A new MM algorithm for constrained estimation in the proportional hazards model

Jieli DING[a], **Guo-Liang TIAN**[b,*] and **Kam Chuen YUEN**[b]

[a] *School of Mathematics and Statistics,*
Wuhan University, Wuhan, Hubei 430072, P.R.China
[b] *Department of Statistics and Actuarial Science,*
The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China
[*] Corresponding author's email: gltian@hku.hk
3 April 2013, submitted; 20 July 2014, the first revision; 8 November 2014, the second revision

Abstract: In this paper, we study the constrained estimation in Cox's model for the right-censored survival data and derive asymptotic properties of the constrained estimator by using the Lagrangian method based on Karush–Kuhn–Tucker conditions. A novel *minorization–maximization* (MM) algorithm is developed for calculating the maximum likelihood estimates of the regression coefficients subject to box or linear inequality restrictions in the proportional hazards model. The first M-step of the proposed MM algorithm is to construct a surrogate function with a diagonal Hessian matrix, which can be reached by utilizing the convexity of the exponential function and the negative logarithm function. The second M-step is to maximize the surrogate function with a diagonal Hessian matrix subject to box constraints, which is equivalent to separately maximizing several one-dimensional concave functions with a lower bound and an upper bound constraint, resulting in an explicit solution via a median function. The ascent property of the proposed MM algorithm under constraints is theoretically justified. Standard error estimation is also presented via a non-parametric bootstrap approach. Simulation studies are performed to compare the estimations with and without constraints. Two real data sets are used to illustrate the proposed methods.

Keywords: Asymptotic properties; Bootstrap approach; Constrained estimation; Karush–Kuhn–Tucker conditions; MM algorithm; Proportional hazards model.

1. Introduction

Survival data arise in a number of fields such as reliability engineering, economics, sociology, public health, epidemiology and medicine (especially, clinical trials). Survival analysis is used to model the relationship between the time-to-event (e.g., death or disease) and a set of covariates or predictors. When the period of observation expires, or an individual is removed from or drops out the study prior to the event occurs, survival data are considered as right-censored. The proportional hazards model originally introduced by Cox (1972) may be the most widely used method for analyzing survival data with censoring. Since the publication of Cox (1972), numerous extensions and developments in various aspects have been proposed during the past 40 years by many authors including Cox (1975), Andersen and Gill (1982), Bickel *et al.* (1993), Lin and Ying (1993), Lin (1994), Huang (1996), Chen and Little (1999), and Chen and Lo (1999). A comprehensive review was given by Kalbfleisch and Prentice (2002).

In many practical problems, it may be available as prior information that restrictions on some model parameters would result in a more reasonable interpretation. Such restrictions cannot be ignored; otherwise the statistical inference may be misled and an underestimate of the effect may be caused (Tan *et al.*, 2005; Fang *et al.*, 2006). Therefore, it is reasonable to expect that the analysis would perform better if parameter constraints are taken into account in the modeling process. However, the complication resulting from such restrictions raises statistical challenges. Statistical inferences on constrained problems have been studied by many authors (e.g., Liew, 1976; Nyquist, 1991; Silvapulle, 1997). For example, Wang (1996, 2000) studied asymptotic properties of constrained estimators in nonlinear regressions. Moore and Sadler (2006) and Moore *et al.* (2008) discussed the asymptotic theory for the constrained *maximum likelihood estimator* (MLE) and presented a constrained Cramér–Rao bound. However, to our knowledge, asymptotic properties of constrained estimators for the regression coefficients in Cox's model have never been studied.

The first objective of this paper is to derive two asymptotic properties of the constrained MLEs for the regression coefficients in the proportional hazards model with right-censored data. These asymptotic results are useful in statistical inferences for Cox's model with box and linear inequality constraints. We use the Karush–Kuhn–Tucker conditions, a well-known approach in optimization with inequality constraints, to overcome the difficulty caused by the constraint. Similar techniques were adopted by Wang (2000), Xu and Wang (2008) for constrained least-squares estimator, and Moore and Sadler (2006), Moore *et al.* (2008) for constrained Cramér–Rao bound in parametric models.

