixed number of parameters. Regarding the asymptotic distribution, a natural conjecture which follows previous work on isotonic estimation, is that our estimator has a $n^{1/3}$ rate of convergence and that $n^{1/3}\{\hat{\psi}_n(z_0)-\psi(z_0)\}$ converges to $C(z_0)\mathbb{Z}$, where $C(z_0)$ is constant depending on z_0 and $\mathbb{Z}$ is a Chernoff distribution random variable. Figure 3.2 shows quantile-quantile plots of the sample quantiles versus theoretical quantiles from Chernoff's distribution (Groeneboom and Wellner 2001). The linearity of the plots as the same size increases support this conjecture. Other scenarios show similar behaviors (results not shown). Future work is needed to rigorously derive the asymptotic properties of the isotonic estimators proposed in this paper.

3.7 Technical Details for Chapter 3

Proof of Theorem 3.1. We first prove the convexity by showing $xH(\psi)x \geq 0$ for any $x \neq 0 \in \mathbb{R}^n$. For notational convenience, we will drop ψ in the Hessian matrix. Since $h_{st} \leq 0$ for $s, t = 1, \ldots, n$ $(s \neq t)$,

$$\begin{aligned} xHx &= \textstyle\sum_{s=1}^n h_{ss}x_s^2 + 2\sum\sum_{s<t} h_{st}x_s x_t \\ &= \textstyle\sum_{s=1}^n h_{ss}x_s^2 + \sum\sum_{s<t}\big\{(-h_{st})^{\frac{1}{2}}x_s + (-h_{st})^{\frac{1}{2}}x_t\big\}^2 + \sum\sum_{s<t}(h_{st}x_s^2 + h_{st}x_t^2) \\ &= \textstyle\sum\sum_{s<t}\big\{(-h_{st})^{\frac{1}{2}}x_s + (-h_{st})^{\frac{1}{2}}x_t\big\}^2 + \sum_{s=1}^n\big\{h_{ss}x_s^2 + \sum_{s<t}(h_{st}x_s^2 + h_{st}x_t^2)\big\} \\ &= \textstyle\sum\sum_{s<t}\big\{(-h_{st})^{\frac{1}{2}}x_s + (-h_{st})^{\frac{1}{2}}x_t\big\}^2 + \sum_{s=1}^n\big\{h_{ss} + \sum_{t=1,t\neq s}^n h_{st}\big\}x_s^2. \end{aligned}$$

The first term is greater than or equal to 0, so the convexity holds by showing that the second term is 0, which is

$$\begin{aligned} h_{ss} + \textstyle\sum_{t=1,t\neq s}^n h_{st} &= \displaystyle\int_0^\infty \Big\{E_s(\psi,u) - E_s(\psi,u)^2 - \textstyle\sum_{t=1,t\neq s}^n E_s(\psi,u)E_t(\psi,u)\Big\}d\bar{N}(u) \\ &= \displaystyle\int_0^\infty E_s(\psi,u)\Big\{1 - \textstyle\sum_{t=1}^n E_t(\psi,u)\Big\}d\bar{N}(u) = 0 \end{aligned}$$

for $s = 1, \ldots, n$. The last equality holds because $\sum_{t=1}^n E_t(\psi,u) = \{\sum_{t=1}^n Y_t(u)e^{\psi_t}\}/\{\sum_{j=1}^n Y_j(u)e^{\psi_j}\} = 1$. Therefore, the Hessian matrix is semi-positive definite.

We next prove the strict convexity by imposing the anchor constraint where $\psi_k = 0$. Similarly, for any $y \neq 0 \in \mathbb{R}^n$ with $y_k = 0$,

$$yHy = \sum_{s<t,s\neq k,t\neq k}\sum \big\{(-h_{st})^{\frac{1}{2}}y_s + (-h_{st})^{\frac{1}{2}}y_t\big\}^2 + \sum_{s=1,s\neq k}^n \big\{h_{ss} + \sum_{t=1,t\neq s,t\neq k}^n h_{st}\big\}y_s^2.$$

The second term is strictly greater than 0 because

$$h_{ss} + \sum_{t=1,t\neq s,t\neq k}^{n} h_{st} = \int_0^\infty \Big\{E_s(\psi,u) - E_s(\psi,u)^2 - \sum_{t=1,t\neq s,t\neq k}^{n} E_s(\psi,u)E_t(\psi,u)\Big\} d\bar{N}(u)$$
$$= \int_0^\infty E_s(\psi,u)\Big\{1 - \textstyle\sum_{t=1,t\neq k}^{n} E_t(\psi,u)\Big\} d\bar{N}(u) \geq E_s(\psi, X_{(1)})\Big\{1 - \textstyle\sum_{t=1,t\neq k}^{n} E_t(\psi, X_{(1)})\Big\}$$
$$\geq \frac{e^{\psi_s}}{\sum_{j=1}^{n} e^{\psi_j}}\left(1 - \frac{\sum_{t=1,t\neq k}^{n} e^{\psi_t}}{\sum_{j=1}^{n} e^{\psi_j}}\right) > 0$$

for $s = 1,\dots,n\,(s \neq k)$. Thus, under the anchor constraint, the Hessian matrix is positive definite so that $l^N(\psi)$ is strictly convex.

Proof of Theorem 3.2. Since $l^N(\psi)$ is a convex function and Ψ^k is a convex cone, Lemma 2.1 (Groeneboom 1996) is directly applicable, where $\hat{\psi}$ minimizes $l^N(\psi)$ over Ψ^k if and only if

$$\textstyle\sum_{i=1,i\neq k}^{n} \psi_i u_i(\hat{\psi}) \geq 0 \text{ for all } \psi \in \Psi^k, \tag{3.13}$$

$$\textstyle\sum_{i=1,i\neq k}^{n} \hat{\psi}_i u_i(\hat{\psi}) = 0. \tag{3.14}$$

Since (3.14) is the same as (3.3), we claim that (3.13) is equivalent to (3.2). Suppose that $\hat{\psi}$ satisfies (3.13). Let $\alpha_i = \psi_i - \psi_{i+1}$ for $i = 1,\dots,k-1$ and $\alpha_i = \psi_i - \psi_{i-1}$ for $i = k+1,\dots,n$. For any $\psi \in \Psi^k$, ith element of ψ is expressed as $\sum_{j=i}^{k-1}\alpha_j$ for $i = 1,\dots,k-1$, or $\sum_{j=k+1}^{i}\alpha_j$ for $i = k+1,\dots,n$. Thus,

$$0 \leq \sum_{i=1,i\neq k}^{n} \psi_i u_i(\hat{\psi}) = \sum_{i=1}^{k-1} \psi_i u_i(\hat{\psi}) + \sum_{i=k+1}^{n} \psi_i u_i(\hat{\psi}) = \sum_{i=1}^{k-1}\{\sum_{j=i}^{k-1}\alpha_j\}u_i(\hat{\psi}) + \sum_{i=k+1}^{n}\{\sum_{j=k+1}^{i}\alpha_j\}u_i(\hat{\psi})$$
$$= \textstyle\sum_{i=1}^{k-1}\{\sum_{j=1}^{i} u_j(\hat{\psi})\}\alpha_i + \sum_{i=k+1}^{n}\{\sum_{j=i}^{n} u_j(\hat{\psi})\}\alpha_i, \tag{3.15}$$

which yields (3.2) because $\alpha_i \leq 0$ for $i = 1,\dots,k-1$ and $\alpha_i \geq 0$ for $i = k+1,\dots,n$. The other direction is trivial, because (3.15) is greater than or equal to zero when (3.2) holds.

If $k = 1$ (or $k = n$), then it is easily shown that the first (or second) inequality in (3.2) is removed, because the left (or right) term in (3.15) is removed.

The uniqueness condition holds, since $l^N(\psi)$ is a strictly convex function from Theorem 3.1, where the same statement is made by Proposition 1.1 (Groeneboom and Wellner 1992).

Proof of Theorem 3.3. The proof is analogous to that of Theorem 3.2. Since $l^P(\psi|\nu)$ satisfies the conditions of Lemma 2.1 (Groeneboom 1996) over the convex cone Ψ, $\hat{\psi}$ minimizes $l^P(\psi|\nu)$ over Ψ if and only if

$$\textstyle\sum_{i=1}^n \psi_i u_i^P(\hat{\psi}_i|\nu) \geq 0 \text{ for all } \psi \in \Psi, \tag{3.16}$$

$$\textstyle\sum_{i=1}^n \hat{\psi}_i u_i^P(\hat{\psi}_i|\nu) = 0. \tag{3.17}$$

Since (3.17) is the same as (3.10), we claim that (3.16) is equivalent to (3.9). Suppose that $\hat{\psi}$ satisfies (3.16). By setting $\alpha_i = \psi_i - \psi_{i-1}$ for $i = 1, \dots, n$ with $\psi_0 = 0$,

$$0 \leq \sum_{i=1}^n \psi_i u_i^P(\hat{\psi}_i|\nu) = \sum_{i=1}^n \Big\{\sum_{j=1}^i \alpha_j\Big\} u_i^P(\hat{\psi}_i|\nu) = \Big\{\sum_{j=1}^n u_j^P(\hat{\psi}_j|\nu)\Big\}\alpha_1 + \sum_{i=2}^n \Big\{\sum_{j=i}^n u_j^P(\hat{\psi}_j|\nu)\Big\}\alpha_i,$$

which yields (3.9) because $\alpha_i \geq 0$ for $i = 2, \dots, n$. The other direction is trivial.

The uniqueness condition holds because $l^P(\psi|\nu)$ is strictly convex due to (3.7).

Proof of Theorem 3.4. Using the max-min formula (Robertson et al. 1988, pp.23-24), the isotonic estimator $\hat{\psi}_i^+$ have the closed form solution of $\max_{s \leq i} \min_{t \geq i} \sum_{j=s}^t \Delta_j / \sum_{j=s}^t w_j$ for $i = 1, \dots, n$. Let $M_b = \{b_1, \dots, b_{n_b}\}$ be a block of consecutive indices $b_1, \dots, b_{n_b}$ for $b = 1, \dots, B$, where $\hat{\psi}^+_{b_1-1} < \hat{\psi}^+_{b_1} = \cdots = \hat{\psi}^+_{b_{n_b}} < \hat{\psi}^+_{b_{n_b}+1}$, with $\hat{\psi}^+_0 = -\infty$ and $\hat{\psi}^+_{n+1} = \infty$. Since $\hat{\psi}^+_i$

is constant on M_b, the max-min formula gives the following inequality, which is

$$\hat{\psi}_a^+ \le \frac{\sum_{i=b_1}^{a} \Delta_i}{\sum_{i=b_1}^{a} w_i} \Leftrightarrow \sum_{i=b_1}^{a} (-\Delta_i + \hat{\psi}_i^+ w_i) \le 0$$

for $a \in M_b$ with equality holding when $a = b_{n_b}$. The above two inequalities are the same because w_i's are strictly positive for $i = 1, \ldots, n$. Since $\sum_{i \in M_b}(-\Delta_i + \hat{\psi}_i^+ w_i) = 0$ for $b = 1, \ldots, B$ and $\cup_{b=1}^{B} M_b = \{1, \ldots, n\}$, then

$$0 \le \sum_{j=1}^{n}(-\Delta_j + \hat{\psi}_j^+ w_j) - \sum_{j=1}^{i-1}(-\Delta_j + \hat{\psi}_j^+ w_j) = \sum_{j=i}^{n}(-\Delta_j + \hat{\psi}_j^+ w_j) \quad (i = 1, \ldots, n) \tag{3.18}$$

with equality holding when $i = 1$, and

$$0 = \sum_{b=1}^{B} \sum_{i \in M_b} (-\Delta_i + \hat{\psi}_i^+ w_i) = \sum_{b=1}^{B} \sum_{i \in M_b} (-\Delta_i + \hat{\psi}_i^+ w_i) \log(\hat{\psi}_i^+) = \sum_{i=1}^{n}(-\Delta_i + \hat{\psi}_i^+ w_i) \log(\hat{\psi}_i^+), \tag{3.19}$$

where $\sum_{j=1}^{0} = 0$. Let $\dot{\psi}_i = \log(\hat{\psi}_i^+)$. The logarithmic transformation is well defined because $\hat{\psi}_i^+ > 0$ for $i = 1, \ldots, n$. Then, (3.18) and (3.19) are expressed as $\sum_{j=i}^{n}\{-\Delta_j + \exp(\dot{\psi}_i) w_j\}$ and $\sum_{i=1}^{n}\{-\Delta_i + \exp(\dot{\psi}_i) w_i\}\dot{\psi}_i$, respectively, which yield Fenchel's duality condition in (3.9) and (3.10). Thus, by Theorem 3.3, $\dot{\psi}$ is the unique minimizer of $l^P(\psi|\nu)$ over Ψ.

Proof of Theorem 3.5. In the proof, we write r instead of $r(\dot{\epsilon})$ for a notational convenience. We first note that

$$\begin{aligned} &\left|u_i(\ddot{\psi}) - u_i^P(\dot{\psi}_i^{(r)}|\dot{\psi}^{(r-1)})\right| = \left|u_i(\dot{\psi}^{(r)}) - u_i^P(\dot{\psi}_i^{(r)}|\dot{\psi}^{(r-1)})\right| \\ &= \int_0^\infty \left|\left\{E_i(\dot{\psi}^{(r)}, t) - E_i^P(\dot{\psi}^{(r)}, t|\dot{\psi}^{(r-1)})\right\} d\bar{N}(t)\right| \le \mu d_e(\dot{\psi}^{(r)}, \dot{\psi}^{(r-1)}) < \mu \dot{\epsilon}, \end{aligned} \tag{3.20}$$

for $i = 1, \dots, n$ $(i \neq k)$, where

$$0 < \mu = \max_{i\in\{1,\dots,n\}} \int_0^\infty \left[\frac{Y_i(t)e^{\dot{\psi}_i^{(r)}} d\bar{N}(t)}{\{\sum_{j=1}^n Y_j(t)e^{\dot{\psi}_j^{(r)}}\}\{\sum_{j=1}^n Y_j(t)e^{\dot{\psi}_j^{(r-1)}}\}} \right] < \infty.$$

Next, since $\dot{\psi}^{(r)}$, which is the unique minimizer of $l^P(\psi \mid \dot{\psi}^{(r-1)})$ over Ψ, satisfies Fenchel's duality condition in (3.9) and (3.10) in Theorem 3.3, we establish the following inequality and equality conditions by using (3.20) in conjunction with the triangle inequality, where the first inequality in (3.9) shows

$$0 \geq \sum_{j=1}^n u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) - \sum_{j=i+1}^n u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) = \sum_{j=1}^i u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \quad (i = 1, \dots, k-1),$$

which implies

$$\sum_{j=1}^i u_j(\ddot{\psi}) \leq \sum_{j=1}^i \left\{ u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \right\} \leq \left| \sum_{j=1}^i u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \right| \leq i\mu\dot{\epsilon}, \tag{3.21}$$

the second inequality in (3.9) shows

$$0 \leq \sum_{j=i}^n u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \quad (i = k+1, \dots, n),$$

which implies

$$\sum_{j=i}^n u_j(\ddot{\psi}) \geq \sum_{j=i}^n \left\{ u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \right\} \geq - \left| \sum_{j=i}^n u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} | \dot{\psi}^{(r-1)}) \right| \geq -(n-i+1)\mu\dot{\epsilon}, \tag{3.22}$$

and the equality in (3.10) shows

$$\begin{aligned} 0 &= \sum_{i=1}^n \dot{\psi}_i^{(r)} u_i^P(\dot{\psi}_i^{(r)} | \dot{\psi}^{(r-1)}) = \sum_{i=1}^n (\dot{\psi}_i^{(r)} - \dot{\psi}_k^{(r)}) u_i^P(\dot{\psi}_i^{(r)} | \dot{\psi}^{(r-1)}) + \dot{\psi}_k^{(r)} \sum_{i=1}^n u_i^P(\dot{\psi}_i^{(r)} | \dot{\psi}^{(r-1)}) \\ &= \sum_{i=1, i\neq k}^n \ddot{\psi} u_i^P(\dot{\psi}_i^{(r)} | \dot{\psi}^{(r-1)}), \end{aligned}$$

which implies

$$\left|\sum_{i=1,i\neq k}^{n} \ddot{\psi}_i u_i(\ddot{\psi})\right| = \left|\sum_{i=1,i\neq k}^{n} \ddot{\psi}_i\{u_i(\ddot{\psi}) - u_i^P(\dot{\psi}_i^{(r)}|\dot{\psi}^{(r-1)})\}\right| \leq \sum_{i=1,i\neq k}^{n} \left|\ddot{\psi}_i\right|\mu\dot{\epsilon}. \qquad (3.23)$$

Finally, suppose that the pseudo iterative convex minorant algorithm converges under the stopping value of $\dot{\epsilon} > 0$. As $\dot{\epsilon}$ converges to zero, all the bounds in (3.21) - (3.23) converge to zero for each fixed n, which yield Fenchel's duality condition in (3.2) and (3.3). Thus, by Theorem 3.2, $\ddot{\psi}$ converges to the unique minimizer of $l^N(\psi)$ over Ψ^k.

Proof of Proposition 3.6. Under censored data, $pl(\psi)$ can be reformulated as

$$pl(\psi) = \prod_{i=1}^{n}\int_0^\infty \left\{\frac{e^{\psi_i}}{\sum_{j=1}^n Y_j(t)e^{\psi_j}}\right\}^{dN_i(t)} = \prod_{i=1}^{n^\star}\int_0^\infty \left\{\frac{e^{\psi_i^\star}}{\sum_{j=1}^n Y_j(t)e^{\psi_j}}\right\}^{dN_i^\star(t)}. \qquad (3.24)$$

Suppose that jth subject is censored, where $Z_{(h)}^\star \leq Z_j \leq Z_{(h+1)}^\star$ for $h = 1,\ldots,n^\star$ with $Z_{(n^\star+1)}^\star = \infty$. Since ψ_j is only included in the denominator in (3.24), $pl(\psi)$ is maximized when $\psi_j = \psi_h^\star$. If $Z_j \leq Z_{(1)}^\star$, then $pl(\psi)$ is maximized when $\psi_j = \psi_1^\star$ by the assumption. This shows that the isotonic estimator has jumps only at $Z_i^\star$. Thus, the denominator in (3.24) can be reduced to $\sum_{j=1}^{n^\star} Y_j^\star(t)\exp(\psi_j^\star)$, and thus, (3.24) can be reduced to $pl^C(\psi)$. It follows that the unique maximizer of $pl^C(\psi)$ is also the unique maximizer of $pl(\psi)$.

Proof of Proposition 3.7. This proof is analogous to that of Proposition 3.6. Under censored data with a time-dependent covariate, $pl(\psi)$ can be reformulated as

$$pl(\psi) = \prod_{i=1}^{n}\int_0^\infty \left\{\frac{e^{\psi\{Z_i(t)\}}}{\sum_{j=1}^n Y_j(t)e^{\psi\{Z_j(t)\}}}\right\}^{dN_i(t)} = \prod_{i=1}^{n^\star}\int_0^\infty \left\{\frac{e^{\psi(Z_i^\star)}}{\sum_{j=1}^n Y_j(t)e^{\psi\{Z_j(t)\}}}\right\}^{dN_i^\star(t)}. \qquad (3.25)$$

Likewise, $pl(\psi)$ is maximized only when $\psi\{Z_i(X_j)\}$ in the denominator in (3.25), $i,j = 1,\ldots,n$, is replaced with one of $\psi(Z_h^\star)$'s for $h = 1,\ldots,n^\star$. This shows that the isotonic estimator has jumps only at Z_h^*. Thus, the denominator in (3.25) can be reduced to $\sum_{j=1}^{n^*} Y_j^*(t)\exp\{\psi(Z_j^*)\}$, and thus, (3.25) can be reduced to $pl^D(\psi)$. It follows that the unique maximizer of $pl^D(\psi)$ is also the unique maximizer of $pl(\psi)$.