Böhning and Lindsay (1988) developed a *quadratic lower bound* (QLB) algorithm with monotone convergence like the EM algorithm for the Cox model without constraints. Since the construction of the quadratic surrogate function in the QLB algorithm is based on the second-order Taylor expansion of the partial log-likelihood function in the neighborhood of the maximum likelihood estimate, this QLB algorithm cannot be applied to the Cox model with box and/or linear inequality constraints. Note that the QLB algorithm is a special case of *minorization–maximization* (MM) algorithms (Becker *et al.*, 1997; Hunter and Lange, 2004; Lange 2004, 2010). In addition, the existing MM algorithms (Lange *et al.*, 2000) such as De Pierro's algorithm (De Pierro, 1995) cannot be applied to the Cox model even for the case without constraints. Hunter and Lange (2002) proposed an MM algorithm for finding the MLEs of the regression coefficients in the semiparametric proportional odds model just for the case without constraints.

Thus, the second objective of this paper is to develop a novel MM algorithm for calculating the MLEs of the regression coefficients with box or linear inequality restrictions in the proportional hazards model. The key to the proposed MM algorithm is to construct a surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ with a diagonal Hessian matrix, which can be reached by utilizing the convexity of the exponential function e^x and the negative logarithm function $-\log x$. Maximizing this surrogate function with a diagonal Hessian matrix subject to box constraints is equivalent to separately maximizing several one-dimensional concave functions with a lower bound and an upper bound constraint, which has an explicit solution via a median function.

The rest of the article is organized as follows. In Section 2, we formulate the proportional hazards model with constraints and derive two asymptotic properties for the constrained estimator. In Section 3, we develop a new MM algorithm for calculating the constrained estimation in Cox's model. The ascent property of the proposed MM algorithm under constraints is derived. Standard error estimation is also presented via a non-parametric bootstrap approach. We conduct several simulation studies in Section 4 to compare the estimations with and without constraints. In Section 5, two real data sets are used to illustrate the proposed methods. A discussion is presented in Section 6. Detailed proofs on asymptotic properties are put in the Appendix.

2. Constrained estimation in Cox's model

2.1 The formulation of the constrained Cox model

Consider Cox's proportional hazards model with constrained regression coefficients. Suppose that there are n subjects drawn randomly from the population of interest. For the i-th subject ($i = 1, \ldots, n$), let $\widetilde{T}_i$, C_i, $T_i = \min(\widetilde{T}_i, C_i)$ and $\mathbf{Z}_i$ denote the failure time, the censoring time, the observed time and the covariate vector of p dimension, respectively. We assume that given the covariate vector $\mathbf{Z}_i$, the failure time $\widetilde{T}_i$ and the censoring time C_i are conditionally independent. Furthermore, let $\Delta_i = I(T_i \leqslant C_i)$, $Y_i(t) = I(T_i \geqslant t)$ and $N_i(t) = \Delta_i I(T_i \leqslant t)$ respectively denote the right censoring indicator, at-risk process and counting process for subject i, where $I(\cdot)$ is the indicator function.

The proportional hazards model specifies the hazard function of the failure time conditional on covariates taking the following form:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}^\top \mathbf{Z}(t)\}, \tag{2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\mathbf{Z}(t)$ is a p-dimensional vector of time-varying covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional vector of regression coefficients. We are interested in estimating the unknown parameter vector $\boldsymbol{\beta}$ subject to the following equality and inequality constraints:

$$\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = \{\boldsymbol{\beta}\colon \;\; \boldsymbol{f}(\boldsymbol{\beta}) = \mathbf{0}_r,\; \boldsymbol{g}(\boldsymbol{\beta}) \leqslant \mathbf{0}_s\}, \tag{2.2}$$

where both $\boldsymbol{f}(\boldsymbol{\beta}) = (f_1(\boldsymbol{\beta}), \ldots, f_r(\boldsymbol{\beta}))^\top$ and $\boldsymbol{g}(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta}), \ldots, g_s(\boldsymbol{\beta}))^\top$ are assumed to have continuous second-order partial derivatives, and functional constraints are consistent; i.e. $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ is a non-empty convex set. Note that if $r = 0$, then there is no equality constraint; if $s = 0$, there is no inequality constraint.