Proof of Theorem 3.8. We follow from Lemma 2.1 in the detailed version of Banerjee (2007) on his webpage (M. Banerjee, University of Michigan). Denote $\tilde{\psi}_n$ minimizes $\mathbb{P}_n\{l_n(\psi(Z)\mid \underline{\psi}_0, \underline{\epsilon}_n)\}$ among all monotonically increasing functions between L and U. Since $(1-\zeta)\tilde{\psi}_n + \zeta\psi_0$ is a monotonically increasing function between L and U for $\zeta > 0$, we have

$$\lim_{\zeta\to 0^+} \frac{\mathbb{P}_n\{l_n(\tilde{\psi}_n(Z)|\underline{\psi}_0, \underline{\epsilon}_n)\} - \mathbb{P}_n\{l_n((1-\zeta)\tilde{\psi}_n(Z) + \zeta\psi_0(Z)|\underline{\psi}_0, \underline{\epsilon}_n)\}}{\zeta} \le 0.$$

It follows that

$$\mathbb{P}_n\big\{u_n(\tilde{\psi}_n(Z)|\underline{\psi}_0, \underline{\epsilon}_n)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\} \le 0. \tag{3.26}$$

We claim that for every $\omega \in \Omega$, there exists a set Ω with $P(\Omega) = 1$ such that

$$\begin{aligned}
&\big|\mathbb{P}_{n,\omega}\big\{u_n(\tilde{\psi}_n(Z)|\underline{\psi}_0, \underline{\epsilon}_n)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\} - P\big\{u(\tilde{\psi}_n(Z)|\underline{\psi}_0)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\}\big| \\
&\le \big|\mathbb{P}_{n,\omega}\big\{u_n(\tilde{\psi}_n(Z)|\underline{\psi}_0, \underline{\epsilon}_n)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\} - \mathbb{P}_{n,\omega}\big\{u_n(\tilde{\psi}_n(Z)|\underline{\psi}_0)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\}\big| \\
&+ \big|\mathbb{P}_{n,\omega}\big\{u_n(\tilde{\psi}_n(Z)|\underline{\psi}_0)(\tilde{\psi}_n(Z) - \psi_0(Z))\big\} - P\big\{u(\tilde{\psi}_n(Z)|\psi_0(Z))(\tilde{\psi}_n(Z) - \psi_0(Z))\big\}\big|
\end{aligned} \tag{3.27}$$

converges to zero almost surely as n go to infinity over all bounded monotone increasing function, where $\mathbb{P}_{n,\omega}$ is the empirical measure on $\{X_i(\omega), \Delta_i(\omega), Z_i(\omega)\}_{i=1}^n$. Take $\epsilon_{n,i} = c_{n,i}/n$, $i = 1, \ldots, n$, such that $\nu_{n,i} = \psi_0(Z_i) + c_{n,i}/n$ satisfies a monotone constraint, where $c_{n,i}$ has finite lower and upper bounds of l and u, respectively. The first term in (3.27) is bounded above by

$$
\begin{aligned}
&= \int_0^\tau \left| \left\{ n^{-1} \textstyle\sum_{j=1}^n Y_j(t) e^{\psi_0(Z_j)} (1 - e^{c_{n,j}/n}) \right\} \frac{\mathbb{P}_{n,\omega}\{Y(t)e^{\psi(Z)}(\psi_0(Z) - \tilde{\psi}_n(Z))\}\mathbb{P}_{n,\omega}\{dN(t)\}}{\mathbb{P}_{n,\omega}\{Y(t)e^{\psi_0(Z)}\}\{n^{-1}\sum_{j=1}^n Y_j(t)e^{\psi_0(Z_j)+c_{n,j}/n}\}} \right| \\
&\le \int_0^\tau \left[\mathbb{P}_{n,\omega}\{Y(t)e^U \max\{\left|1 - e^{u/n}\right|, \left|1 - e^{l/n}\right|\}\} \frac{\mathbb{P}_{n,\omega}\{Y(t)e^U(U - L)\}\mathbb{P}_{n,\omega}\{dN(t)\}}{\mathbb{P}_{n,\omega}\{Y(t)e^L\}\mathbb{P}_{n,\omega}\{Y(t)e^{L+l/n}\}} \right] \\
&= \max\{\left|e^{-l/n} - e^{(u-l)/n}\right|, \left|e^{-l/n} - 1\right|\} \int_0^\tau \mathbb{P}_{n,\omega}\{dN(t)\}e^{2(U-L)}(U - L), \qquad (3.28)
\end{aligned}
$$

since $\tilde{\psi}_n(\cdot)$ and $\psi_0(\cdot)$ have finite lower and upper bounds of L and U, respectively, by Assumption (A4). Since $\exp(-1/n)$ converges to 1 as n go to infinite, (3.28) converges to zero, where the other terms in (3.28) converges to constant times $\int_0^\tau P\{dN(t)\}$, which is finite by Assumption (A1). Thus, the first term in (3.27) converges to zero. For the second term in (3.27), since a class of bounded monotone increasing function is P-Glivenko-Cantelli by Lemma 2.1 in the detailed version of Banerjee (2007) on his webpage (M. Banerjee, University of Michigan) and N is P-Glivenko-Cantelli by Lemma 4.1 (Kosorok 2007), then $u_n(\tilde{\psi}_n(Z)\,|\,\underline{\psi}_0)(\psi_0(Z)-\tilde{\psi}_n(Z))$, which is a continuous function of P-Glivenko-Cantelli functions, is P-Glivenko-Cantelli by standard preservation property of Glivenko-Cantelli class in Theorem 3 (van der Vaart and Wellner 2000, pp. 115-133) provided that it has an integral envelope. The integrable envelope is simply given by, for example, constant times $|u(U \mid L)|$ by Assumption (A5). The denominator in $u(\tilde{\psi}_n(Z)\,|\,\psi_0(Z))$ is bounded away from 0 by Assumption (A1). Thus, the second term converges to zero almost surely, where a similar statement in (3.27) is made by Kosorok (2007, p.56).

Fix an $\omega \in \Omega$. By the Helly selection theorem, $\{\tilde{\psi}_{n,\omega}(\cdot)\}$ has a convergent subsequence $\{\tilde{\psi}_{n_k,\omega}(\cdot)\}$ that converges to a monotonically increasing right continuous function $\bar{\psi}_\omega(\cdot)$

bounded by L and U. This implies that

$$\lim_{k\to\infty} \mathbb{P}_{n_k,\omega}\{u_n(\tilde{\psi}_{n_k,\omega}(Z)\,|\,\underline{\psi}_0)(\psi_0(Z)-\tilde{\psi}_{n_k,\omega}(Z))\} = P\{u(\bar{\psi}_\omega(Z)\,|\,\psi_0(Z))(\psi_0(Z)-\bar{\psi}_\omega(Z))\}. \tag{3.29}$$

To show this, (3.29) is expressed as

$$\begin{aligned}&\mathbb{P}_{n_k,\omega}\{u_n(\tilde{\psi}_{n_k,\omega}(Z)\,|\,\underline{\psi}_0)(\psi_0(Z)-\tilde{\psi}_{n_k,\omega}(Z))\} - P\{u(\tilde{\psi}_{n_k,\omega}(Z)\,|\,\psi_0(Z))(\psi_0(Z)-\tilde{\psi}_{n_k,\omega}(Z))\}\\ &+ P\{u(\tilde{\psi}_{n_k,\omega}(Z)\,|\,\psi_0(Z))(\psi_0(Z)-\tilde{\psi}_{n_k,\omega}(Z))\} - P\{u(\bar{\psi}_\omega(Z)\,|\,\psi_0(Z))(\psi_0(Z)-\bar{\psi}_\omega(Z))\},\end{aligned}$$

where the first two term converges to zero by (3.27), and the second term is expressed as

$$\int_{z\in I_z}\big[E_{\psi_0(z)}\{u(\tilde{\psi}_{n_k,\omega}(z))\}(\psi_0(z)-\tilde{\psi}_{n_k,\omega}(z)) - E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z))\}(\psi_0(z)-\bar{\psi}_\omega(z))\big]p_z(z)dz,$$

which converges to zero by the dominated convergence theorem.

In conjunction (3.26) with (3.29),

$$0 \geq \int_{z\in I_z} E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z)\,|\,\psi_0(z))\}(\bar{\psi}_\omega(z)-\psi_0(z))p_z(z)dz = \int_{z\in I_z}\eta(z)dz, \tag{3.30}$$

where

$$\eta(z) = \big[E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z)\,|\,\psi_0(z))\} - E_{\psi_0(z)}\{u(\psi_0(z)\,|\,\psi_0)\}\big]\big[\bar{\psi}_\omega(z)-\psi_0(z)\big]p_z(z).$$

Suppose that $\bar{\psi}_\omega(z) > \psi_0(z)$. Then, $u(\bar{\psi}_\omega(z) \mid \psi_0(z)) \geq u(\psi_0(z) \mid \psi_0(z))$. Consequently, $E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z))\} \geq E_{\psi_0(z)}\{u(\psi_0(z))\}$ and $[E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z))\} - E_{\psi_0(z)}\{u(\psi_0(z))\}][\bar{\psi}_\omega(z) - \psi_0(z)] \geq 0$. The same inequality holds when $\bar{\psi}_\omega(z) < \psi_0(z)$. It is obvious that (3.30) equals zero when $\bar{\psi}_\omega(z) = \psi_0(z)$. Thus, (3.30) is greater than or equal to 0, and it is equal to 0.

Fix $a < b$ in the interior of I_z such that $a < z_0 < b$, where $\eta(z)$ is a right continuous function on (a,b). Suppose that $\eta(z) \neq 0$ at any point $z \in (a,b)$. Then $\eta(z) > 0$, because $[E_{\psi_0(z)}\{u(\bar{\psi}_\omega(z) \mid \psi_0(z))\} - E_{\psi_0(z)}\{u(\psi_0(z) \mid \psi_0(z))\}]$ and $[\bar{\psi}_\omega(z) - \psi_0(z)]$ are both strictly positive. The right continuity of $\eta(z)$ with Assumptions (A2) and (A6) implies that (3.30) is strictly positive, which is a contradiction. Hence, we conclude that $\eta(z) = 0$ for $z \in (a,b)$, and thus, $\bar{\psi}_\omega(z) = \psi_0(z)$ for $z \in (a,b)$. Since $\psi_0(\cdot)$ is a continuous function by Assumption (A3), $\sup_{z\in[\sigma_1,\sigma_2]} |\tilde{\psi}_{n,\omega}(z) - \psi_0(z)|$ converges to zero on any compact set $[\sigma_1,\sigma_2]$ of (a,b). Furthermore, since $\psi_0(\cdot)$ has a strictly positive derivative in a neighborhood of z_0 by Assumption (A3) then $\tilde{\psi}_{n,\omega}(z)$ lies strictly between L and U for each $z \in [\sigma_1,\sigma_2]$ and for each $\omega \in \Omega$ by choosing σ_1 and σ_2 sufficiently close to z_0. Now, we compute $\dot{\psi}^{(1)}_{n,\omega}(z)$, and then, set $\tilde{\psi}_{n,\omega}(z) = \dot{\psi}^{(1)}_{n,\omega}(z)$ if $L < \dot{\psi}^{(1)}_{n,\omega}(z) < U$; $\tilde{\psi}_{n,\omega}(z) = L$ if $\dot{\psi}^{(1)}_{n,\omega}(z) < L$; $\tilde{\psi}_{n,\omega}(z) = U$ if $\dot{\psi}^{(1)}_{n,\omega}(z) > U$. Consequently, $\tilde{\psi}_{n,\omega}(z) = \dot{\psi}^{(1)}_{n,\omega}(z)$ if $L < \tilde{\psi}_{n,\omega}(z) < U$. It follows that $\sup_{z\in[\sigma_1,\sigma_2]} |\dot{\psi}^{(1)}_{n,\omega}(z) - \psi_0(z)|$ converges to zero on any compact set $[\sigma_1,\sigma_2]$ of (a,b). This shows strong consistency of $\dot{\psi}^{(1)}(\cdot)$ for $\psi_0(\cdot)$.

For censored data, the same argument can be made for strong consistency by defining empirical and probability measures on $\{X_i, \Delta_i = 1, Z_i\}$ and $\{X, \Delta = 1, Z\}$, respectively, where the isotonic estimator has only a jump at Z_i with $\Delta_i = 1$, $i = 1, \ldots, n$, by Proposition 3.6.

CHAPTER 4: SHAPE RESTRICTED ADDITIVE HAZARD MODELS: MONOTONE, UNIMODAL AND U-SHAPE HAZARD FUNCTIONS

4.1 Introduction

In many survival studies, a hazard function is assumed to have a shape restriction on a covariate such as monotone increasing or monotone decreasing hazard. In Chapter 3, we suggested the isotonic proportional hazard models denoted by $\lambda(t \mid Z) = \lambda_0(t)\exp\{\phi(Z)\}$, where $\lambda_0(\cdot)$ was an unspecified baseline hazard function, $\phi(\cdot)$ was a monotone function, and Z was a scalar continuous covariate. This model is only valid under the proportional hazard assumption. Furthermore, it might be computationally inefficient and unstable to maximize the partial likelihood (Cox 1972) with large sample sizes owing to the complicated nonlinear structured of the partial likelihood, with a large dimensionality (Gorst-Rasmussen and Scheike 2012). In cases where the proportionality assumption is violated or sample size is large, the additive hazards model may be a useful alternative. The model assumes that the effect of a risk factor is added to the hazard, defined as

$$\lambda(t \mid Z) = \lambda_0(t) + \phi(Z). \tag{4.1}$$

By following Lin and Ying (1994)'s approach, its loss function is defined as a quadratic function, which is formulated in Subsection 4.2.1. This simple structure may simplify computations, compared to the complicated structure of the partial likelihood.

The purpose of this paper is efficient and theoretically justified computation for estimating $\phi(\cdot)$ in (4.1) under a shape restriction such as a monotone, unimodal or U-shaped constraint. Specifically, our focus is on a unimodal hazard function, where the hazard is monotone increasing and monotone decreasing on the intervals of $(-\infty, M]$ and $[M, +\infty)$, respectively. The point M is called a mode which is generally unknown. We consider the approach of Shoung and Zhang (2001), who minimized a least squares function over a class of all unimodal functions. They proposed a profiling algorithm that estimated unimodal functions at all hypothetical modes and then a mode at which the loss function had a minimum. The profiling algorithm is directly applicable to our model. The simple structure of our quadratic loss function yields a global Hessian matrix that does not depend on any parameter. Thus, once the global Hessian matrix is computed, we can perform a standard quadratic programming method with a unimodal constraint within profiling the mode.

With large sample sizes, the global Hessian matrix may be quite time consuming to invert in the quadratic programming method owing to the high dimensionality, where the dimension of Hessian matrix and number of hypothetical modes are the same order as the sample size. Furthermore, unlike the least square function in the unimodal regression, our loss function is not separable in terms of observed covariate values so that the global Hessian matrix is not a diagonal but rather a full matrix. To overcome this challenge, we propose the quadratic pool-adjacent-violators algorithm which estimates the unimodal hazard function by minimizing a sequence of pseudo approximated loss functions. Computational efficiency is gained from the fact that each pseudo loss function has a closed form solution, and an efficient computation of the pool-adjacent-violators algorithm (Ayer et al. 1955) is easily implemented to the profiling algorithm.

We define the loss function in Subsection 4.2.1. The quadratic programming method and quadratic pool-adjacent-violators algorithm are described in Subsection 4.2.2 and Subsection 4.2.3 respectively, without censoring. Censoring and time-dependent covariates are described in Subsections 4.2.4 and 4.2.5, respectively. Extensions to estimation of monotone or U-shape hazard, estimation of baseline hazard function and inclusion of additional covariates are described in Section 4.3. In simulation results reported in Section 4.4, the quadratic pool-adjacent-violators algorithm improves computational speeds compared to the quadratic programming method with bias and mean square error reductions. An analysis of a cardiovascular disease dataset in Section 4.5 illustrates the practical utility of our methodology in estimating mode with a nonlinear covariate effect. All proofs are given in Section 4.6.

4.2 Shape restricted additive hazard model

4.2.1 Data set-up and loss function

Suppose that T is a failure time and C is a censoring time. Assume that T and C are conditionally independent on Z. Let $X = min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. The observed data consist of n independent and identically distributed replicates of (X, Δ, Z), denoted by $\{X_i, \Delta_i, Z_i\}$ for $i = 1, \dots, n$. Define $N_i(t) = \Delta_i I(X_i \leq t)$ as a counting process and $Y_i(t) = I(X_i \geq t)$ as an at-risk process for the ith subject. Denote $Z_{(i)}$ as the ith smallest value amongst $Z_1, \dots, Z_n$. Under the isotonic proportional hazard model, the ith element of the (negative) score function $u_i^N(\phi)$ is defined as

$$u_i^N(\phi) = \int_0^\infty \left\{ -dN_{(i)}(t) + Y_{(i)}(t) e^{\phi(Z_{(i)})} d\tilde{\Lambda}_0(\phi, t) \right\} \tag{4.2}$$

for $i = 1, \dots, n$, where $N_{(i)}(t)$ and $Y_{(i)}(t)$ are the counting and at-risk processes corresponding to $Z_{(i)}$, respectively, and where

$$\tilde{\Lambda}_0(\phi, t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n Y_j(u) e^{\phi(Z_j)}}.$$

Based on Lin and Ying (1994)'s approach to the additive hazard model in (4.1), we mimic the score function (4.2), which is

$$u_i(\phi) = \int_0^\infty \left\{-dN_{(i)}(t) + Y_{(i)}(t)d\hat{\Lambda}_0(\phi, t) + Y_{(i)}(t)\phi(Z_{(i)})dt\right\} \tag{4.3}$$

for $i = 1, \ldots, n$, where

$$\hat{\Lambda}_0(\phi, t) = \int_0^t \frac{\sum_{i=1}^n \{dN_i(u) - Y_i(u)\phi(Z_i)du\}}{\sum_{j=1}^n Y_j(u)}. \tag{4.4}$$

By plugging $\hat{\Lambda}_0(\phi, t)$ in (4.4) into $u_i(\phi)$ in (4.3), we reformulate the score function in (4.3) as

$$u_i(\phi) = \textstyle\sum_{j=1}^n \left\{h_{ij}\phi(Z_{(j)})\right\} - q_i, \tag{4.5}$$

where

$$h_{ij} = \int_0^\infty \left\{Y_{(i)}(t)I(i = j) - \frac{Y_{(i)}(t)Y_{(j)}(t)}{\sum_{s=1}^n Y_{(s)}(t)}\right\}dt, \tag{4.6}$$

$$q_i = \int_0^\infty dN_{(i)}(t) - \int_0^\infty \left\{Y_{(i)}(t)\frac{\sum_{l=1}^n dN_{(l)}(t)}{\sum_{s=1}^n Y_{(s)}(t)}\right\} \tag{4.7}$$

for $i, j = 1, \ldots, n$. Accordingly, we define the quadratic loss function as

$$l(\phi) = \frac{1}{2}\phi^T H\phi - \phi^T q, \tag{4.8}$$

where $\phi = \{\phi(Z_{(1)}), \ldots, \phi(Z_{(n)})\}^T$. Here, H and q are an $n \times n$ matrix and an $n \times 1$ vector with elements in (4.6) and (4.7), respectively. An important point is that the loss function is a quadratic function and H does not involve any unknown parameters in ϕ.

4.2.2 Quadratic programming method with no censoring

We want to find a minimizer in ϕ of the loss function in (4.8) under shape restrictions, such as a monotone, unimodal or U-shaped constraint on ϕ. In this section, we focus on the unimodal function ϕ where is monotone increasing and monotone decreasing on $(-\infty, M]$ and $[M, +\infty)$, respectively, along with the mode M. The monotone or U-shaped function ϕ is described in Section 4.3.

We first assume that the mode M is known. As stated in Theorem 4.1 below, we impose an anchor constraint $\phi(M) = \delta$ for a constant M to guarantee a unique minimizer of the loss function. Under the anchor constraint, the model being fitted in (4.1) is reformulated by

$$\lambda(t \mid Z) = \{\lambda_0(t) + \delta\} + \psi(Z), \tag{4.9}$$

where $\psi(\cdot) = \phi(\cdot) - \delta$ with $\psi(M) = 0$. Since the baseline hazard function absorbs δ, our focus is on estimation of $\psi(\cdot)$, not $\phi(\cdot)$. The vertical shift parameter δ is regarded as a nuisance parameter, because the only difference between $\phi(\cdot)$ and $\psi(\cdot)$ is the reference group for defining the hazard difference parameters. In other words, hazard differences based on $\phi(\cdot)$ and $\psi(\cdot)$ are identical, where $\phi(\cdot) - \phi(Z_R) = \psi(\cdot) - \psi(Z_R)$ for any reference value Z_R. In practice, since $\psi(\cdot)$ is only estimable at the observed $Z_{(i)}$'s, we set $\psi(Z_{(m)}) = 0$, where $Z_{(m)}$ is the largest $Z_{(i)} \leq M$.