It is well known that the inference on $\boldsymbol{\beta}$ can be based on the partial likelihood function (Cox, 1972)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_i(T_i)}}{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)}} \right)^{\Delta_i},$$

for which the corresponding partial log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left\{ \boldsymbol{\beta}^\top \mathbf{Z}_i(T_i) - \log\left(\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)} \right) \right\}. \tag{2.3}$$

The constrained MLE $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})} \ell(\boldsymbol{\beta}), \tag{2.4}$$

which can be viewed as the solution to the following constrained optimization problem:

$$\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \quad \text{subject to} \quad \boldsymbol{f}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{and} \quad \boldsymbol{g}(\boldsymbol{\beta}) \leqslant \mathbf{0}. \tag{2.5}$$

By allowing inequality constraints, the Karush–Kuhn–Tucker (KKT) approach to constrained optimizations generalizes the method of Lagrange multipliers, which allows only equality constraints. By using the KKT approach, we can build the Lagrangian of the optimization problem in (2.5) as

$$H(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \ell(\boldsymbol{\beta}) + \sum_{j=1}^{r} \mu_j f_j(\boldsymbol{\beta}) + \sum_{k=1}^{s} \nu_k g_k(\boldsymbol{\beta}), \tag{2.6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)^\top$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_s)^\top$ are KKT multipliers. Any potential solution to (2.5) must be a stationary point of (2.6); that is, it must be a point $\boldsymbol{\beta}^*$ satisfying the following KKT necessary conditions (Boyd and Vandenberghe, 2004; Madsen *et al.*, 2004):

$$\begin{cases} \text{Stationarity:} & \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^*) + \sum_{j=1}^{r} \mu_j^* \nabla_{\boldsymbol{\beta}} f_j(\boldsymbol{\beta}^*) + \sum_{k=1}^{s} \nu_k^* \nabla_{\boldsymbol{\beta}} g_k(\boldsymbol{\beta}^*) = \mathbf{0}, \\ \text{Primal feasibility:} & f_j(\boldsymbol{\beta}^*) = 0, \qquad \forall\, j = 1, \ldots, r, \\ & g_k(\boldsymbol{\beta}^*) \leqslant 0, \qquad \forall\, k = 1, \ldots, s, \\ \text{Dual feasibility:} & \nu_k^* \geqslant 0, \qquad \forall\, k = 1, \ldots, s, \\ \text{Complementary slackness:} & \nu_k^* g_k(\boldsymbol{\beta}^*) = 0, \quad \forall\, k = 1, \ldots, s, \end{cases} \tag{2.7}$$

where

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^*) \mathrel{\hat=} \left. \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^*}.$$

The k-th inequality constraint $g_k(\boldsymbol{\beta}) \leqslant 0$ is referred to be *active* at a feasible point $\boldsymbol{\beta}$ if $g_k(\boldsymbol{\beta}) = 0$; otherwise it is called *inactive*. For an equality constraint $f_j(\boldsymbol{\beta}^*) = 0$, the μ_j^* can have any sign. For an active inequality constraint, we have $g_k(\boldsymbol{\beta}^*) = 0$ and $\nu_k^* \geqslant 0$. For an inactive inequality constraint $g_k(\boldsymbol{\beta}^*) < 0$, we must have $\nu_k^* = 0$ to confirm the last s equations in (2.7), indicating that these inactive inequality constraints have no influence on the stationarity equation. Due to these remarks, the stationarity equation in (2.7) can be simplified by incorporating all active constraints into $\boldsymbol{h}(\boldsymbol{\beta}) = (h_1(\boldsymbol{\beta}), \ldots, h_q(\boldsymbol{\beta}))^\top$. Then, we can rewrite the stationarity equation as

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^*) + \boldsymbol{\mu}^{*\top} \mathbf{F}(\boldsymbol{\beta}^*) = \mathbf{0}, \qquad \boldsymbol{\beta}^* \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}), \tag{2.8}$$

where $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_q^*)^\top$, $\mathbf{F}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \boldsymbol{h}(\boldsymbol{\beta})$ is the $q \times p$ gradient matrix of all q active constraints and have full-row rank. By using the Lagrangian method based on KKT conditions, the asymptotic properties of constrained estimator $\hat{\boldsymbol{\beta}}$ can be established.