Theorem 4.1. *Suppose that there is no censoring. The loss function $l(\psi)$ is convex. It is strictly convex when an anchor constraint is imposed that $\psi(Z_{(m)}) = 0$.*

Let $\psi_i = \psi(Z_{(i)})$, $i = 1, \ldots, n$. Denote $\Psi^m = \{\psi \in \mathbb{R}^n : \psi_1 \leq \ldots \leq \psi_m, \psi_{m+1} \geq \ldots \geq \psi_n, \psi_m = 0\}$ which is a convex cone. Then, the problem of minimizing the loss function under the unimodal and anchor constraints is equivalent to the problem of minimizing the strictly convex quadratic function $l(\psi)$ over the convex cone Ψ^m. We denote $\hat{\psi} = (\hat{\psi}_1, \ldots, \hat{\psi}_n)$ as the unimodal minimizer of $l(\psi)$ over Ψ^m. To estimate ψ at covariates values other that those in $Z_{(1)},\dots,Z_{(n)}$, we assume that the unimodal estimator is a right continuous step function with potential jumps at $Z_{(i)}$'s. Under this assumption, the uniqueness of the unimodal estimator is established in the following theorem:

Theorem 4.2. *Suppose that there is no censoring. The unimodal estimator $\hat{\psi}$ minimizes $l(\psi)$ over the convex cone Ψ^m if and only if Fenchel's duality condition holds that $\hat{\psi} \in \Psi^m$ satisfies*

$$\textstyle\sum_{j=1}^{i} u_j(\hat{\psi}) \leq 0 \quad (i = 1,\dots,m-1), \ \sum_{j=i}^{n} u_j(\hat{\psi}) \leq 0 \quad (i = m+1,\dots,n) \tag{4.10}$$

with equality holding if $i = m+1$, and

$$\textstyle\sum_{i=1,i\neq m}^{n} \hat{\psi}_i u_i(\hat{\psi}) = 0 \tag{4.11}$$

Moreover, $\hat{\psi}$ is uniquely determined by (4.10) and (4.11).

A quadratic programming method can be performed with equality and inequality constraints, which gives the unique minimizer that satisfies Fenchel's duality conditions in (4.10) and (4.11) in Theorem 4.2.

We next assume that M is unknown, which is a more general case. We apply a profiling algorithm that estimates the unimodal hazard functions at all hypothetical modes and estimates the mode to be the value at which the loss function has a minimum value, which is formalized as

$$\hat{M} = \left[Z_{(\hat{m})} : \hat{m} = \arg\min_{m\in(1,\dots,n)}\{\min_{\psi\in\Psi^m} l(\psi)\}\right]. \tag{4.12}$$

Since the global Hessian matrix is available in (4.6), which does not depend on any parameters in ψ, the quadratic programing method is easily performed by profiling the mode.

4.2.3 Quadratic pool-adjacent-violators algorithm with no censoring

Unlike unimodal linear regression under the additive hazards model, the global Hessian matrix is not a diagonal but rather a full matrix. When sample sizes increase, it is a challenge to handle the full Hessian matrix in the quadratic programming method. In particular, when the mode is unknown, it may be quite time consuming to perform the quadratic programing method at every hypothetical mode owing to high dimensionality, where both the dimension of the Hessian matrix and the number of hypothetical modes are of the same order as the sample size.

To improve the computational speed, we suggest the quadratic pool-adjacent-violators algorithm that estimates the unique minimizer of the loss function $l(\psi)$ over Ψ^m by minimizing a sequence of pseudo loss functions. By regarding some parameter ψ as a known constant ν, we approximate $l(\psi)$ in (4.8) by

$$\begin{aligned} l(\psi) &= \sum_{i=1}^{n}\Big\{\frac{1}{2}h_{ii}\psi_i^2 + \Big(\sum_{j=1,j\neq i}^{n} h_{ij}\psi_i\psi_j\Big) - q_i\psi_i\Big\} \\ &\approx \sum_{i=1}^{n}\Big\{\frac{1}{2}h_{ii}\psi_i^2 + \Big(\sum_{j=1,j\neq i}^{n} h_{ij}\psi_i\nu_j\Big) - q_i\psi_i\Big\} \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\psi_i - \frac{q_i - \sum_{j=1,j\neq i}^{n} h_{ij}\nu_j}{h_{ii}}\right)^2 h_{ii} + g(\nu) = l^P(\psi \mid \nu), \end{aligned} \tag{4.13}$$

where $g(\nu) = 2^{-1}\sum_{i=1}^{n}(q_i - \sum_{j=1,j\neq i}^{n} h_{ij}\nu_j)^2 h_{ii}^{-1}$ which does not depend on ψ. The pseudo loss function $l^P(\psi \mid \nu)$ in (4.13) is a strictly convex function because the pseudo Hessian matrix is a diagonal matrix having the diagonal elements of $h_{ii}^P = h_{ii} > 0$, $i = 1, \dots, n$, and the pseudo score function is defined as $u_i^P(\psi \mid \nu) = h_{ii}\psi_i + \sum_{j=1,j\neq i}^{n} h_{ij}\nu_j - q_i$.

Let $\dot{\Psi}^m = \{\psi \in \mathbb{R}^n : \psi_1 \leq \ldots \leq \psi_m, \psi_{m+1} \geq \ldots \geq \psi_n\}$ be the convex cone obtained by removing the anchor constraint from Ψ^m. Denote $\dot{\psi}$ as the unimodal minimizer of $l^P(\psi|\nu)$ over $\dot{\Psi}^m$. The procedure of quadratic pool-adjacent-violators algorithm is described by the following steps:

Step 4.1: Choose an initial value of $\psi^{(0)} \in \dot{\Psi}^m$ (or $\psi^{(0)} \in \Psi^m$).

Step 4.2 Update $\dot{\psi}^{(r)}$ such that $\dot{\psi}^{(r)} = \arg\min_{\psi \in \dot{\Psi}^m} l(\psi \mid \dot{\psi}^{(r-1)})$.

Step 4.3: Repeat Step 4.2 until convergence, where its convergence criteria is $d(\dot{\psi}^{(r)}, \dot{\psi}^{(r-1)}) = \sum_{i=1}^{n} |\dot{\psi}_i^{(r)} - \dot{\psi}_i^{(r-1)}| < \dot{\epsilon}$ for small $\dot{\epsilon} > 0$.

Step 4.4: Do a vertical shift, $\ddot{\psi}_i = \dot{\psi}_i^{(r)} - \dot{\psi}_m^{(r)}$, $i = 1, \ldots, n$, where $\ddot{\psi} \in \Psi^m$.

The pseudo loss function $l^P(\psi \mid \nu)$ in (4.13) has the form of the least square function. Thus, $\dot{\psi} \in \Psi^m$ has a closed form solution that satisfies Fenchel's duality condition in (4.14) and (4.15) Theorem 4.3 below, which can be easily computed by using increasing and decreasing pool-adjacent-violators algorithm on the convex cones $\{\psi \in \mathbb{R}^m : \psi_1 \leq \ldots \leq \psi_m\}$ and $\{\psi \in \mathbb{R}^{n-m} : \psi_{m+1} \geq \ldots \geq \psi_n\}$, separately.

Theorem 4.3. *The unimodal estimator $\hat{\psi}$ minimizes $l(\psi)$ over the convex cone Ψ if and only if Fenchel's duality condition holds that $\dot{\psi} \in \Psi$ satisfies*

$$\sum_{j=1}^{i} u_j^P(\dot{\psi} \mid \nu) \leq 0 \quad (i = 1, \ldots, m), \ \sum_{j=i}^{n} u_j^P(\dot{\psi} \mid \nu) \leq 0 \quad (i = m+1, \ldots, n), \tag{4.14}$$

with equality holding if $i = m$ or $i = m + 1$, and

$$\sum_{i=1}^{n} \dot{\psi}_i u_i^P(\dot{\psi} \mid \nu) = 0 \tag{4.15}$$

Moreover, $\dot{\psi}$ is uniquely determined by (4.14) and (4.15).

After the iterative pool-adjacent-violators algorithm converges in Step 4.3, the anchor constraint is imposed by the vertical shift in Step 4.4. It guarantees that the vertically shifted unimodal estimate $\ddot{\psi}$ is the unique minimizer of $l(\psi)$ over Ψ^m, as stated in the following theorem:

Theorem 4.4. *Suppose that for any $\dot{\epsilon} > 0$, there exists $r(\dot{\epsilon})$ such that the quadratic pool-adjacent-violators algorithm converges at $r(\dot{\epsilon})$th iteration under the distance stopping criteria $d(\dot{\psi}^{(r(\dot{\epsilon}))}, \dot{\psi}^{(r(\dot{\epsilon}))-1}) < \dot{\epsilon}$. Then, as $\dot{\epsilon} \to 0$, $\ddot{\psi} = (\dot{\psi}_1^{(r(\dot{\epsilon}))} - \dot{\psi}_k^{(r(\dot{\epsilon}))}, \ldots, \dot{\psi}_1^{(r(\dot{\epsilon}))} - \dot{\psi}_k^{(r(\dot{\epsilon}))})$ converges to the unique minimizer of $l(\psi)$ over Ψ^m.*

4.2.4 Censoring

Suppose that some subjects' failure times are censored. Under the proportional hazard assumption, the isotonic estimator jumps only at the covariate values associated with observed failure events by Proposition 3.6. Since our approach was based on a mimicked score function from the proportional hazard model in Subsection 4.2.1, we also restricted our unimodal estimator to a right continuous step function with jumps at the covariate values associated with uncensored subjects. Based on this restriction, we apply the replacement parameters algorithm that replaces a parameter for a censored subject with the parameter for an uncensored subject having covariate value which is closest to that for the censored subject amongst all uncensored subjects having smaller covariate values than the censored subject. This approach is formalized below.

Let $n^\star$ be the number of subjects having observed failure time out of the total n subjects, and $Z_i^\star$ be their covariates for $i = 1, \dots, n^\star$. Define $n^\star$ disjoint intervals by $I_1^\star = (-\infty, Z_{(1)}^\star) \cup [Z_{(1)}^\star, Z_{(2)}^\star), I_2^\star = [Z_{(2)}^\star, Z_{(3)}^\star), \dots, I_{n^\star}^\star = [Z_{(n^\star)}^\star, +\infty)$, where $Z_{(i)}^\star$ is the ith smallest value among $Z_1^\star, \dots, Z_{n^\star}^\star$. We can then apply the replacement parameters algorithm that replaces $\psi(Z_h)$ with $\psi(Z_i^\star)$ if $Z_h \in I_i^\star$ for $h = 1, \dots, n$, $i = 1, \dots, n^\star$. Accordingly, the loss function for censored data is defined by

$$l^C(\psi^\star) = \frac{1}{2}\psi^{\star T} H^\star \psi^\star - \psi^{\star T} q^\star,$$

where $\psi^\star = \{\psi(Z_{(1)}^\star), \dots, \psi(Z_{(n^\star)}^\star)\}^T$, $h_{ij}^\star = \sum_{s \in R_i} \sum_{t \in R_j} h_{st}$, $q_i^\star = \sum_{s \in R_i} q_s$ and $R_i = \{s : Z_s \in I_i^\star, s = 1, \dots, n\}$. We assume that $\psi_j = \psi_1^\star$ if $Z_j < Z_{(1)}^\star$, $j = 1, \dots, n$, for censored subjects whose covariate value is smaller then the smallest value for uncensored subjects. This is needed to estimate $\psi(\cdot)$ at all values of Z including $[Z_{(1)}, Z_{(1)}^\star)$. As stated in Proposition 4.5, $l^C(\psi^\star)$ is a strictly convex quadratic function when an anchor constraint is imposed. Thus, Theorems 4.2–4.4 are all valid for $l^C(\psi^\star)$, so that either the profiling quadratic programming method or profiling quadratic pool-adjacent-violators algorithm is directly applicable to $l^C(\psi^\star)$.

Proposition 4.5. *Assume that* $\psi_j = \psi_1^\star$ *if* $Z_j < Z_{(1)}^\star$, $j = 1, \ldots, n$. *The loss function* $l^C(\psi^\star)$ *is strictly convex when an anchor constraint is imposed that* $\psi_m^\star = \psi(Z_{(m)}^\star) = 0$.

4.2.5 Time-dependent covariates

We consider a time-dependent covariate $Z(t)$ in the additive hazard model of $\lambda(t \mid Z(t), \psi) = \lambda_0(t) + \psi(Z(t))$. We assume that $\psi(\cdot)$ does not change over time. Under the proportional hazard assumption, the isotonic estimator jumps only at the time-dependent covariate values associated with uncensored subjects only at their failure times. Since our approach was based on a mimicked score function from the proportional hazard model in Subsection 4.2.1, we also restricted our unimodal estimator to a right continuous step function with jumps at the time-dependent covariate values at their observed failure times. Based on this restriction, we apply the replacement parameters algorithm where the parameters for subjects having their failure times observed are substituted for other parameters in the loss function. This approach is formalized below.

A challenge is that the number of parameters in the loss function is potentially much larger than the total sample size because the time-varying covariate may result in different $\psi(Z_i(X_j))$ at each observed failure time for a given subject $i = 1, \ldots, n$. To alleviate this problem, we assume that the unimodal estimator is a right continuous step function which jumps only at $Z_i(X_i^\star)$, where $X_i^\star$ is the ith subject's failure time. Based on this assumption, one may use replacement parameters algorithm where the parameters for subjects having their failure times observed are substituted for other parameters in the loss function.

Formally, let $n^\star$ be the number of subjects having observed failure time, and let $Z_i^\star(t)$ be their covariates for $i = 1, \ldots, n^\star$. Let $Z_i^\star = Z_i^\star(X_i^\star)$ be the ith subject's covariate at time of failure. Define $n^\star$ disjoint intervals by $I_1^* = (-\infty, Z_{(1)}^*) \cup [Z_{(1)}^*, Z_{(2)}^*), I_2^* = [Z_{(2)}^*, Z_{(3)}^*), \ldots, I_{n^\star}^* = [Z_{(n^\star)}^*, +\infty)$, where $Z_{(i)}^*$ is the ith smallest value among $Z_1^*, \ldots, Z_{n^\star}^*$. We can then apply the replacement parameters algorithm that replaces $\psi(Z_h(X_j))$ with $\psi(Z_i^*)$ if $Z_h(X_j) \in I_i$ for $h, j = 1, \ldots, n$ and $i = 1, \ldots, n^\star$. Accordingly, the loss function is then defined as

$$l^D(\psi^*) = \frac{1}{2}\psi^{*T} H^* \psi^* - \psi^{*T} q^*,$$

where $h_{ij}^* = \int_0^\infty \{\sum_{s \in R_i(u)} \sum_{t \in R_j(u)} h_{st}(u)\} dt$, $q_i^* = \int_0^\infty \{\sum_{s \in R_i(u)} q_s(u)\}$, $R_i(u) = \{s : Z_s(u) \in I_i^*, s = 1, \ldots, n\}$ and $\psi^* = \{\psi(Z_{(1)}^*), \ldots, \psi(Z_{(n^\star)}^*)\}^T$. Here, $h_{st}(u)$ and $q_s(u)$ are the quantities inside the integrals in (4.6) and (4.7) respectively, where $h_{st}(u) = Y_s(u) I(s = t) - \{Y_s(u) Y_t(u)\} / \{\sum_{l=1}^n Y_l(u)\}$ and $q_s(u) = dN_s(u) - Y_s(u)\{\sum_{v=1}^n dN_v(u)\} / \{\sum_{l=1}^n Y_l(u)\}$ for $s, t = 1, \ldots, n$. Since Z_i^*'s are only defined for those who have observed failure time, $l^D(\psi^*)$ with replacement parameters algorithm is also applicable for censored data with the time-dependent covariate. However, unlike the censoring case with a time independent covariate that replaces parameters for censored subjects only, both the uncensored and cenceored cases with the time-dependent covariate require replacing, which may lead to certain parameters vanishing from the loss function. To ensure valid estimations, we further assume that the unimodal estimator does not have jumps at a covariate value associated with the excluded parameters. Then, Theorems 4.2–4.4 are all valid for $l^D(\psi^*)$, so that either the profiling quadratic programming method or profiling quadratic pool-adjacent-violators algorithm is directly applicable to $l^D(\psi^*)$.

Proposition 4.6. *Assume that $\psi(Z_i(X_j)) = \psi_1^*$ if $Z_i(X_j) < Z_{(1)}^\star$, $i, j = 1, \ldots, n$. The loss function of $l^D(\psi^*)$ is strictly convex when an anchor constraint is imposed that $\psi_m^* = \psi(Z_{(m)}^*) = 0$.*

4.3 Extension

4.3.1 Monotone and U-shape hazard functions

The unimodal shape restricted hazard model is allowed to include a monotone hazard function by setting the mode M to the left or right boundary of the covariate values. In other words, the unimodal function $\psi(\cdot)$ is changed to a monotone increasing (or monotone decreasing) function if we set M to $Z_{(n)}$ (or $Z_{(1)}$). A U-shape hazard function is accommodated by reversing the order of the unimodal constraint before and after the mode. Since the monotone or U-shape constraint is expressed as a convex cone, the quadratic pool-adjacent-violators algorithm is applicable to compute the constrained estimator.

4.3.2 Baseline hazard function

It is not possible to estimate the baseline hazard function $\lambda_0(t)$ and vertical shift parameter δ. They are not identifiable because $\{\lambda_0(t), \phi(Z)\}$ and $\{\lambda_0(t) + \delta, \phi(Z) - \delta\}$ give the same model in (4.1). Under the anchor constraint $\psi(Z) = \phi(Z) - \delta$ in Subsection 4.2.2, one may estimate $\lambda_0^\star(t)$, where $\lambda_0^\star(t) = \lambda_0(t) + \delta$ is a baseline hazard function including an anchor effect. This approach is needed for the the standard additive hazard model to estimate a baseline hazard function at a reference group. Hence, the baseline hazard function including the anchor effect can be computed by plugging $\hat{\psi}$ into $\hat{\Lambda}_0(\psi, t)$ in (4.4), where $\hat{\psi}$ is the constrained estimator from the quadratic loss function.

4.3.3 Additional covariates

Suppose that there exist an additional p covariates. We include those covariates in the model $\lambda(t \mid Z, W(t)) = \lambda_0(t) + \psi(Z) + \beta^T W(t)$, where $W(\cdot)$ is a $p \times 1$ dimensional covariate process and β is a $p \times 1$ vector of regression parameters. Taking the same approach as in Subsection 4.2.1, we mimic the score function from the isotonic proportional hazard model with additional covariates. We define the quadratic loss function as

$$l^A(\theta) = \frac{1}{2}\theta^T H^A \theta - \theta^T q^A,$$

where $\theta = (\psi^T, \beta^T)^T$, $q^A = (q^T, q^{\circ T})^T$ and

$$H^A = \begin{bmatrix} H & H^{\Diamond} \\ (H^{\Diamond})^T & H^{\circ} \end{bmatrix},$$

where H° and q° are an $p \times p$ matrix and an $p \times 1$ vector, respectively, defined as

$$\begin{aligned} H^\circ &= \sum_{i=1}^n \int_0^\infty Y_i(t)\big\{W_i(t) - \bar{W}(t)\big\}^{\otimes 2} dt, \\ q^\circ &= \sum_{i=1}^n \int_0^\infty \big\{W_i(t) - \bar{W}(t)\big\} dN_i(t), \end{aligned}$$

and $H^{\Diamond}$ is an $n \times p$ matrix with elements

$$h_{ij}^{\Diamond} = \int_0^\infty Y_i(t)\big\{W_{ij}(t) - \frac{\sum_{s=1}^n Y_s(t) W_{sj}(t)}{\sum_{l=1}^n Y_l(t)}\big\} dt$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$. Here, $\bar{W}(t) = \sum_{j=1}^n Y_j(t) W_j(t) / \sum_{l=1}^n Y_l(t)$ and $W^{\otimes 2} = W^T W$. Detailed derivations are available in Section 4.6.

For fixed β, $l^A(\theta)$ reduces to the quadratic function $l^A(\psi \mid \beta) = \psi^T H \psi / 2 - \psi^T q^{\Diamond}(\beta)$, where $q^{\Diamond}(\beta) = q - H^{\diamond}\beta$. Similarly, for fixed ψ, $l^A(\theta)$ reduces to the quadratic function $l^A(\beta \mid \psi) = \beta^T H^\circ \beta / 2 - \beta^T q^{\Box}(\psi)$, where $q^{\Box}(\psi) = q^\circ - (H^{\diamond})^T \psi$. Thus, we can estimate ψ and β by the following steps. Set initial values of $(\psi^{(0)}, \beta^{(0)}) \in \Psi^m \times \mathbb{R}^p$. Update $\psi^{(m)}$ given $\beta = \beta^{(m-1)}$ by using the quadratic pool-adjacent-violators algorithm for $l^A(\psi \mid \beta)$, and update $\beta^{(m)}$ given $\psi = \psi^{(m)}$ which has the closed form solution $\beta^{(m)} = H^{\circ -1} q^{\Box}(\psi^{(m)})$. Repeat the updates until convergence, where the convergence criteria is $d(\psi^{(m)}, \psi^{(m-1)}) + d(\beta^{(m)}, \beta^{(m-1)}) < \epsilon$ for a small positive ϵ. Since $l^A(\psi \mid \beta)$

has the same form as the quadratic loss function in (4.8), the replacement parameters algorithm described in Subsections 4.2.4 and 4.2.5 is directly applicable to censored data with time independent covariate and time-dependent covariate, respectively.

4.4 Simulations

We investigated the performance of quadratic programming method and quadratic pool-adjacent-violators algorithm through simulation studies. As a gold standard, one-step quadratic pool-adjacent-violators algorithm was considered by setting ν to the true value in the pseudo loss function. For the first part of the simulation studies, we considered a time independent covariate Z. The time independent covariate was generated from a uniform distribution on (0,1). Three forms of unimodal functions on the interval (0,1) were considered: $\phi(Z) = -|Z - M|$, $\phi(Z) = -|Z - M|^{1/2}$ and $\phi(Z) = -|Z - M|^2$, where M=0·25, 0·50 and 0·75. The failure time was then generated from an additive hazards model with a constant baseline hazard function. The censoring time was independently generated from a uniform distribution yielding approximately 30% censoring. For the second set of simulation studies, the same scenarios were used with a time-dependent covariate $Z(t)$, where $Z(t)$ was piecewise constant. We constructed $Z(t)$ by generating independent uniform (0,1) random variables on disjoint time intervals $(x_{j-1}, x_j]$, where $x_0 = 0, x_1 = 0{\cdot}22, x_2 = 0{\cdot}44, \ldots, x_9 = 2, x_{10} = +\infty$. We simulated 500 replicates with sample size n = 100, 500 and 1000. To demonstrate Theorem 4.4, we additionally checked Fenchel's duality conditions in Theorem 4.2 for the converged value of the quadratic pool-adjacent-violators algorithm: $max_{j\in(1,\ldots,m-1)}\{\sum_{j=1}^{i} u_j(\ddot{\psi})\} < \epsilon$; $\max_{j\in(m+1,\ldots,n)}\{\sum_{j=i}^{n} u_j(\ddot{\psi})\} < \epsilon$; $|\sum_{j=m+1}^{n} u_j(\ddot{\psi})| < \epsilon$; $|\sum_{i=1,i\neq m}^{n} \ddot{\psi}_i u_i(\ddot{\psi})| < \epsilon$. The stopping values of ϵ and $\dot{\epsilon}$ were set to 10^{-3} and 10^{-5}, respectively. For each data set an initial value for the quadratic pool-adjacent-violators algorithm was set to $|\hat{\gamma}|\bar{Z}_i I(i \leq m) - |\hat{\gamma}|\bar{Z}_i I(i \geq m)$, where $\bar{Z}_i = Z_{(i)} - Z_{(m)}$ and $\hat{\gamma}$ was estimated coefficient of $\bar{Z}_i$ from the standard additive hazard models.

First we conducted the simulations with known mode M and evaluated the performance of the algorithms by computing the integrated mean square error $\int_0^1 E[\phi^\circ(Z) - \hat{\psi}^\circ(Z)]^2 dZ$, where $\phi^\circ(Z) = \phi(Z) - \phi(0)$ and $\hat{\psi}^\circ(Z) = \hat{\psi}(Z) - \hat{\psi}(0)$. This was approximated by $\sum_{r=1}^{R} \sum_{g=1}^{G} [\phi^\circ(z_g) - \hat{\psi}_r^\circ(z_g)]^2/(GR)$ based on equally spaced grid points of z_g's between (0·001, 0·999), where $G = 1000$ grid points and $R = 500$ simulation runs. Here, points in $M \pm 0{\cdot}005$ were excluded from the grid points to compute integrated mean square error because the unimodal estimator might be unstable around the mode (Robertson et al. 1988, p.326). Second we conducted the simulations with unknown mode M using the profiling algorithm. We then additionally evaluated bias and mean squared error, $\sum_{r=1}^{R}(\hat{M}_r - M)/R$ and $\sum_{r=1}^{R}(\hat{M}_r - M)^2/R$, respectively, where $\hat{M}_r$ was the estimated mode for rth simulated dataset. Likewise, $M \pm 0{\cdot}005$ points from the unimodal estimators were excluded to compare the loss functions during the profiling algorithm in (4.12).

Tables 4.1 and 4.2 show simulations results for time independent and time-dependent covariates, respectively. For known mode, integrated mean square error decreases as sample size increases for both quadratic programming method and quadratic pool-adjacent-violators algorithm with reasonable computation speeds. The quadratic pool-adjacent-violators algorithm converges 100% for all scenarios, which is in agreement with Theorem 4.4. For unknown mode, both quadratic programming and quadratic pool-adjacent-violators algorithms have relatively large integrated mean square error, bias and mean square error in small sample size, but these decreases as sample size increases. The quadratic programming method is quite slow for large sample sizes. For example, when sample size is 1000, it takes approximately 5 seconds for known mode and $1000 \times 5 = 5000$ seconds (83 minutes) for unknown mode, since the quadratic programming method is performed at every hypothetical mode. On the other hand, the quadratic pool-adjacent-violators algorithm dramatically improves computational speed. When sample size is

1000, it takes less than one minute for all scenario and is orders of magnitude faster than the quadratic programming method. Here the computation time does not include time for calculating Hessian matrix in order to compare the algorithms. Both quadratic programming method and quadratic pool-adjacent-violators algorithm have similar performances as the gold standard one-step quadratic pool-adjacent-violators algorithm in terms of integrated mean square error of the unimodal estimator, bias and mean square error of the mode estimator.

4.5 Folic acid for vascular outcome reduction in transplantation study

Folic Acid for Vascular Outcome Reduction in Transplantation (FAVORIT) study was a multicenter double-blind randomized controlled clinical trial to investigate if vitamin supplementation reduces risk of cardiovascular disease (CVD) in kidney transplant recipients (Bostom et al. 2011). Four thousand one hundred ten study participants were enrolled between August 2002 and January 2007 and followed up every six months thorough January 2010. Patients were randomized to a multivitamin that included either a high-dose or low-dose of folic acid (5 or 0 mg), vitamin B6 (50 or 1.4 mg), and vitamin B12 (1000 or 2 microg). The outcome of interest was any of the following nine events: (1) CVD death, (2) myocardial infarction, (3) resuscitated sudden death, (4) stroke, (5) coronary artery revascularization, (6) lower extremity revascularization or amputation above the ankle for severe arterial disease, (7) carotid endarterectomy or angioplasty, (8) abdominal aortic aneurysm repair, or (9) renal artery revascularization. A total of 584 CVD events were observed. Diastolic and systolic blood pressure was measured at baseline on all participants. The mean arterial pressure (Gevers et al. 1993) was then calculated as a weighted average of diastolic and systolic blood pressure with a weight of one third and two thirds, respectively. The mean arterial pressure ranged from 55·0 to 167·8 with the sample mean of 97·6±12·9 $mmHg$.

Table 4.1: Simulation results for time independent covariate: IMSE multiplied by 10^5 (CPU time in seconds), bias multiplied by 10^3 and MSE multiplied by 10^3 for known and unknown modes, where $\phi_1 = -|Z - M|$, $\phi_2 = -|Z - M|^{1/2}$ and $\phi_3 = -|Z - M|^2$.

			QPM			QPAVA			QPAVA True		
M	$\phi(Z)$	n	Known	Unknown		Known	Unknown		Known	Unknown	
			IMSE	IMSE	B/M	IMSE	IMSE	B/M	IMSE	IMSE	B/M
0·25	ϕ_1	100	79 (0)	230 (0)	125/77	82 (0)	230 (0)	125/77	80 (0)	224 (0)	125/77
		500	16 (1)	55 (225)	14/20	15 (0)	55 (7)	14/20	15 (0)	55 (2)	14/19
		1000	12 (5)	18 (3648)	7/12	12 (0)	18 (47)	7/12	11 (0)	18 (19)	7/12
	ϕ_2	100	84 (0)	311 (0)	183/105	87 (0)	311 (0)	183/105	85 (0)	303 (0)	184/106
		500	23 (1)	64 (224)	77/47	21 (0)	64 (6)	77/47	21 (0)	64 (2)	78/47
		1000	17 (6)	26 (3684)	55/31	17 (0)	26 (44)	55/31	17 (0)	26 (19)	55/31
	ϕ_3	100	55 (0)	129 (0)	86/56	58 (0)	129 (0)	86/56	56 (0)	126 (0)	86/56
		500	10 (1)	29 (220)	0/9	9 (0)	29 (7)	0/9	8 (0)	29 (2)	-1/9
		1000	6 (5)	15 (3620)	-2/4	6 (0)	15 (49)	-2/4	6 (0)	15 (19)	-2/4
0·5	ϕ_1	100	40 (0)	216 (0)	6/64	42 (0)	216 (0)	6/64	40 (0)	209 (0)	6/64
		500	10 (1)	42 (222)	-4/28	10 (0)	42 (6)	-4/28	10 (0)	41 (2)	-5/28
		1000	7 (5)	13 (3663)	-5/13	8 (0)	13 (44)	-5/13	7 (0)	13 (19)	-5/13
	ϕ_2	100	51 (0)	350 (0)	12/74	53 (0)	350 (0)	12/74	52 (0)	341 (0)	12/74
		500	16 (1)	64 (218)	-12/50	16 (0)	64 (6)	-12/50	16 (0)	63 (2)	-12/50
		1000	13 (5)	23 (3574)	-3/37	13 (0)	23 (41)	-3/37	13 (0)	23 (18)	-4/37
	ϕ_3	100	27 (0)	99 (0)	4/53	28 (0)	99 (0)	4/53	26 (0)	96 (0)	4/53
		500	6 (1)	20 (220)	-1/11	6 (0)	20 (6)	-1/11	5 (0)	20 (2)	-1/12
		1000	3 (5)	7 (3610)	0/5	4 (0)	7 (46)	0/5	4 (0)	7 (19)	-1/5
0·75	ϕ_1	100	30 (0)	220 (0)	-127/74	34 (0)	220 (0)	-127/74	32 (0)	213 (0)	-127/74
		500	7 (1)	38 (219)	-29/23	10 (0)	38 (7)	-29/23	9 (0)	37 (2)	-28/23
		1000	3 (5)	9 (3623)	-11/12	3 (0)	9 (47)	-11/12	3 (0)	8 (19)	-10/12
	ϕ_2	100	40 (0)	328 (0)	-161/95	38 (0)	328 (0)	-161/95	36 (0)	319 (0)	-162/95
		500	12 (1)	52 (212)	-80/45	13 (0)	52 (6)	-80/45	13 (0)	51 (2)	-80/45
		1000	6 (5)	15 (3503)	-46/28	6 (0)	15 (43)	-46/28	6 (0)	15 (18)	-47/28
	ϕ_3	100	22 (0)	115 (0)	-93/55	23 (0)	115 (0)	-93/55	22 (0)	112 (0)	-93/55
		500	6 (1)	21 (181)	-3/10	7 (0)	21 (6)	-3/10	7 (0)	20 (2)	-4/10
		1000	2 (5)	6 (3007)	0/5	2 (0)	6 (43)	0/5	2 (0)	5 (17)	0/5

QPM: quadratic programming method; QPAVA: quadratic pool-adjacent-violators algorithm; QPAVA True: one-step QPAVA from true initial value; Known: mode is known; Unknown: mode is unknown; IMSE: integrated mean squared error; B/M: bias/mean squared error.

Table 4.2: Simulation results for time-dependent covariate: IMSE multiplied by 10^5 (CPU time in seconds), bias multiplied by 10^3 and MSE multiplied by 10^3 for known and unknown modes, where $\phi_1 = -|Z(t) - M|$, $\phi_2 = -|Z(t) - M|^{1/2}$ and $\phi_3 = -|Z(t) - M|^2$.

			QPM			QPAVA			QPAVA True		
M	$\phi(Z(t))$	n	Known	Unknown		Known	Unknown		Known	Unknown	
			IMSE	IMSE	B/M	IMSE	IMSE	B/M	IMSE	IMSE	B/M
0·25	ϕ_1	100	21 (0)	55 (1)	104/57	21 (0)	55 (0)	103/57	22 (0)	56 (0)	103/57
		500	15 (1)	26 (244)	15/18	15 (0)	26 (7)	15/18	16 (0)	26 (2)	15/18
		1000	14 (5)	15 (4077)	18/11	14 (0)	15 (46)	18/11	14 (0)	16 (20)	18/11
	ϕ_2	100	33 (0)	60 (1)	151/85	33 (0)	60 (0)	151/85	35 (0)	62 (0)	152/85
		500	23 (1)	39 (331)	71/43	23 (0)	39 (8)	71/43	23 (0)	39 (3)	71/43
		1000	23 (5)	24 (5394)	42/28	23 (0)	24 (55)	42/28	23 (0)	24 (24)	41/28
	ϕ_3	100	11 (0)	21 (0)	47/35	11 (0)	21 (0)	47/35	12 (0)	22 (0)	47/35
		500	7 (1)	15 (133)	0/10	7 (0)	15 (5)	0/10	7 (0)	15 (2)	0/10
		1000	6 (5)	8 (2241)	-3/4	6 (0)	8 (30)	-3/4	6 (0)	8 (13)	-3/4
0·5	ϕ_1	100	15 (0)	38 (1)	-4/52	14 (0)	38 (0)	-4/52	15 (0)	39 (0)	-4/52
		500	8 (1)	16 (280)	9/20	8 (0)	16 (7)	9/20	8 (0)	16 (3)	8/20
		1000	7 (5)	9 (4616)	3/13	7 (0)	9 (50)	3/13	7 (0)	9 (22)	3/13
	ϕ_2	100	25 (0)	62 (1)	6/68	25 (0)	62 (0)	6/68	27 (0)	63 (0)	6/68
		500	16 (1)	29 (361)	-11/46	16 (0)	29 (8)	-11/46	16 (0)	30 (3)	-11/46
		1000	15 (5)	18 (5877)	-5/39	15 (0)	18 (58)	-5/39	15 (0)	18 (26)	-5/38
	ϕ_3	100	8 (0)	19 (0)	-19/35	8 (0)	19 (0)	-19/35	8 (0)	19 (0)	-18/35
		500	3 (1)	7 (160)	2/9	3 (0)	7 (5)	2/9	3 (0)	7 (2)	2/9
		1000	3 (5)	4 (2688)	-7/4	3 (0)	4 (34)	-7/4	3 (0)	4 (15)	-7/4
0·75	ϕ_1	100	9 (0)	27 (0)	-107/59	10 (0)	27 (0)	-107/59	10 (0)	28 (0)	-107/59
		500	3 (1)	13 (243)	-31/20	3 (0)	13 (7)	-31/20	3 (0)	13 (2)	-31/20
		1000	3 (5)	4 (4138)	-8/11	3 (0)	4 (46)	-8/11	3 (0)	4 (20)	-9/11
	ϕ_2	100	15 (0)	47 (1)	-154/90	15 (0)	47 (0)	-154/90	16 (0)	48 (0)	-154/90
		500	7 (1)	18 (333)	-81/43	7 (0)	18 (8)	-81/43	7 (0)	18 (3)	-81/43
		1000	6 (5)	8 (5454)	-41/29	6 (0)	8 (55)	-41/29	6 (0)	8 (24)	-41/29
	ϕ_3	100	6 (0)	12 (0)	-61/36	6 (0)	12 (0)	-61/36	7 (0)	12 (0)	-61/36
		500	2 (1)	9 (136)	-12/11	2 (0)	9 (5)	-12/11	2 (0)	9 (2)	-12/11
		1000	1 (5)	2 (2208)	4/5	1 (0)	2 (29)	4/5	1 (0)	2 (13)	4/5

QPM: quadratic programming method; QPAVA: quadratic pool-adjacent-violators algorithm; QPAVA True: one-step QPAVA from true initial value; Known: mode is known; Unknown: mode is unknown; IMSE: integrated mean squared error; B/M: bias/mean squared error.

The mean arterial pressure is known to have a U-shape relationship with the CVD with low and high values associated with increased risk (Berbari and Manci 2010, p.95). Table 4.3 displays the relationship between mean arterial pressure and CVD estimated from the FAVORIT data. The rate of CVD is higher when the mean arterial pressure is below the 20th percentile and when the mean pressure is above the 80th percentile. The CVD rate is lower for the value of the mean arterial pressure between the 20th and 80th percentiles. In the standard Cox model, the proportional hazards assumption is violated for the mean arterial pressure ($P = 0{\cdot}02$), and the additive hazards assumption would be an alternative.

We fit an additive hazards model assuming a U-shaped relationship between the mean arterial pressure and CVD. The profiling quadratic pool-adjacent-violators algorithm was used to estimate the mode and U-shape hazard function, with the method described in Subsection 4.3.3 to adjust for treatment effect. Figure 4.1 displays the estimated U-shape hazard function in the mean arterial pressure where location of the mode is estimated to be at 77. The black dots indicate the values of the mean arterial pressure associated with observed CVD event, which are potential jump points for the U-shape estimate, as discussed in Subsection 4.2.4. The estimated hazard function shows the U-shaped relationship, with relatively large jumps at 60 and 146 of the mean arterial pressure.

We additionally fit the standard additive hazards models with polynomials of degree 2 ($\alpha_1\bar{Z}+\alpha_2\bar{Z}^2+\alpha_3 Trt$), piecewise linear ($\beta_1\bar{Z}_{lt}+\beta_2\bar{Z}_{rt}+\beta_3 Trt$) and piecewise polynomials of degree 2 ($\gamma_1\bar{Z}_{lt}+\gamma_2\bar{Z}_{lt}^2+\gamma_3\bar{Z}_{rt}+\gamma_4\bar{Z}_{rt}^2+\gamma_5 Trt$) in Figure 4.1, using the estimated mode, 77, from the U-shape additive hazards model. Here, $\bar{Z}$ is the mean arterial pressure centered at 77, $\bar{Z}_{lt} = \bar{Z}I(\bar{Z} \leq 0)$, $\bar{Z}_{rt} = \bar{Z}I(\bar{Z} > 0)$ and Trt is a treatment group indicator with a reference group of the low folic acid. The parametric polynomials do not provide a good fit particularly for the increased risk at the lower mean arterial pressure, except maybe for the piecewise 2nd degree polynomials. However, the piecewise 2nd degree

Table 4.3: Rate of CVD events by quantiles.