2.2 Asymptotic properties of the constrained estimator

We now study asymptotic properties of the constrained estimator $\hat{\boldsymbol{\beta}}$. Inspired with the methods developed by Moore and Sadler (2006) and Moore *et al.* (2008), we introduce a $p \times (p-q)$ matrix $\mathbf{U}(\boldsymbol{\beta})$ whose elements being continuous of $\boldsymbol{\beta}$ such that for each $\boldsymbol{\beta}$,

$$\mathbf{F}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta}) = 0 \quad \text{and} \quad \mathbf{U}(\boldsymbol{\beta})^\top \mathbf{U}(\boldsymbol{\beta}) = \mathbf{I}_{p-q}. \tag{2.9}$$

That is, the columns of $\mathbf{U}(\boldsymbol{\beta})$ form an orthonormal null space of the range space of the row vectors in $\mathbf{F}(\boldsymbol{\beta})$. To present asymptotic results, we define

$$\mathbf{S}^{(j)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{l=1}^{n} Y_l(t) \mathbf{Z}_l(t)^{\otimes j} \, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(t)}, \quad j = 0, 1, 2,$$

where $\boldsymbol{x}^{\otimes 0} = \mathbf{1}$, $\boldsymbol{x}^{\otimes 1} = \boldsymbol{x}$ and $\boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\boldsymbol{x}^\top$ for a vector $\boldsymbol{x}$. Let $\boldsymbol{\beta}^0$ be the true parameter vector, and τ be the stoping time for the survival study. We make the following assumptions throughout this paper.

(A1) The parameter space $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ is a compact and convex set, and the space of covariate, $\mathbb{Z}$, is also compact;

(A2) $\int_0^\tau \lambda_0(t) \, \mathrm{d}t < \infty$;

(A3) There exists a positive number δ such that

$$\frac{1}{\sqrt{n}} \sup_{l=1,\ldots,n;\ t \in [0,\tau]} |\mathbf{Z}_l(t)| Y_l(t) I\left(\boldsymbol{\beta}^{0\top} \mathbf{Z}_l(t) > \delta |\mathbf{Z}_l(t)|\right) \xrightarrow{\mathrm{P}} 0;$$

(A4) There exist three matrix functions $\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)$, $\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t)$ and $\boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t)$ defined on $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) \times [0, \tau]$ which satisfy the following conditions:

  (a) $\sup_{\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g});\ t \in [0,\tau]} \|\mathbf{S}^{(j)}(\boldsymbol{\beta}, t) - \boldsymbol{s}^{(j)}(\boldsymbol{\beta}, t)\| \xrightarrow{\mathrm{P}} 0$, $j = 0, 1, 2$;

  (b) For $j = 0, 1, 2$, the functions $\boldsymbol{\beta} \to \boldsymbol{s}^{(j)}(\boldsymbol{\beta}, t)$ are continuous on $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ uniformly in $t \in [0, \tau]$, and the equalities $\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t) = \nabla_{\boldsymbol{\beta}} \boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)$ and $\boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t) = \nabla_{\boldsymbol{\beta}}^2 \boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)$ hold for any $\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ and $t \in [0, \tau]$;

(c) $\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from zero for any $\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ and $t \in [0, \tau]$;

(d) The matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \int_0^\tau \boldsymbol{v}(\boldsymbol{\beta}, t)\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)\lambda_0(t)\,\mathrm{d}t$$

is positive definite at $\boldsymbol{\beta}^0$, where

$$\boldsymbol{v}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t)}{\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t)}{\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2}.$$

Under these conditions, we have the following results with detailed proofs given in Appendix.

Theorem 1 (Consistency). *Under Assumptions (A1)–(A4), we have the consistency of the constrained estimator* $\hat{\boldsymbol{\beta}}_n$*; that is,* $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\beta}^0$ *as* $n \to \infty$.