Quintile (%)	Range of MAP	The number of patients	Patient year at risk	The number of CVD events	Rate (per 1000)
0-20	55-87	808	3038	125	41
20-40	87-94	802	3178	103	32
40-60	94-100	810	3171	119	38
60-80	100-107	802	3062	114	37
80-100	107-168	799	2736	121	44

MAP: mean arterial pressure.

polynomials result in a W-shape rather than a U-shape. In particular, the risk increases sharply between 70 and 80 of the mean arterial pressure, which is not supported by the data. Thus, polynomials appear not to be well suited to capture the nonlinear effect of the mean arterial pressure. Furthermore, such models might be hard to interpret ($\hat{\alpha}_1 = -4 \times 10^{-4}$, $P = 0{\cdot}20$; $\hat{\alpha}_2 = 1 \times 10^{-5}$, $P = 0{\cdot}06$; $\hat{\beta}_1 = -7 \times 10^{-4}$, $P = 0{\cdot}70$; $\hat{\beta}_2 = 2 \times 10^{-4}$, $P = 0{\cdot}23$; $\hat{\gamma}_1 = 7 \times 10^{-2}$, $P = 0{\cdot}06$; $\hat{\gamma}_2 = 1 \times 10^{-3}$, $P = 0{\cdot}10$; $\hat{\gamma}_3 = -8 \times 10^{-4}$, $P = 0{\cdot}06$; $\hat{\gamma}_4 = 2 \times 10^{-5}$, $P = 0{\cdot}02$). Note that one needs to know the value of the mode to have a good parametric estimation. We used the estimate of the mode obtained from fitting the U-shape additive hazard model. It is not clear how to specify the value of the mode for a parametric model if a preliminary estimate of the mode is not available. The treatment effect was not significant in all models ($\hat{\alpha}_3 = 7 \times 10^{-3}$, $P = 0{\cdot}82$; $\hat{\beta}_3 = 5 \times 10^{-4}$, $P = 0{\cdot}85$; $\hat{\gamma}_5 = 5 \times 10^{-4}$, $P = 0{\cdot}87$; 3×10^{-3} for the U-shape additive hazards model).

4.6 Technical Details for Chapter 4

Proof of Theorem 4.1 It is obvious that the loss function is convex because it is a quadratic function. Under the anchor constraint $\psi_m = \psi(Z_{(m)}) = 0$, we prove the strict convexity by showing $x^T Hx \geq 0$ for any $x \neq 0 \in \mathbb{R}^{n-1}$. Since $h_{ij} \leq 0$ for $i, j = 1, \ldots, n$ $(i \neq$

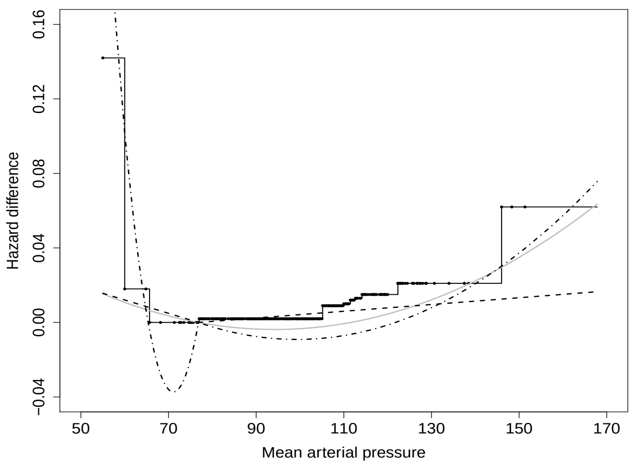

Figure 4.1: FAVORIT study: U-shape (black solid) versus standard additive hazards models with polynomials of degree 2 (grey solid), piecewise linear (dashed), piecewise polynomials of degree 2 (dot-dashed). The black dots indicate CVD events.

$j, i \neq m, j \neq m$),

$$\begin{aligned}
x^T Hx &= \sum_{i=1,i\neq m}^{n} h_{ii}x_i^2 + 2\sum\sum_{i<j,i\neq m,j\neq m} h_{ij}x_ix_j \\
&= \sum_{i=1,i\neq m}^{n} h_{ii}x_i^2 + \sum\sum_{i<j,i\neq m,j\neq m} \left\{(-h_{ij})^{\frac{1}{2}}x_i - (-h_{ij})^{\frac{1}{2}}x_j\right\}^2 + \sum\sum_{i<j,i\neq m,j\neq m} (h_{ij}x_i^2 + h_{ij}x_j^2) \\
&= \sum\sum_{i<j,i\neq m,j\neq m} \left\{(-h_{ij})^{\frac{1}{2}}x_i - (-h_{ij})^{\frac{1}{2}}x_j\right\}^2 + \sum_{i=1,i\neq m}^{n} \left\{h_{ii}x_i^2 + \sum_{i<j,i\neq m,j\neq m} (h_{ij}x_i^2 + h_{ij}x_j^2)\right\} \\
&= \sum\sum_{i<j,i\neq m,j\neq m} \left\{(-h_{ij})^{\frac{1}{2}}x_i - (-h_{ij})^{\frac{1}{2}}x_j\right\}^2 + \sum_{i=1,i\neq m}^{n} \left\{h_{ii} + \sum_{j=1,j\neq i}^{n} h_{ij}\right\}x_i^2.
\end{aligned}$$

The first term is greater than or equal to 0, so the strict convexity holds by showing that the second term is strictly greater than 0 because

$$\begin{aligned}
h_{ii} + \textstyle\sum_{j=1,i\neq j,j\neq m}^{n} h_{ij} &= \int_0^\infty \left\{Y_i(u) - \frac{Y_i(u)^2}{\sum_{s=1}^n Y_s(u)} - \frac{Y_i(u)\sum_{j=1,j\neq i,j\neq m}^n Y_j(u)}{\sum_{s=1}^n Y_s(u)}\right\}du \\
&= \int_0^\infty \frac{Y_i(u)Y_m(u)}{\sum_{s=1}^n Y_s(u)}du \geq \frac{1}{n}X_{(1)} > 0. \qquad (4.16)
\end{aligned}$$

for $i = 1, \ldots, n$ ($i \neq m$).

Proof of Theorem 4.2 Since $l(\psi)$ is a convex function and Ψ^m is a convex cone, Lemma 2.1 (Groeneboom 1996) is directly applicable, where $\hat{\psi}$ minimizes $l(\psi)$ over Ψ^m if and only if

$$\textstyle\sum_{i=1,i\neq m}^{n} \psi_i u_i(\hat{\psi}) \geq 0 \quad {}^{\forall}\psi \in \Psi^m, \tag{4.17}$$

$$\textstyle\sum_{i=1,i\neq m}^{n} \hat{\psi}_i u_i(\hat{\psi}) = 0. \tag{4.18}$$

Since (4.18) is the same as (4.11), we claim that (4.17) is equivalent to (4.10). Suppose that $\hat{\psi}$ satisfies (4.17). Let $\alpha_i = \psi_i - \psi_{i+1}$ for $i = 1, \ldots, m-1$ and $\alpha_i = \psi_i - \psi_{i-1}$ for $i = m+1, \ldots, n$. For any $\psi \in \Psi^m$, ith element of ψ is expressed as $\sum_{j=i}^{m-1} \alpha_j$ for $i = 1, \ldots, m-1$, or $\sum_{j=m+1}^{i} \alpha_j$ for $i = m+1, \ldots, n$. Thus,

$$\begin{aligned}
0 \leq \sum_{i=1,i\neq k}^{n} \psi_i u_i(\hat{\psi}) &= \sum_{i=1}^{m-1} \psi_i u_i(\hat{\psi}) + \sum_{i=m+1}^{n} \psi_i u_i(\hat{\psi}) \\
&= \sum_{i=1}^{m-1} \{\sum_{j=i}^{m-1} \alpha_j\} u_i(\hat{\psi}) + \sum_{i=m+1}^{n} \{\sum_{j=m+1}^{i} \alpha_j\} u_i(\hat{\psi}) \\
&= \textstyle\sum_{i=1}^{m-1} \{\sum_{j=1}^{i} u_j(\hat{\psi})\} \alpha_i + \sum_{i=m+2}^{n} \{\sum_{j=i}^{n} u_j(\hat{\psi})\} \alpha_i + \{\sum_{j=m+1}^{n} u_j(\hat{\psi})\} \alpha_{m+1},
\end{aligned} \tag{4.19}$$

which yields (4.10) because $\alpha_i \leq 0$ for $i = 1, \ldots, m-1, m+2, \ldots, n$. The other direction is trivial.

The uniqueness condition holds, since $l(\psi)$ is a strictly convex function from Theorem 4.1, where the same statement is made by Proposition 1.1 (Groeneboom and Wellner 1992).

Proof of Theorem 4.3 The proof is analogous to that of Theorem 4.2. Since $l^P(\psi \mid \nu)$ satisfies the conditions of Lemma 2.1 (Groeneboom 1996) over the convex cone Ψ, $\dot{\psi}$

minimizes $l(\psi|\nu)$ over Ψ if and only if

$$\textstyle\sum_{i=1}^{n} \psi_i u_i^P(\hat{\psi}_i \mid \nu) \geq 0 \ \ {}^{\forall}\psi \in \Psi, \tag{4.20}$$

$$\textstyle\sum_{i=1}^{n} \hat{\psi}_i u_i^P(\hat{\psi}_i \mid \nu) = 0. \tag{4.21}$$

Since (4.18) is the same as (4.15), we claim that (4.20) and (4.14) are equivalent. Suppose that $\dot{\psi}$ satisfies (4.20). Let $\alpha_i = \psi_i - \psi_{i+1}$ for $i = 1, \ldots, m-1$ and $\alpha_i = \psi_i - \psi_{i-1}$ for $i = m+2, \ldots, n$ with $\alpha_m = \psi_m$ and $\alpha_{m+1} = \psi_{m+1}$. Then

$$\begin{aligned}
0 &\leq \textstyle\sum_{i=1}^{n} \psi_i u_i^P(\hat{\psi} \mid \nu) = \sum_{i=1}^{m} \psi_i u_i^P(\hat{\psi} \mid \nu) + \sum_{i=m+1}^{n} \psi_i u_i^P(\hat{\psi} \mid \nu) \\
&= \textstyle\sum_{i=1}^{m} \big\{\sum_{j=i}^{m} \alpha_j\big\} u_i^P(\hat{\psi} \mid \nu) + \sum_{i=m+1}^{n} \big\{\sum_{j=m+1}^{i} \alpha_j\big\} u_i^P(\hat{\psi} \mid \nu) \\
&= \textstyle\sum_{i=1}^{m-1} \big\{\sum_{j=1}^{i} u_j^P(\hat{\psi} \mid \nu)\big\}\alpha_i + \big\{\sum_{j=1}^{m} u_j^P(\hat{\psi} \mid \nu)\big\}\alpha_m \\
&+ \textstyle\sum_{i=m+2}^{n} \big\{\sum_{j=i}^{n} u_j^P(\hat{\psi} \mid \nu)\big\}\alpha_i + \big\{\sum_{j=m+1}^{n} u_j^P(\hat{\psi} \mid \nu)\big\}\alpha_{m+1},
\end{aligned}$$

which yields (4.14) because $\alpha_i \leq 0$ for $i = 1, \ldots, m-1, m+2, \ldots, n$. The other direction is trivial.

The uniqueness condition holds because $l^P(\psi \mid \nu)$ is a strictly convex function.

Proof of Theorem 4.4 In the proof, we write r instead of $r(\dot{\epsilon})$ for a notational convenient. First, we show that for $i = 1, \ldots, n$ $(i \neq m)$,

$$u_i(\ddot{\psi}) = \{\textstyle\sum_{j=1}^{n} h_{ij} \ddot{\psi}_j\} - q_i = \{\sum_{j=1}^{n} h_{ij} \dot{\psi}_j^{(r)}\} - \dot{\psi}_m^{(r)} \{\sum_{j=1}^{n} h_{ij}\} - q_i = u_i(\dot{\psi}^{(r)}), \tag{4.22}$$

where the last equality holds because

$$\begin{aligned}\textstyle\sum_{j=1}^{n} h_{ij} = h_{ii} + \sum_{j=1,j\neq i}^{n} h_{ij} &= \int_0^\infty \left\{ Y_i(u) - \frac{Y_i(u)^2}{\sum_{s=1}^{n} Y_s(u)} - \frac{Y_i(u)\sum_{j=1,i\neq j}^{n} Y_j(u)}{\sum_{s=1}^{n} Y_s(u)} \right\} du \\ &= \int_0^\infty \frac{Y_i(u)}{\sum_{s=1}^{n} Y_s(u)} \left\{ \textstyle\sum_{s=1}^{n} Y_s(u) - Y_i(u) - \sum_{j=1,n\neq i}^{n} Y_j(u) \right\} du = 0. \qquad (4.23)\end{aligned}$$

Thus, by (4.22), we show that

$$\begin{aligned}&\left|u_i(\ddot{\psi}) - u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)})\right| = \left|u_i(\dot{\psi}^{(r)}) - u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)})\right| = \textstyle\sum_{j=1,j\neq i}^{n} \left| h_{ij}\{\dot{\psi}_j^{(r)} - \dot{\psi}_j^{(r-1)}\}\right| \\ &\leq \mu d(\dot{\psi}^{(r)}, \dot{\psi}^{(r)}) < \mu\dot{\epsilon} \qquad (4.24)\end{aligned}$$

for $i = 1, \ldots, n$ $(i \neq m)$, where $\mu = \max_{i\in\{1\ldots,n\}} \max_{j\in\{1,\ldots,n\}} |h_{ij}| < \infty$.

Next, since $\dot{\psi}_i^{(r)}$, which is the unique minimizer of $l^P(\psi|\dot{\psi}_i^{(r-1)})$, satisfies Fenchel's duality condition in (4.14) and (4.15) in Theorem 4.3, we establish the following inequality and equality conditions by using (4.24) in conjunction with the triangle inequality, where the first inequality in (4.14) shows

$$\sum_{j=1}^{i} u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)}) \leq 0 \quad (i = 1, \ldots, m-1),$$

which implies

$$\sum_{j=1}^{i} u_j(\ddot{\psi}) \leq \sum_{j=1}^{i} \{u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)})\} \leq |\sum_{j=1}^{i} \{u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)})\}| \leq i\mu\dot{\epsilon}, \quad (4.25)$$

the second inequality in (4.14) shows

$$\sum_{j=i}^{n} u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)}) \leq 0 \quad (i = m+1, \ldots, n),$$

which implies

$$\sum_{j=i}^{n} u_j(\ddot{\psi}) \leq \sum_{j=i}^{n}\{u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)})\} \leq |\sum_{j=i}^{n}\{u_j(\ddot{\psi}) - u_j^P(\dot{\psi}_j^{(r)} \mid \dot{\psi}_j^{(r-1)})\}| \leq (n-m+1)\mu\dot{\epsilon}, \tag{4.26}$$

and the equality in (4.15) shows

$$\begin{aligned}0 &= \sum_{i=1}^{n} \dot{\psi}_i^{(r)} u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) = \sum_{i=1}^{n}(\dot{\psi}_i^{(r)} - \dot{\psi}_m^{(r)}) u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) + \dot{\psi}_m^{(r)} \sum_{i=1}^{n} u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) \\ &= \sum_{i=1, i\neq m}^{n} \ddot{\psi}_i u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) + \dot{\psi}_m^{(r)} \left\{ \sum_{i=1}^{m} u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) + \sum_{i=m+1}^{n} u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) \right\} \\ &= \sum_{i=1, i\neq m}^{n} \ddot{\psi}_i u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)})\end{aligned}$$

which implies

$$\left| \sum_{i=1, i\neq m}^{n} \ddot{\psi}_i u_i(\ddot{\psi}) \right| = \left| \sum_{i=1, i\neq m}^{n} \ddot{\psi}_i \{ u_i(\ddot{\psi}) - u_i^P(\dot{\psi}_i^{(r)} \mid \dot{\psi}^{(r-1)}) \} \right| \leq \sum_{i=1, i\neq m}^{n} |\ddot{\psi}_i| \mu\dot{\epsilon}. \tag{4.27}$$

By (4.24), we choose $n\mu\dot{\epsilon}$ as an upper bounce for (4.25) and (4.26) and $\pm\{\sum_{i=1, i\neq k}^{n} \ddot{\psi}_i\}\mu\dot{\epsilon}$ as upper and lower bounces for (4.27).

Finally, suppose that the quadratic pool-adjacent-violators algorithm converges under the stopping value of $\dot{\epsilon} > 0$. As $\dot{\epsilon}$ converges to zero, all the bounds in (4.25) - (4.27) converge to zero for each fixed n, which yield Fenchel's duality condition in (4.10) and (4.11). Thus, by Theorem 4.2, $\ddot{\psi}$ converges to the unique minimizer of $l^N(\psi)$ over Ψ^m.

Proof of Proposition 4.5 By the assumption, R_i is not an empty set for $i = 1, \ldots, n^\star (i \neq m)$. Since $l^C(\psi^\star)$ has the same form as $l(\psi)$, we can directly follow the proof in Theorem

4.1 by showing (4.16), which is

$$\begin{aligned}
h^\star_{ii} + \textstyle\sum_{j=1,i\neq j,j\neq m}^{n^\star} h^\star_{ij} &= \textstyle\sum_{s\in R_i}\big\{h_{ss} + \sum_{t\in R_i,t\neq s} h_{st} + \sum_{j=1,j\neq i,j\neq m}^{n^\star}\sum_{t\in R_j} h_{st}\big\}\\
&= \textstyle\sum_{s\in R_i}\big\{h_{ss} + \sum_{t\in R_i,t\neq s} h_{st} + \sum_{t\in\cup_{j=1,j\neq i,j\neq m}^{n^\star}R_j} h_{st}\big\}\\
&= \textstyle\sum_{s\in R_i}\big\{h_{ss} + \sum_{t=1,t\neq s,t\notin R_m}^{n} h_{st}\big\} > 0 \qquad (4.28)
\end{aligned}$$

for $i = 1,\dots,n^\star$ $(i \neq m)$. The last equality holds because R_i's are mutually exclusive set with $\cup_{j=1}^{n^\star} R_j = \{1,\dots,n\}$, and the last inequality holds by (4.23) with $h_{st} < 0$ for $t = 1,\dots,n$ $(s \neq t)$. It is easy to check that $h^\star_{ij} < 0$ for $i,j = 1,\dots,n$ $(i \neq j, i \neq m, j \neq m)$, since off diagonal elements of h_{ij} in (4.6) is strictly negative. Thus, $l^C(\psi^\star)$ is a strictly convex function under an anchor constraint.

Proof of Proposition 4.6 Assume that there exists at least one time point of $\dot{X}_i$ among $X_1^*,\dots,X_{n^\star}^*$ such that $R_i(\dot{X}_i) \neq \varnothing$ nor $R_m(\dot{X}_i) \neq \varnothing$ for $i = 1,\dots,n^\star$ $(i \neq m)$. Since $l^D(\psi^\star)$ has the same form as $l(\psi)$, we can directly follow the proof in Theorem 4.1 by showing (4.16), which is

$$\begin{aligned}
h^*_{ii} + \sum_{j=1,i\neq j,j\neq m}^{n^\star} h^*_{ij} &= \int_0^\infty \sum_{s\in R_i(u)}\big\{h_{ss}(u) + \sum_{t\in R_i(u),t\neq s} h_{st}(u) + \sum_{j=1,j\neq i,j\neq m}^{n^\star}\sum_{t\in R_j(u)} h_{st}(u)\big\}du\\
&= \int_0^\infty \textstyle\sum_{s\in R_i(u)}\big\{h_{ss}(u) + \sum_{t\in R_i(u),t\neq s} h_{st}(u) + \sum_{t\in\cup_{j=1,j\neq i,j\neq m}^{n^\star}R_j(u)} h_{st}(u)\big\}du\\
&= \int_0^\infty \textstyle\sum_{s\in R_i(u)}\big\{h_{ss}(u) + \sum_{t\in\cup_{j=1,j\neq m}^{n}R_j(u),t\neq s} h_{st}(u)\big\}du\\
&= \int_0^\infty \textstyle\sum_{s\in R_i(u)}\left\{Y_s(u) - \dfrac{Y_s(u)^2}{\sum_{l=1}^n Y_l(u)} - \dfrac{Y_s(u)\sum_{t\in\cup_{j=1,j\neq m}^{n^\star}R_j(u),t\neq s} Y_t(u)}{\sum_{l=1}^n Y_l(u)}\right\}du\\
&= \int_0^\infty \textstyle\sum_{s\in R_i(u)}\left\{\dfrac{Y_s(u)\sum_{t\in\{1,\dots,n\}\setminus\{\cup_{j=1,j\neq m}^{n^\star}R_j(u)\}} Y_t(u)}{\sum_{l=1}^n Y_l(u)}\right\}du\\
&\geq \frac{\{\sum_{s\in R_i(\dot{X}_i} Y_s(\dot{X}_i)\}\{\sum_{t\in R_m(\dot{X}_i)} Y_t(\dot{X}_i)\}}{\sum_{l=1}^n Y_l(\dot{X}_i)}(\dot{X}_i - \dot{X}_{i-1}) > 0 \qquad (4.29)
\end{aligned}$$

for $i = 1, \ldots, n^\star$ $(i \neq m)$, where $\setminus$ is set difference and $X_0 = 0$. The third equality holds because $R_j(u)$'s are mutually exclusive sets for $u > 0$. It is easy to check that $h^\star_{ij} < 0$ for $i, j = 1, \ldots, n$ $(i \neq j, i \neq m, j \neq m)$. Thus, $l^D(\psi^\star)$ is a strictly convex function under an anchor constraint.

Detailed Derivation for Subsection 4.3.3 We show the derivation for the loss function of the shape restricted hazard function with additional covariates in Subsection 4.3.3. We consider the isotonic proportional hazard model $\lambda(u \mid Z, W(u)) = \lambda_0(u) \exp\{\psi(Z) + \beta^T W(u)\}$, where the negative log likelihood function is defined as

$$l^N(\psi, \beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ -\psi_i - \beta^T W_i(u) + log\left(\sum_{j=1}^{n} Y_j(u) e^{\psi_j + \beta^T W_j(u)}\right) \right\} dN_i,$$

and the score function of $l^N(\psi, \beta)$ is derived as

$$\begin{aligned}
u_i^{N,\psi} &= \frac{\partial l^N(\psi, \beta)}{\partial \psi_i} = \int_0^\infty \left\{ -dN_i(u) + \frac{Y_i(u) e^{\psi_i + \beta^T W_i(u)}}{\sum_{j=1}^{n} Y_j(u) e^{\psi_i + \beta^T W_i(u)}} \sum_{s=1}^{n} dN_s(u) \right\} \\
&= \int_0^\infty \left\{ -dN_i(u) + Y_i(u) e^{\psi_i + \beta^T W_i(u)} d\tilde{\Lambda}_0(\phi, \beta, u) \right\} \\
u_s^{N,\beta} &= \frac{\partial l^N(\psi, \beta)}{\partial \beta_s} = \sum_{i=1}^{n} \int_0^\infty \{ -W_{is}(u) + \frac{\sum_{j=1}^{n} Y_j(u) e^{\psi_j + \beta^T W_j(u)} w_{js}(u)}{\sum_{j=1}^{n} Y_j(u) e^{\psi_j + \beta^T W_j(u)}} \} dN_i(u) \\
&= \sum_{i=1}^{n} \int_0^\infty W_{is}(u) \left\{ -dN_i(u) + Y_i(u) e^{\psi_i + \beta^T W_i(u)} d\tilde{\Lambda}_0(\phi, \beta, u) \right\}
\end{aligned}$$

for $i = 1, \ldots, n$ and $s = 1, \ldots, p$, where $\tilde{\Lambda}_0(\phi, \beta, t) = \int_0^t \{\sum_{i=1}^n dN_i(u)\} / \{\sum_{j=1}^n Y_j(u) e^{\psi_i} + \beta^T W_i(u)\}$. By mimicking the score functions $u^{N,\psi}$ and $u^{N,\beta}$, we define a score function in our model, which is

$$
\begin{aligned}
u_i^{\psi} &= \int_0^{\infty} \big\{-dN_i(u) + Y_i(u)\psi_i du + Y_i(u)\beta^T W_i(u)du + Y_i(u)\hat{\Lambda}_0(\phi,\beta,t)\big\} \\
&= \int_0^{\infty} \sum_{j=1}^{n} h_{ij}(u)\psi_j du - \int_0^{\infty} q_i(u) + \int_0^{\infty} Y_i(u) \sum_{s=1}^{p} \big\{W_{is}(u) - \bar{W}_s(u)\big\}\beta_s du \\
u_s^{\beta} &= \sum_{i=1}^{n} \int_0^{\infty} W_{is}(u)\big\{-dN_i(u) + Y_i(u)\psi_i du + Y_i(u)\beta^T W_i(u)du + Y_i(u)\hat{\Lambda}_0(\phi,\beta,t)\big\} \\
&= \sum_{i=1}^{n} \int_0^{\infty} Y_i(u)\big\{W_{is}(u) - \bar{W}_s(u)\big\} \sum_{t=1}^{p} \big\{W_{it}(u) - \bar{W}_t(u)\big\}\beta_t du \\
&\quad - \sum_{i=1}^{n} \int_0^{\infty} \big\{W_{is}(u) - \bar{W}_s(u)\big\} dN_i(u) + \sum_{i=1}^{n} \int_0^{\infty} Y_i(u)\big\{W_{is}(u) - \bar{W}_s(u)\big\}\psi_i
\end{aligned}
$$

for $i = 1, \dots, n$ and $s = 1, \dots, p$, where $\bar{W}_s(u) = \big\{\sum_{j=1}^{n} Y_j(u) w_{js}(u)\big\} / \big\{\sum_{j=1}^{n} Y_j(u)\big\}$ and $\hat{\Lambda}_0(\phi,\beta,t) = \int_0^t \{\sum_{i=1}^{n}\{dN_i(u) - Y_i(u)\phi_i du - Y_i(u)\beta^T W_i(u)du\}\}/\{\sum_{j=1}^{n} Y_j(u)\}$. Correspondingly, the loss function is defined as $\psi^T H\psi/2 + \beta^T H^{\circ}\beta/2 + \psi^T H^{\diamond}\beta - \psi^T q - \beta^T q^{\circ}$, which is the same as $l^A(\theta)$.

CHAPTER 5: ADDITIVE ISOTONIC PROPORTIONAL HAZARDS MODELS

5.1 Introduction

The isotonic proportional hazards model is a useful nonparametric method to estimate an isotonic (or monotone) covariate effect on a hazard function under the natural assumption that the the hazard was isotonic in a continuous covariate. In Chapter 3, an efficient computation of pseudo iterative convex minorant algorithm was proposed to estimate the isotonic covariate effect. The method, however, only handled a single continuous covariate. When more than one continuous covariate exists, multivariate extension of the isotonic regression (Robertson et al. 1988, p.12) have been studied. Denote $\phi_j(\cdot)$ as an isotonic function, where $\phi_j(x) \leq \phi_j(y)$ whenever $x \leq y$, $j = 1, \ldots, p$ with $p \geq 2$. The function having the additive isotonic structure $\phi(\cdot) = \sum_{j=1}^{p} \phi_j(\cdot)$ is said to be isotonic with respect to a partial order, where $\phi(\boldsymbol{x}) \leq \phi(\boldsymbol{y})$ whenever $x_j \leq y_j$ for all $j = 1, \ldots, p$. Here, $\boldsymbol{x} = (x_1, \ldots, x_p)$ and $\boldsymbol{y} = (y_1, \ldots, y_p)$. The additive isotonic structure has been well-studied in a standard uncensored regression setting with independent and identically distributed data (Bacchetti 1989, Morton-Jones et al. 2000, Mammen and Yu 2007, Cheng 2009) but not for the right censored data.

In this paper, we suggest the additive isotonic proportional hazards model by incorporating the additive isotonic function in the proportional hazards model. That is, we assume

$$\lambda(t \mid \boldsymbol{Z}_i) = \lambda_0(t) e^{\phi_1(Z_{i1}) + \cdots + \phi_p(Z_{ip})}, \tag{5.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ is the ith subject's $p \times 1$ continuous covariate vector with $p \geq 2$. Accordingly, the hazard function is defined as the isotonic function with respect to the partial order on the covariates. The partial likelihood (Cox 1972) is then well defined, as formulated in Subsection 5.2.1. Thus, $\phi(\cdot)$ in (5.1) can be estimated by maximizing the partial likelihood under additive isotonic constraints, without simultaneous estimation of the baseline hazard function. Classically, $\phi(\cdot)$ is specified parametrically such as low order polynomials of $\boldsymbol{Z}_i$. In our case, however, $\phi(\cdot)$ is isotonic but otherwise unspecified, and thus, a special technique is needed to obtain the constrained estimator.

Bacchetti (1989) considered the additive isotonic regression that incorporated the additive structure of isotonic functions to the standard regression setting. He suggested a cyclic algorithm that updated a univariate $\phi_j(\cdot)$ by iterating, $j = 1, \ldots, p, 1, \ldots, p, \ldots$, until convergence, holding the other parameters $\{\phi_1(\cdot), \ldots, \phi_{j-1}(\cdot), \phi_{j+1}(\cdot), \ldots, \phi_p(\cdot)\}$ fixed. The simple structure of the least squares function gave a closed form solution for $\phi_j(\cdot)$ in each univariate optimization, which could be computed by the pool-adjacent-violators algorithm (Ayer et al. 1955). The cyclic algorithm is directly applicable to our model with the partial likelihood and has a convergence property as shown in Theorem 5.2. On the other hand, the complicated structure of the partial likelihood does not yield a closed form solution, and more complex computation is needed, i.e. double iterations are needed where one is from the univariate optimization and the other is from the cyclic optimization. Computational efficiency and stability may depend on the univariate optimization method. We implement the pseudo iterative convex minorant algorithm in the cycling algorithm.

In Subsection 5.2.1, we define the partial likelihood with multiple covariates having an additive isotonic structure. The cyclic and univariate optimization methods are described in Subsections 5.2.2 and 5.2.3, respectively, without censorship. Our model is extended to allow censorship and multiple time-dependent covariates in Subsections 5.2.4 and 5.2.5, respectively. A separate estimator for baseline hazard function and the inclusion of additional covariates are described in Section 5.3. In simulation study in Section 5.4, the cyclic pool-adjacent-violators algorithm increases computational speed, with moderate bias and mean square error reductions. An analysis of a cardiovascular disease dataset demonstrates the practical utility of our methodology in estimating nonlinear covariate effects under the additive isotonic structure.

5.2 Additive isotonic proportional hazards models

5.2.1 Data set-up and partial likelihood

Suppose that T is a failure time, C is a censoring time and $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ is a $p \times 1$ vector of continuous covariates with $p \geq 2$, where T and C are conditionally independent on $\boldsymbol{Z}$. Define $X = \min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. The observed data consist of n replicates of $(X, \Delta, \boldsymbol{Z})$, denoted by $(X_i, \Delta_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$, where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})$. Under the additive isotonic proportional hazards model in (5.1), the partial likelihood is defined as

$$pl(\boldsymbol{\phi}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{e^{\phi_1(Z_{i1}) + \cdots + \phi_p(Z_{ip})}}{\sum_{s=1}^{n} Y_s(t) e^{\phi_1(Z_{s1}) + \cdots + \phi_p(Z_{sp})}} \right\}^{dN_i(t)}, \tag{5.2}$$

where $\boldsymbol{\phi} = \{\phi_j(Z_{ij}), i = 1, \ldots, n, j = 1, \ldots, p\}$, $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ is a counting process and $Y_i(t) = I(X_i \geq t)$ is an at-risk process for the ith subject, $i = 1, \ldots, n$. Since the parameter $\boldsymbol{\phi}$ only enters the partial likelihood at the observed covariate values in the dataset, we restrict the estimator to be piecewise constant with potential jumps at Z_{ij}'s for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

5.2.2 Cyclic optimization

Denote $\hat{\boldsymbol{\phi}}$ as the maximizer of the partial likelihood under the additive isotonic constraint $\phi_j(Z_{(1)j}) \leq \cdots \leq \phi_j(Z_{(n)j})$ for $j = 1, \ldots, p$, where $Z_{(i)j}$ is the ith smallest value among Z_{1j}, …, Z_{nj}. Similarly to the univariate case, we impose an anchor constraint that $\phi_j(K_j) = \delta_j$ by prespecifying a constant K_j. Otherwise, $\{\phi_j(Z_{(1)j}) \pm \delta_j, \ldots, \phi_j(Z_{(n)j}) \pm \delta_j\}$ yields the same value to $pl(\boldsymbol{\phi})$. Under the anchor constraint, the model in (5.1) is reformulated as

$$\lambda(t \mid \boldsymbol{Z_i}) = \lambda_0(t) e^{\phi_1(Z_{i1}) + \cdots + \phi_p(Z_{ip})} = \{\lambda_0(t) e^{\delta}\} e^{\psi_1(Z_{i1}) + \cdots + \psi_p(Z_{ip})}, \tag{5.3}$$

where $\delta = \delta_1 + \cdots + \delta_p$ and $\psi_j(\cdot) = \phi_j(\cdot) - \delta_j$ with $\psi_j(K_j) = 0$ for $j = 1, \ldots, p$. Since the baseline hazard function absorbs δ, what we actually estimate is $\psi_j(\cdot)$, not $\phi_j(\cdot)$. We regard δ as a nuisance parameter, with the only difference between $\psi_j(\cdot)$ and $\phi_j(\cdot)$ being the reference group defining the hazard ratio parameters. In other words, hazard ratios based on $\phi_j(\cdot)$ and $\psi_j(\cdot)$ are identical, i.e., $\exp\{\phi_j(\cdot) - \phi(K_j)\} = \exp\{\psi_j(\cdot) - \psi_j(K_j)\}$ for $j = 1, \ldots, p$. Practically we set K_j to $Z_{(k_j)j}$, where $Z_{(k_j)j}$ is the closest value to K_j among $Z_{(i)j}$'s, because $\psi_j(\cdot)$ is only identifiable at $Z_{(i)j}$'s. Denote the negative log partial likelihood as

$$lpl^N(\boldsymbol{\psi}) = \textstyle\sum_{i=1}^n \displaystyle\int_0^\infty \left[-\{\psi_{i1} + \cdots + \psi_{ip}\} + \log\left\{\textstyle\sum_{s=1}^n Y_s(u) e^{\psi_{s1} + \cdots + \psi_{sp}}\right\} \right] dN_i(u), \tag{5.4}$$

where $\boldsymbol{\psi} = \{\psi_{ij}, i = 1, \ldots, n, j = 1, \ldots, p\}$, and $\psi_{ij} = \psi_j(Z_{ij})$.

Lemma 5.1. *The negative log partial likelihood $lpl^N(\boldsymbol{\psi})$ is convex.*

Denote $\Psi = \Psi_1^{k_1} \times \cdots \times \Psi_p^{k_p}$, where $\Psi_j^{k_j} = \{\boldsymbol{\psi} \in \mathbb{R}^n \mid \psi_{(1)j} \leq \ldots \leq \psi_{(n)j}, \psi_{(k_j)j} = 0\}$ is a convex cone for $j = 1, \ldots, p$, where $\psi_{(i)j} = \psi_j(Z_{(i)j})$. The problem of maximizing the partial likelihood under the anchor and isotonic constraints is equivalent to minimizing the convex function $lpl^N(\boldsymbol{\psi})$ over the convex cone Ψ. We compute this by using the cyclic algorithm, as stated in the following steps:

Step 5.1: Set an initial value $\boldsymbol{\psi}_j^{(0)} \in \Psi_j$ for $j = 1, \ldots, p$

Step 5.2: Update $\boldsymbol{\psi}_j^{(r)} \in \Psi_j^{k_j}$, $j = 1, \ldots, p$, iteratively by regarding the other parameters $\{\boldsymbol{\psi}_1^{(r)}, \ldots, \boldsymbol{\psi}_{j-1}^{(r)}, \boldsymbol{\psi}_{j+1}^{(r-1)}, \cdots \boldsymbol{\psi}_p^{(r-1)}\}$ are fixed in the partial likelihood.

Step 5.3: Repeat the cycle $r = 1, 2, \ldots$ in Step 5.2 until convergence, where the convergence criterion is $\sum_{j=1}^{p} d(\boldsymbol{\psi}_j^{(r)}, \boldsymbol{\psi}_j^{(r-1)}) < \epsilon$ for small $\epsilon > 0$ and Euclidean distance $d(\cdot, \cdot)$.

Theorem 5.2. *Assume that* $lpl^N(\boldsymbol{\psi}) > -\infty$ *on* Ψ. *Let* $\boldsymbol{\psi}^{(r)} = \{\psi_{ij}^{(r)}, i = 1, \ldots, n, j = 1, \ldots, p\}$ *generated by the cyclic algorithm from any starting values in* Ψ. *Then,* $lpl^N(\boldsymbol{\psi}^{(r)})$ *converges to* $min_{\boldsymbol{\psi} \in \Psi} lpl^N(\boldsymbol{\psi})$ *as* $r \to \infty$.

The univariate optimization in Step 5.2 is further explained in the next subsection. As stated in Theorem 5.2, the cyclic algorithm has a convergence property, regardless of the starting value. On the other hand, it does not guarantee the uniqueness of the isotonic estimator, because different isotonic estimators may possibly yield the same minimum value to $lpl^N(\boldsymbol{\psi})$ in (5.4). The same situation occurs for the additive isotonic regression (Bacchetti 1989) where different isotonic estimators may yield the same minimum value for the least squares function.

5.2.3 Univariate optimization without censoring

In this subsection, we focus on the univariate estimation $\hat{\boldsymbol{\psi}}_j$ in Step 5.2 for the cyclic algorithm. Holding the other parameters $\{\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_{j-1}, \boldsymbol{\psi}_{j+1}, \cdots, \boldsymbol{\psi}_p\}$ fixed, we reduce $lpl^N(\boldsymbol{\psi})$ in (5.4) to

$$l^N(\boldsymbol{\psi}_j) = \sum_{i=1}^{n} \int_0^\infty \left[-\psi_{(i)j} + \log\{\sum_{s=1}^{n} Y_{(s)j}(u) e^{\psi_{(s)j}}\} \right] dN_{(i)}(u) + C, \qquad (5.5)$$

where $Y_{(s)j}(u) = Y_{(s)}(u)\exp\{\sum_{l=1,l\neq j}^{n}\psi_{(i)l}\}$, and $C = \sum_{i=1}^{n}\int_0^\infty\{\sum_{s=1,s\neq j}^{n}\psi_{(i)s}\}dN_{(i)}(u)$ which does not involve the parameter $\boldsymbol{\psi}_j$. Here, $N_{(i)}(t)$ and $Y_{(i)j}(t)$ are the counting and at-risk processes corresponding to the subject whose covariate is $Z_{(i)j}$. In the sequel, we drop the subparentheses for notational convenience, as needed. By including other fixed parameters $\{\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_{j-1}, \boldsymbol{\psi}_{j+1}, \cdots, \boldsymbol{\psi}_p\}$ to $Y_{sj}(u)$, we define $l^N(\boldsymbol{\psi}_j)$ in (5.5), which is the same likelihood for the univariate isotonic proportional hazards model. Thus, we may apply the pseudo iterative convex minorant algorithm to compute $\hat{\boldsymbol{\psi}}_j$.

The procedure of the pseudo iterative convex minorant algorithm is described as follows. Let $w_i(\boldsymbol{\psi}_j) = \int_0^\infty[\{Y_{ij}(u)\sum_{l=1}^{n}dN_l(u)\}/\{\sum_{s=1}^{n}Y_{sj}(u)\exp(\psi_{sj})\}]$ and $\Delta_i = \int_0^\infty dN_i(u)$ be a censoring indicator for $i = 1, \ldots, n$. Set an initial value of $\dot{\boldsymbol{\psi}}_j^{(0)} \in \Psi_j$, where $\Psi_j = \{\psi \in \mathbb{R}^n \mid \psi_{1j} \leq \cdots \leq \psi_{nj}\}$. We then solve $\boldsymbol{\psi}_j^+ = \arg\min_{\boldsymbol{\psi}_j \in \Psi_j}\sum_{i=1}^{n}\{\psi_{ij} - w_i(\dot{\boldsymbol{\psi}}_j^{(a-1)})\}^2 w_i(\dot{\boldsymbol{\psi}}_j^{(a-1)})$, which can be easily computed by using the pool-adjacent-violators algorithm, and take log transformation $\dot{\boldsymbol{\psi}}_j^{(a)} = \{log(\psi_{1j}^+), \ldots, log(\psi_{nj}^+)\}$. Repeat this until convergence, where the convergence criterion is $d_e(\dot{\boldsymbol{\psi}}_j^{(a)}, \dot{\boldsymbol{\psi}}_j^{(a-1)}) = \sum_{i=1}^{n}|\exp(\dot{\psi}_{ij}^{(a)}) - \exp(\dot{\psi}_{ij}^{(a-1)})| < \dot{\epsilon}$ for small $\epsilon > 0$. After it converges at the bth step, the anchor constraint is imposed by vertical shift, $\ddot{\psi}_{ij} = \dot{\psi}_{ij}^{(b)} - \dot{\psi}_{k_j j}^{(b)}$ for $i = 1, \ldots, n$. This guarantees that whenever $\dot{\epsilon}$ converges to zero, $\ddot{\boldsymbol{\psi}}_j$ converges to $\hat{\boldsymbol{\psi}}_j$.

An advantage of the pseudo iterative convex minorant algorithm is its efficient computation, using the pool-adjacent-violators algorithm iteratively. On the other hand, it may not have a global convergence property, due to the complicated structure of the partial likelihood. The other existing computations of iterative quadratic programming and iterative convex minorant algorithm (Groeneboom and Wellner 1992, pp. 69-73) are known to be unstable owing to inverting a large Hessian matrix and imposing an anchor constraint, respectively. We examine the performances of the algorithms in simulation studies in Section 5.4.

5.2.4 Censoring

Suppose that failure times of some subjects are censored. Since censored subjects contribute limited information to the partial likelihood, the isotonic estimator jumps only at the covariate value associated with the failure events, as stated in Proposition 5.3 below. Let $n^\star$ be the number of subjects with observed failure time out of the total n subjects and $\boldsymbol{Z}_i^\star = (Z_{i1}^\star, \dots, Z_{ip}^\star)$ be their covariate vector for $i = 1, \dots, n^\star$. Let $\psi_{ij}^\star = \psi_j(Z_{(i)j}^\star)$ for $i = 1, \dots, n^\star$ and $j = 1, \dots, p$, where $Z_{(i)j}^\star$ is the ith order statistic amongst $Z_{1j}^\star, \dots, Z_{n^\star j}^\star$.

Proposition 5.3. *Assume that $\psi_{hj} = \psi_{(1)j}^\star$ if $Z_{hj} < Z_{(1)j}^\star$, $h = 1, \dots, n$. Then, the isotonic estimator jumps only at $\boldsymbol{Z}_i^\star$, $i = 1, \dots, n^\star$.*

It does not affect that the partial likelihood is a convex function in Lemma 5.1, and the cyclic algorithm is directly applicable to optimize the partial likelihood. On the other hand, one modification is needed for the univariate optimization, because the number of parameters is reduced from n to $n^\star$. Define $n^\star$ disjoint intervals of $I_{1j}^\star = (-\infty, Z_{(1)j}^\star) \cup [Z_{(1)j}^\star, Z_{(2)j}^\star), I_{2j}^\star = [Z_{(2)j}^\star, Z_{(3)j}^\star), \dots, I_{n^\star j}^\star = [Z_{(n^\star)j}^\star, +\infty)$. Under Proposition 5.3, we define the partial likelihood by

$$pl^C(\boldsymbol{\psi}_j^\star) = \prod_{i=1}^{n^\star} \prod_{t \geq 0} \left\{ \frac{e^{\psi_{ij}^\star}}{\sum_{s=1}^n Y_{sj}^\star(t) e^{\psi_{sj}^\star}} \right\}^{dN_i^\star(t)},$$

where $Y_{ij}^\star(t) = \sum_{h \in R_{ij}^\star} Y_{hj}(t)$, $R_{ij}^\star = \{h : Z_{hj} \in I_{ij}^\star, h = 1, \dots, n\}$, and $N_i^\star(t)$ is the counting process corresponding to $Z_{(i)j}^\star$. The assumption, $\psi_{hj} = \psi_{(1)j}^\star$ if $Z_{hj} < Z_{(1)j}^\star$ for $h = 1, \dots, n$, allows estimation on the all values of $\boldsymbol{Z}$ including the left side of $Z_{(1)j}^\star$. Since log of $pl^C(\boldsymbol{\psi}_j^\star)$ has the same form as $l^N(\boldsymbol{\psi}_j)$ in (5.5), the pseudo iterative convex minorant algorithm is directly applicable for $pl^C(\boldsymbol{\psi}_j^\star)$,

5.2.5 Time-dependent covariates

Consider the model $\lambda(t \mid \boldsymbol{Z}_i(t)) = \lambda_0(t)\exp\{\psi_1\{Z_{i1}(t)\} + \cdots + \psi_p\{Z_{ip}(t)\}\}$, where $\boldsymbol{Z}_i(t) = \{Z_{i1}(t), \ldots, Z_{ip}(t)\}$ is a $p \times 1$ time-dependent covariates vector for the ith subject, $i = 1, \ldots, n$. We assume that the isotonic function $\psi_j(\cdot)$ does not change over time for $j = 1, \ldots, p$. The partial likelihood is defined as

$$pl(\boldsymbol{\psi}) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{e^{\psi_1\{Z_{i1}(t)\}+\cdots+\psi_p\{Z_{ip}(t)\}}}{\sum_{s=1}^{n} Y_s(t)e^{\psi_1\{Z_{s1}(t)\}+\cdots+\psi_p\{Z_{sp}(t)\}}}\right\}^{dN_i(t)}. \tag{5.6}$$

The values of the time-dependent covariates prior to the observed failure time contribute limited information to the partial likelihood, which restricts the form of the isotonic partial likelihood estimator, as stated in the Proposition 5.4 below. Formally let $n^\star$ be the number of subjects with observed failure time, and $Z_{ij}^*(t)$ be their covariate vector value for $i = 1, \ldots, n^\star$ and $j = 1, \ldots, p$. Let $\boldsymbol{Z}_i^* = (Z_{i1}^*, \ldots, Z_{ip}^*)$, where $Z_{ij}^* = Z_{ij}^*(X_i^*)$ and X_i^* is the ith subject's failure time. Let $\psi_{ij}^* = \psi_j(Z_{(i)j}^*)$ for $i = 1, \ldots, n^\star$ and $j = 1, \ldots, p$, where $Z_{(i)j}^*$ is the ith order statistic amongst $Z_{1j}^*, \ldots, Z_{n^\star j}^*$.

Proposition 5.4. *Assume that $\psi_{hj} = \psi_{1j}^*$ if $Z_{hj}(X_i) < Z_{(1)j}^*$ for $i, h = 1, \ldots, n$, Then, the isotonic estimator jumps only at $\boldsymbol{Z}_i^*$, $i = 1, \ldots, n^\star$.*

Similarly to the censored data in Subsection 5.2.4, the cycling algorithm is directly applicable, but one modification is needed for the univariate optimization. Define $n^\star$ disjoint intervals of $I_{1j}^* = (-\infty, Z_{(1)j}^*) \cup [Z_{(1)j}^*, Z_{(2)j}^*)$, $I_{2j}^* = [Z_{(2)j}^*, Z_{(3)j}^*), \ldots, I_{n^\star j}^* = [Z_{(n^\star)j}^*, +\infty)$. Under Proposition 5.4, we defined the partial likelihood by

$$pl^D(\boldsymbol{\psi}_j^*) = \prod_{i=1}^{n^\star}\prod_{t\geq 0}\left\{\frac{e^{\psi_{ij}^*}}{\sum_{s=1}^{n} Y_{sj}^*(t)e^{\psi_{sj}^*}}\right\}^{dN_i^*(t)},$$

where $Y_{ij}^*(t) = \sum_{h\in R_{ij}^*} Y_{hj}(t)$, $R_{ij}^* = \{h : Z_{hj} \in I_{ij}^*, h = 1, \ldots, n\}$, and $N_i^*(t)$ is the counting process corresponding to $Z_{(i)j}^*$. Since log of $pl^D(\boldsymbol{\psi}_j^*)$ has the same form of $l^N(\boldsymbol{\psi}_j)$ in

(5.5), the pseudo iterative convex minorant algorithm is directly applicable to $pl^D(\boldsymbol{\psi}_j^*)$.

5.3 Extension

5.3.1 Baseline hazard function

We are not able to estimate the baseline hazard function $\lambda_0(t)$ and vertical shift parameter δ in (5.3), because they are not identifiable as discussed in Subsection 5.2.2 above. Instead, we can estimate the baseline hazard function including an anchor effect $\lambda_0^\star(t)$, where $\lambda_0^\star(t) = \lambda_0(t)\exp(\delta)$. This is the same approach to estimate a baseline hazard function at a reference group in the standard Cox model. Thus, Breslow (1972)'s estimator can be directly applicable, which is

$$\hat{\Lambda}_0^\star(t,\hat{\psi}) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n Y_j(u)e^{\hat{\psi}_1\{Z_{j1}(u)\}+\cdots+\hat{\psi}_p\{Z_{jp}(u)\}}},$$

where $\Lambda_0^\star(t,\psi)$ is a profiled estimator of the cumulative baseline hazard function including an anchor effect, and $\hat{\psi}(\cdot)$ is the isotonic estimator from the partial likelihood.

5.3.2 Additional covariates

Suppose that there exists additional q covariates. We include those covariates to the model $\lambda(t \mid \boldsymbol{Z}_i(t), \boldsymbol{W}_i(t)) = \lambda_0(t)\exp\{\psi_1\{Z_{i1}(t)\} + \cdots + \psi_p\{Z_{ip}(t)\} + \boldsymbol{\beta}^T\boldsymbol{W}_i(t)\}$, where $\boldsymbol{\beta}$ is $q \times 1$ regression parameter and $\boldsymbol{W}_i(t) = \{W_{i1}(t), \cdots, W_{iq}(t)\}$ is $q \times 1$ covariates vector for $i = 1, \ldots, n$. The partial likelihood is defined as

$$pl(\boldsymbol{\phi}) = \prod_{i=1}^n \prod_{t\geq 0} \left\{ \frac{e^{\psi_1\{Z_{i1}(t)\}+\cdots+\psi_p\{Z_{ip}(t)\}+\boldsymbol{\beta}^T\boldsymbol{W}_i(t)}}{\sum_{s=1}^n Y_s(t)e^{\psi_1\{Z_{s1}(t)\}+\cdots+\psi_p\{Z_{sp}(t)\}+\boldsymbol{\beta}^T\boldsymbol{W}_i(t)}} \right\}^{dN_i(t)}.$$

The partial likelihood can be maximized by the following step. We set an initial value of $(\boldsymbol{\psi}^{(0)}, \boldsymbol{\beta}^{(0)}) \in \Psi \times \mathbb{R}^q$. We then update $\boldsymbol{\psi}^{(m)}$ given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}$ using the cyclic pseudo iterative convex minorant algorithm, and update $\boldsymbol{\beta}^{(m)}$ given $\boldsymbol{\psi} = \boldsymbol{\psi}^{(m)}$ using the

Newton-Raphson algorithm $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} - H(\boldsymbol{\psi}^{(m)}, \boldsymbol{\beta}^{(m-1)})^{-1} U(\boldsymbol{\psi}^{(m)}, \boldsymbol{\beta}^{(m-1)})$, where

$$\begin{aligned} U(\boldsymbol{\psi}, \boldsymbol{\beta}) &= \textstyle\sum_{i=1}^{n} \displaystyle\int_0^\infty \left\{ \boldsymbol{W}_i(t) - \frac{\sum_{l=1}^{n} Y_l^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_l(t)} \boldsymbol{W}_l(t)}{\sum_{s=1}^{n} Y_s^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_s(t)}} \right\} dN_i(t), \\ H(\boldsymbol{\psi}, \boldsymbol{\beta}) &= \textstyle\sum_{i=1}^{n} \displaystyle\int_0^\infty \left[- \frac{\sum_{l=1}^{n} Y_l^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_l(t)} \boldsymbol{W}_l(t)^{\otimes 2}}{\sum_{s=1}^{n} Y_s^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_s(t)}} \right. \\ &\qquad \left. + \frac{\left\{ \sum_{l=1}^{n} Y_l^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_l(t)} \boldsymbol{W}_l(t) \right\}^{\otimes 2}}{\left\{ \sum_{s=1}^{n} Y_s^\circ(t, \boldsymbol{\psi}) e^{\boldsymbol{\beta}^T \boldsymbol{W}_s(t)} \right\}^2} \right] dN_i(t), \end{aligned}$$

where $Y_i^\circ(t, \boldsymbol{\psi}) = Y_i(t) \exp\{\psi_1(Z_{i1}(t)) + \cdots + \psi_p(Z_{ip}(t))\}$ and $W^{\otimes 2} = W^T W$. Repeat these updates until convergence, where the convergence criteria is $\sum_{j=1}^{p} d(\boldsymbol{\psi}_j^{(m)}, \boldsymbol{\psi}_j^{(m-1)}) + d(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m-1)}) < \epsilon$ for small $\epsilon > 0$.

5.4 Simulations

We performed simulation studies to examine the performance of the cyclic pseudo iterative convex minorant algorithm, as well as the cyclic iterative quadratic programming and cyclic iterative convex minorant algorithm. As a gold standard, we also evaluated one-step update using the pseudo iterative convex minoramt algorithm from the true initial value. For the first part of the simulations, we considered time independent covariates with $p = 2$, where $\boldsymbol{Z} = (Z_1, Z_2)$ were independently generated from a uniform distribution on (0,1). Three combinations of isotonic functions were considered on $(0, 1)$: $\{\psi_1(z) = z, \psi_2(z) = z\}$, $\{\psi_1(z) = z, \psi_2(z) = z^2\}$, and $\{\psi_1(z) = z^2, \psi_2(z) = z\}$. The failure time was then generated from the additive isotonic proportional hazard with a constant baseline hazard function. The right censoring time was then independently generated from a uniform distribution with approximately 30% censoring. We repeated the simulations 500 times with $n = 100$, 500 and 1000. The anchor constraints were set to $K_1 = K_2 = 0{\cdot}5$. For each data set, the initial values of $\boldsymbol{\psi}^0 \in \Psi$ was set from the stnadard Cox model with a linear function of $\gamma_1 Z_1 + \gamma_2 Z_2$, i.e. $\boldsymbol{\psi}_j^{(0)} = \{|\gamma_j| \bar{Z}_{1j}, \ldots, |\gamma_j| \bar{Z}_{nj}\}$ for $j = 1, 2$, where $\bar{Z}_{ij} = Z_{(i)j} - Z_{(k_j)j}$. For the second part of the simulations, we consider time-dependent covariates $\boldsymbol{Z} = \{Z_1(t), Z_2(t)\}$. By assuming the time-dependent covariates are piecewise constant, we generate each $Z_j(t)$, $j = 1, 2$, from uniform distribution on (0,1), independently, with disjoint time intervals. Other scenarios are the same as the setting in the first simulation.

For the evaluation of the performance among the algorithms, we computed integrate mean square error, $\int_0^1 E\{\psi_j(Z) - \hat{\psi}_j(Z)\}^2 dZ$ for $j = 1, 2$, where $\psi_j(Z) = \phi_j(Z) - \phi_j(K_j)$ and $\hat{\psi}_r(\cdot)$ is an estimated isotonic function for the rth data set for $r = 1, \ldots, 500$. It is approximated by $\sum_{r=1}^{R} \sum_{g=1}^{G} \{\psi_j(z_g) - \hat{\psi}_{j,r}(z_g)\}^2/(GR)$ based on equally spaced grid points of z_g's between 0·001 and 0·999, with $G = 1000$ grid points and $R = 500$ simulation runs. Then, the total integrated mean square error is computed by summing the two integrate mean square error for $\hat{\boldsymbol{\psi}}_1$ and $\hat{\boldsymbol{\psi}}_2$.

Tables 5.1 and 5.2 show simulations results for time independent and time-dependent covariates. We exclude non-convergent cases for calulating the integrated mean squared error and the computing time. The pseudo iterative convex minorant algorithm has almost 100% convergence results in the cycling algorithm, except few non-convergence cases for time-dependent covariate. The other existing methods of the iterative quadratic programming and iterative convex minorant algorithm give unstable results, approximately 10-30% convergence failures. The iterative quadratic programming requires the inversion of a high dimensional Hessian matrix from the partial likelihood, which leads to unstable convergence results and large computational burdens. The anchor constraint is originally designed for a sparse Hessian matrix without the anchor constraint, and it does not perform well with our partial likelihood having full Hessian matrix, with an anchor constraint imposed. The pseudo iterative convex minorant algorithm dramatically improves computational speed, e.g. approximately 20 seconds, 9 minutes and 1 hours from the pseudo iterative convex minorant algorithm, iterative convex minorant algorithm and

Table 5.1: Simulation results for time independent covariates: IMSE multiplied by 10^5, CPU time in seconds and convergence percentage.

Type	ϕ_1	ϕ_2	n	IQM		ICM		PICM		PICM.true
				IMSE(1/2)	Cpu(Cv)	IMSE(1/2)	Cpu(Cv)	IMSE(1/2)	Cpu(Cv)	IMSE(1/2)
Comp	Z	Z	100	19(9/10)	8(75)	270(130/140)	2(84)	74(39/35)	0(100)	14(7/7)
			500	10(5/5)	491(85)	20(7/12)	90(87)	12(6/6)	5(100)	9(6/3)
			1000	9(4/4)	4945(86)	10(5/5)	597(88)	9(5/5)	21(100)	8(6/2)
	Z	Z^2	100	20(10/11)	9(75)	253(136/117)	1(83)	73(40/34)	0(100)	14(7/7)
			500	10(5/5)	431(85)	14(8/6)	74(88)	11(6/5)	4(100)	8(6/2)
			1000	8(5/3)	4115(85)	10(7/3)	431(88)	9(5/4)	17(100)	8(6/2)
	Z^2	Z	100	20(10/10)	8(75)	233(63/170)	2(83)	68(32/36)	0(100)	14(7/7)
			500	10(4/5)	497(84)	18(6/13)	88(87)	12(5/7)	5(100)	9(6/3)
			1000	8(3/5)	4689(86)	9(4/5)	607(88)	9(4/5)	20(100)	8(6/2)
Cens	Z	Z	100	24(12/11)	11(80)	47(16/30)	1(83)	47(25/22)	0(100)	17(8/8)
			500	11(5/6)	225(86)	12(6/6)	50(87)	12(6/6)	3(100)	9(6/3)
			1000	9(5/5)	1752(86)	9(5/5)	314(87)	10(5/5)	12(100)	9(6/2)
	Z	Z^2	100	24(12/12)	13(80)	30(13/18)	1(83)	46(25/22)	0(100)	17(9/9)
			500	11(6/5)	208(86)	11(6/5)	44(88)	12(6/6)	2(100)	9(6/3)
			1000	9(5/4)	1418(86)	9(5/4)	265(87)	9(5/4)	10(100)	8(6/2)
	Z^2	Z	100	23(12/11)	11(79)	45(15/30)	1(83)	47(26/21)	0(100)	17(8/8)
			500	11(5/6)	222(86)	16(5/11)	49(87)	13(6/7)	3(100)	10(6/3)
			1000	9(4/5)	1774(85)	9(4/5)	330(87)	9(4/5)	12(100)	9(6/2)

IQM: iterative quadratic programming; ICM: iterative convex minorant algorithm; PICM: pseudo iterative convex minorant algorithm; PICM.true: One step PICM from true initial value; Comp: complete case; Cens: censoring case (about 30%); IMSE: (total) integrated mean squared error (IMSE for ϕ_1 / IMSE for ϕ_2); Cpu(Cv): computing time in second. (convergence percentage).

iterative quadratic programing method for the complete data with time-independent covariate and $n = 1000$. The integrated mean squared errors decrease when sample size increases for all methods. As expected, the one step pseudo iterative convex minorant algorithm has the smallest integrated mean squared error.

5.5 Data analysis

Folic Acid for Vascular Outcome Reduction in Transplantation (FAVORIT) study was a multicenter double-blind randomized controlled clinical trial to investigate if vitamin supplementation reduces risk of cardiovascular disease (CVD) in kidney transplant recipients (Bostom et al. 2011). Four thousand one hundred ten study participants were

Table 5.2: Simulation results for time-dependent covariates: IMSE multiplied by 10^5, CPU time in seconds and convergence percentage.

Type	ϕ_1	ϕ_2	n	IQM		ICM		PICM		PICM.true
				IMSE(1/2)	Cpu(Cv)	IMSE(1/2)	Cpu(Cv)	IMSE(1/2)	Cpu(Cv)	IMSE(1/2)
Comp	Z	Z	100	21(11/10)	1(33)	68(31/36)	0(40)	186(119/67)	0(100)	18(9/8)
			500	6(3/3)	222(48)	9(3/6)	73(62)	24(10/14)	29(100)	6(3/3)
			1000	4(2/2)	1996(70)	5(2/3)	552(80)	5(2/3)	244(99)	4(2/2)
	Z	Z^2	100	20(11/10)	1(39)	90(55/35)	1(46)	206(104/102)	0(100)	18(9/9)
			500	7(4/3)	251(71)	11(6/6)	91(78)	90(73/17)	28(100)	6(3/3)
			1000	4(2/2)	2166(86)	4(2/2)	572(88)	4(2/2)	242(99)	4(2/2)
	Z^2	Z	100	21(11/10)	1(41)	147(35/112)	1(48)	248(103/145)	0(100)	18(9/9)
			500	7(3/3)	257(69)	10(3/7)	93(79)	104(55/50)	29(100)	6(3/3)
			1000	4(2/2)	2108(86)	4(2/2)	600(88)	6(3/2)	242(99)	4(2/2)
Cens	Z	Z	100	23(12/11)	4(77)	57(46/11)	1(80)	110(50/60)	0(100)	20(11/10)
			500	8(4/4)	128(90)	8(4/4)	58(90)	58(36/22)	23(100)	8(4/4)
			1000	5(3/3)	952(86)	5(3/3)	379(86)	6(3/3)	181(100)	5(3/2)
	Z	Z^2	100	23(12/12)	2(71)	42(12/30)	1(74)	260(163/98)	0(100)	22(11/11)
			500	8(4/4)	129(85)	8(4/4)	58(85)	35(16/19)	21(100)	8(4/4)
			1000	5(3/2)	954(86)	5(3/2)	357(87)	14(3/11)	175(100)	5(3/2)
	Z^2	Z	100	24(11/12)	2(74)	31(11/20)	1(76)	216(107/108)	0(100)	21(10/11)
			500	8(4/4)	124(84)	8(4/4)	59(85)	61(49/12)	22(100)	8(4/4)
			1000	5(2/3)	933(86)	5(2/3)	386(87)	22(5/16)	178(100)	5(3/3)

IQM: iterative quadratic programming; ICM: iterative convex minorant algorithm; PICM: pseudo iterative convex minorant algorithm; PICM.true: One step PICM from true initial value; Comp: complete case; Cens: censoring case (about 30%); IMSE: (total) integrated mean squared error (IMSE for ϕ_1 / IMSE for ϕ_2); Cpu(Cv): computing time in second (convergence percentage).

enrolled between August 2002 and January 2007 and followed up every six months thorough January 2010. Each patient was randomized to a multivitamin that included either a high-dose or low-dose of folic acid (5 or 0 mg), vitamin B6 (50 or 1.4 mg), and vitamin B12 (1000 or 2 microg). The outcome of interest was any of the following nine events: (1) CVD death, (2) myocardial infarction, (3) resuscitated sudden death, (4) stroke, (5) coronary artery revascularization, (6) lower extremity revascularization or amputation above the ankle for severe arterial disease, (7) carotid endarterectomy or angioplasty, (8) abdominal aortic aneurysm repair, or (9) renal artery revascularization. A total of 584 CVD events were observed.

It is of interest to examine the effect of the systolic blood pressure (SBP) and age, both measured at baseline, on CVD. An average age was 52 years with the range from 32 to 84. The average SBP was 136·0 mm Hg with SBP ranging from 70·0 to 247·5. We fit the isotonic proportional hazards model with polynomials 1, α_1SBP+$\alpha_2$age+α_3Trt, and polynomias 2, β_1SBP+β_2SBP^2 + $\beta_3$age+$\beta_4$age^2 + β_5Trt, where Trt is the treatment group with a reference group of the low folic acid. In Figure 5.1, the polynomials 1 show that both SBP and age have significant effect on the risk of CVD with the increase direction ($\hat{\alpha}_1 = 13 \times 10^{-3}, P < 0{\cdot}01$; $\hat{\alpha}_2 = 37 \times 10^{-3}, P < 0{\cdot}01$; $\hat{\alpha}_3 = 27 \times 10^{-3}, P = 0{\cdot}75$). The polynomials 2 also shows the increase direction except that the risk of CVD decreases after age 70 ($\hat{\beta}_1 = 86 \times 10^{-4}, P = 0{\cdot}62$; $\hat{\beta}_2 = 13 \times 10^{-6}, P = 0{\cdot}82$; $\hat{\beta}_3 = 13 \times 10^{-2}, P < 0{\cdot}01$; $\hat{\beta}_4 = -85 \times 10^{-5}, P = 0{\cdot}049$; $\hat{\beta}_5 = 28 \times 10^{-3}, P = 0{\cdot}73$). A parametric polynomial model, however, may be too simplistic to capture nonlinear effects of SBP and age on CVD.

Alternatively, we fit the additive isotonic proportional hazards model only assuming that the risk of CVD is monotonically increasing in SBP and age. First, we fit a univariate isotonic proportional hazards model with SBP, and separately with age, adjusted for the treatment group. Second, we fit the additive isotonic proportional hazards model including both SBP and age and adjusted for the treatment group using the algorithm described in Subsection 5.3.2. The anchor points were set to the maximum value for both covariates: 247·5 for SBP and 84 for age.

Figure 5.1 displays the estimated isotonic functions from univariate and additive isotonic proportional hazards models. Black dots are the values associated with observed CVD, which are potential jump points. The risk of CVD increases gradually in SBP in the univariate isotonic proportional hazards model. On the other hand, the additive isotonic proportional hazards model gives constant CVD risk in SBP up to 157 mm Hg with three possible cut-point values around 160, 175 and 190 mm Hg. The constant risk of SBP before 157 mm Hg agrees with Port et al. (2000) where the risk of CVD mortality was constant before the lower 70% of the SBP, with a sharp increase at the upper 20% of the systolic blood pressure ((lower 70%, upper 20%): (141, 148) for males and (142, 151) for females, respectively, for 45-54 years; (148, 159) for males and (158, 167) for females, respectively, for 55-64 years; (158, 169) for males and (168, 177) for females, respectively, for 65-74 years). In addition, the identified cut-point values agree with the American Heart Association recommendation (as shown in the American Heart Association webpage titled Understanding Blood Pressure Readings) that suggested three stages of hypertension 1, hypertension 2 and hypertensive crisis at the systolic blood pressure 140, 160 and 180, respectively. According to the additive isotonic proportional hazards model, people older than 60 have a constant risk of CVD after adjusting for SBP.

The effect of treatment was not significant ($\exp(\hat{\alpha}_3)$ = 1·027, P = 0·75; $\exp(\hat{\beta}_5)$ = 1·029, P = 0·73; Hazard ratio=1·027 from the additive isotonic proportional hazards model).

5.6 Technical Details for Chapter 5

Proof of Lemma 5.1 Since a sum of convex functions is a convex function, we can prove the convexity of $lpl^N(\boldsymbol{\psi})$ in (5.4) by showing that the left and right parts of $lpl^N(\boldsymbol{\psi})$ are

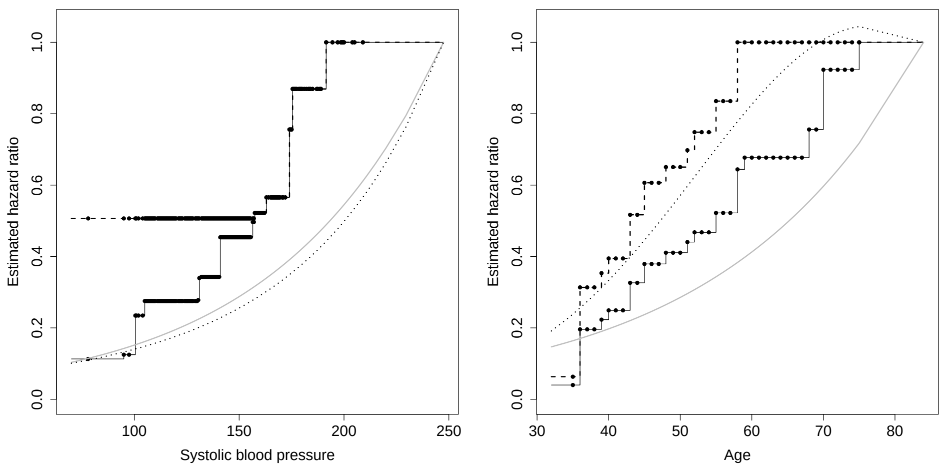

Figure 5.1: FAVORIT study: Estimated hazard ratios. Univariate isotonic proportional hazards model (black solid), additive isotonic proportional hazards model (dashed) and standard additive proportional hazards mode with polynomials of 1 (grey solid) and 2 (dot dashed). Left is for systolic blood pressure. Right is for age. The circles indicate CVD events.

convex, respectively. It is obvious that the left part is convex because it is a linear function. The right part is convex because it is a composition of strictly increasing convex functions.

Proof of Theorem 5.2 The proof is directly followed from the theorem 1 (Bacchetti 1989) by showing the following three conditions. First, $lpl^N(\boldsymbol{\psi})$ is a convex function by Lemma 5.1. Second, Ψ is a convex and compact cone, because each $\Psi_j^{k_j}$ is a convex and compact cone, $j = 1, \ldots, p$. Last, for each univariate optimization, $l^N(\boldsymbol{\psi}_j)$ has a unique minimizer over $\Psi_j^{k_j}$ by Theorem 3.1.

Proof of Proposition 5.3 Under censored data, the partial likelihood in (5.2) reduces to

$$pl(\boldsymbol{\psi}) = \prod_{i=1}^{n^\star} \prod_{t\geq 0} \left\{ \frac{e^{\psi_1(Z_{i1}^\star)+\cdots+\psi_p(Z_{ip}^\star)}}{\sum_{s=1}^n Y_s(t) e^{\psi(Z_{s1})+\cdots+\psi(Z_{sp})}} \right\}^{dN_i^\star(t)}. \tag{5.7}$$

In (5.7), the parameters for uncensored subjects are included in both numerator and denominator, while the parameters for censored subjects are only included in the denominator. Thus, the partial likelihood is maximized when the parameters for uncensored subjects are minimized, which occur among the parameters for uncensored subjects because of the order restriction. When $Z_{hj} < Z^{\star}_{(1)j}$ for $h = 1, \ldots, n$, ψ_{hj} is set to $\psi^{\star}_{(1)j}$ for each $j = 1, \ldots, p$ by the assumption. It shows that the isotonic estimator jumps only at $\boldsymbol{Z}^{\star}_i$, $i = 1, \ldots, n^{\star}$.

Proof of Proposition 5.4 The proof is analogous to that of Proposition 5.3. Under censored data with time-dependent covariates, the partial likelihood in (5.6) reduces to

$$\prod_{i=1}^{n^{\star}} \prod_{t \geq 0} \left\{ \frac{e^{\psi_1(Z^*_{i1}) + \cdots + \psi_p(Z^*_{ip})}}{\sum_{s=1}^{n} Y_s(t) e^{\psi_1\{Z_{s1}(t)\} + \cdots + \psi_p\{Z_{sp}(t)\}}} \right\}^{dN^*_i(t)} \tag{5.8}$$

In (5.8), the parameters at $\boldsymbol{Z}^*_1, \ldots, \boldsymbol{Z}^*_n$ are included in both numerator and denominator, while the other parameters are only included in the denominator. Thus, the partial likelihood is maximized, when the other parameters are minimized, which occur among the parameters in the numerator because of the order restriction. When $Z_{hj}(X_i) < Z^*_{(1)j}$ for $i, h = 1, \ldots, n$, $\psi_j(Z_{hj})$ is set to $\psi_j(Z^*_{1j})$ for each $j = 1, \ldots, p$ by the assumption. It shows that the isotonic estimator jumps only at $\boldsymbol{Z}^*_i$, $i = 1, \ldots, n^{\star}$.

CHAPTER 6: SUMMARY AND FUTURE RESEARCH

In this dissertation, we have studied order restricted inference for survival data analysis, where a hazard function has an order restriction on continuous covariates. In Chapter 3, we proposed the isotonic proportional hazards model with an algorithm that can handle large datasets. The proposed model captured a nonlinear and isotonic effect of a covariate for a hazard function, with completely unspecified baseline hazard function. In Chapter 4, we proposed a shape restricted additive hazard model. This model is particularly useful when a unimodal hazard function with an unknown mode is being estimated. We also proposed the quadratic pool-adjacent-violators algorithm to use when a standard quadratic programming may be computationally limiting in this model. In Chapter 5, we generalized the isotonic proportional hazard model to include multiple continuous covariates under the additive structure. We developed and efficient way to compute the estimates by combining the pseudo iterative convex minorant algorithm and the cycling algorithm.

Classical order restricted inference has focused on independent and identically distributed observations, where likelihood functions are separable in terms of observed covariate values. Under the separable structure of the likelihood function, large sample properties were well-studied. An efficient computation was also developed to compute isotonic estimator, such as the pool-adjacent-violators algorithm. The order restricted inference has been extended to the survival data analysis, where the hazard function was isotonic in time. In this case, the likelihood function is separable in terms of observed time points, and therefore, the isotonic regression techniques can be easily extended. We investigated the estimation of the order restricted hazard function in continuous covariate without making any assumptions on relationship with time. The partial likelihood, therefore, does not have a separable structure in terms of the observed covariate values or time points. This brings a number of challenges.

One challenge is to establish the consistency and asymptotic properties of the isotonic partial likelihood estimator in Chapter 3. We conjectured that the isotonic estimator converged to the Chernoff distribution, as supported by our simulation studies in Section 3.6. A rigorous proof is not yet available as the log partial likelihood is not a sum of independent terms and therefore it is not clear how to apply existing theory developed for the case of independent terms. For the same reason, it was a challenge to establish the consistency property for the mode estimator from the shape restricted additive hazard function in Chapter 4. In Chapter 5, we studied isotonic hazard function in multiple covariates under the additive isotonic structure. This model assumes independent effects of the covariates but can be extended to include interaction terms.

REFERENCES

Ancukiewicz, M., Finkelstein, D. M., and Schoenfeld, D. A. (2003), "Modelling the relationship between continuous covariates and clinical events using isotonic regression," *Statistics in Medicine*, 22, 3151–3159.

Aragón, J. and Eberly, D. (1992), "On convergence of convex minorant algorithms for distribution estimation with interval-censored data," *Journal of Computational and Graphical Statistics*, 1, 129–140.

Ayer, M., Brunk, H., Ewing, G., Reid, W., and Silverman, E. (1955), "An empirical distribution function for sampling with incomplete information," *The Annals of Mathematical Statistics*, 26, 641–647.

Bacchetti, P. (1989), "Additive isotonic models," *Journal of the American Statistical Association*, 84, 289–294.

Banerjee, M. (2007), "Likelihood based inference for monotone response models," *The Annals of Statistics*, 35, 931–956.

— (2008), "Estimating monotone, unimodal and U-shaped failure rates using asymptotic pivots," *Statistica Sinica*, 18, 467–492.

Berbari, A. E. and Manci, G. (2010), *Cardiorenal Syndrome: Mechanisms, Risk and Treatment*, Springer.

Bostom, A. G., Carpenter, M. A., Kusek, J. W., Levey, A. S., Hunsicker, L., Pfeffer, M. A., Selhub, J., Jacques, P. F., Cole, E., Gravens-Mueller, L., et al. (2011), "Homocysteine-Lowering and Cardiovascular Disease Outcomes in Kidney Transplant Recipients Primary Results From the Folic Acid for Vascular Outcome Reduction in Transplantation Trial," *Circulation*, 123, 1763–1770.

Breslow, N. (1972), "Discussion of the paper by D. R. Cox," *Journal of the Royal Statistical Society, Series B*, 34, 216–217.

Cheng, G. (2009), "Semiparametric additive isotonic regression," *Journal of Statistical Planning and Inference*, 139, 1980–1991.

Cox, D. R. (1972), "Regression models and life-tables (with discussion)," *Journal of the Royal Statistical Society. Series B*, 34, 187–220.

Gevers, M., Hack, W., Ree, E., Lafeber, H., and Westerhof, N. (1993), "Calculated mean arterial blood pressure in critically ill neonates," *Basic Research in Cardiology*, 88, 80–85.

Gorst-Rasmussen, A. and Scheike, T. H. (2012), "Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model," *Journal of Statistical Software*, 47, 1–17.

Grenander, U. (1956), "On the theory of mortality measurement, Part II," *Scandinavian Actuarial Journal*, 1956, 125–153.

Groeneboom, P. (1996), "Lectures on inverse problems. *Lectures on Probability Theory and Statistics: Ecole d'Ete de Probabilites de Saint Flour XXIV -1994. Lecture Notes in Math.*" 1648, 67–164. Springer, Berlin.

Groeneboom, P. and Wellner, J. (1992), "*Information Bounds and Nonparametric Maximum Likelihood Estimation, Part II.*" 19, 35–121, Basel: Birkhauser.

Groeneboom, P. and Wellner, J. A. (2001), "Computing Chernoff's distribution," *Journal of Computational and Graphical Statistics*, 10, 388–400.

Huang, J. and Wellner, J. A. (1995), "Estimation of a monotone density or monotone hazard under random censoring," *Scandinavian Journal of Statistics*, 22, 3–33.

Jamieson, D. J., Chasela, C. S., Hudgens, M. G., King, C. C., Kourtis, A. P., Kayira, D., Hosseinipour, M. C., Kamwendo, D. D., Ellington, S. R., Wiener, J. B., et al. (2012), "Maternal and infant antiretroviral regimens to prevent postnatal HIV-1 transmission: 48-week follow-up of the BAN randomised controlled trial," *The Lancet*, 379, 2449–2458.

Jongbloed, G. (1998), "The iterative convex minorant algorithm for nonparametric estimation," *Journal of Computational and Graphical Statistics*, 7, 310–321.

Kaplan, E. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Kosorok, M. R. (2007), *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer.

Lin, D. and Ying, Z. (1994), "Semiparametric analysis of the additive risk model," *Biometrika*, 81, 61–71.

Lopuhaä, H. P. and Nane, G. F. (2013), "Shape Constrained Non-parametric Estimators of the Baseline Distribution in Cox Proportional Hazards Model," *Scandinavian Journal of Statistics*, 40, 619–646.

Mammen, E. and Yu, K. (2007), "Additive isotone regression," *Institute of Mathematical Statistics Lecture Notes-Monograph Series: Asymptotics: Particles, Processes and Inverse Problems*, 55, 179–195.

Marshall, A. W. and Proschan, F. (1965), "Maximum likelihood estimation for distributions with monotone failure rate," *The Annals of Mathematical Statistics*, 36, 69–77.

Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H. O., and Binks, K. (2000), "Additive isotonic regression models in epidemiology," *Statistics in medicine*, 19, 849–859.

Mukerjee, H. and Wang, J.-L. (1993), "Nonparametric maximum likelihood estimation of an increasing hazard rate for uncertain cause-of-death data," *Scandinavian Journal of Statistics*, 17–33.

Parzen, M. and Lipsitz, S. R. (1999), "A Global Goodness-of-Fit Statistic for Cox Regression Models," *Biometrics*, 55, 580–584.

Port, S., Demer, L., Jennrich, R., Walter, D., and Garfinkel, A. (2000), "Systolic blood pressure and mortality," *The Lancet*, 355, 175–180.

Rao, B. P. (1970), "Estimation for distributions with monotone failure rate," *The Annals of Mathematical Statistics*, 41, 507–519.

Robertson, T., Wright, F., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, Chichester: Wiley.

Shoung, J.-M. and Zhang, C.-H. (2001), "Least squares estimators of the mode of a unimodal regression function," *The Annals of Statistics*, 29, 648–665.

van der Vaart, A. and Wellner, J. A. (2000), "High dimensional probability II. *Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes*." 47, 115–133. Springer.

Wellner, J. A. and Zhan, Y. (1997), "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data," *Journal of the American Statistical Association*, 92, 945–959.

Wellner, J. A. and Zhang, Y. (2000), "Two estimators of the mean of a counting process with panel count data," *The Annals of Statistics*, 28, 779–814.