Theorem 2 (Asymptotic normality). *Under Assumptions (A1)–(A4), we have the asymptotic normality of* $\hat{\boldsymbol{\beta}}_n$*; that is,* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) \xrightarrow{\mathrm{D}} N_p(\mathbf{0},\ \boldsymbol{\Omega}(\boldsymbol{\beta}^0))$*, where the asymptotic variance matrix is* $\boldsymbol{\Omega}(\boldsymbol{\beta}^0) = \mathbf{U}(\boldsymbol{\beta}^0)\left[\mathbf{U}(\boldsymbol{\beta}^0)^\top\boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\mathbf{U}(\boldsymbol{\beta}^0)\right]^{-1}\mathbf{U}(\boldsymbol{\beta}^0)^\top$.

3. A new MM algorithm for constrained estimation in Cox's model

In Section 2, we obtained the large-sample theory for the constrained estimator $\hat{\boldsymbol{\beta}}$ of the regression coefficients in Cox's model by using the Lagrangian method based on KKT conditions. However, in general, the computation of the optimization problem in (2.4) are very complicated, which motivates the development of efficient algorithms to obtain the solution to the constrained optimization problem. In this section, we propose a new MM algorithm for the computation of the constrained estimation in Cox's model. We transfer the constrained problem for maximizing the partial log-likelihood function $\ell(\boldsymbol{\beta})$ in (2.4) to maximizing a surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ with a diagonal Hessian matrix subject to the constraints $\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$, which has an explicit solution via a median function.

3.1 Construction of the surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$

Since $x \to \mathrm{e}^x$ is a convex function, for any positive weights $\{\alpha_k\}_{k=1}^p$ satisfying $\sum_{k=1}^p \alpha_k = 1$, we have

$$\exp\left(\sum_{k=1}^p \alpha_k x_k\right) \leqslant \sum_{k=1}^p \alpha_k \exp(x_k). \tag{3.1}$$

Let $\mathbf{Z}_l(T_i) = (Z_{l1}(T_i), \ldots, Z_{lp}(T_i))^\top$ and $\boldsymbol{\beta}^{(m)} = (\beta_1^{(m)}, \ldots, \beta_p^{(m)})^\top$ denote the m-th approximation of the constrained MLE $\hat{\boldsymbol{\beta}}$ defined in (2.4). For the logarithm term in the partial log-likelihood function (2.3), we have

$$
\begin{aligned}
& \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)} \\
= \;& \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)}\, \mathrm{e}^{(\boldsymbol{\beta}-\boldsymbol{\beta}^{(m)})^\top \mathbf{Z}_l(T_i)} \\
= \;& \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)} \exp\left\{ \sum_{k=1}^{p} \lambda_{lk} \left[\lambda_{lk}^{-1} Z_{lk}(T_i)(\beta_k - \beta_k^{(m)}) \right] \right\} \\
\overset{(3.1)}{\leqslant} \;& \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)} \left\{ \sum_{k=1}^{p} \lambda_{lk} \exp\left[\lambda_{lk}^{-1} Z_{lk}(T_i)(\beta_k - \beta_k^{(m)}) \right] \right\} \\
= \;& \sum_{l=1}^{n} Y_l(T_i) \left\{ \sum_{k=1}^{p} \lambda_{lk} \exp\left[\lambda_{lk}^{-1} Z_{lk}(T_i)(\beta_k - \beta_k^{(m)}) + \boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i) \right] \right\},
\end{aligned} \tag{3.2}
$$

for all $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(m)} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ and arbitrary positive weights $\{\lambda_{lk}\}_{k=1}^{p}$. In practice, similar to the suggestion in Becker *et al.* (1997), for each $l \in \{1, \ldots, n\}$, we can set

$$
\lambda_{lk} = \frac{|Z_{lk}(T_i)|}{\sum_{k'=1}^{p} |Z_{lk'}(T_i)|}, \quad k = 1, \ldots, p. \tag{3.3}
$$

If $Z_{lk}(T_i) = 0$, then $\lambda_{lk}^{-1} \mathrel{\hat{=}} 0$.

Furthermore, in the well-known inequality $-\log x \geqslant 1 - \log y - x/y$, let

$$
x = \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)} \quad \text{and} \quad y = \sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)},
$$

we obtain

$$
\begin{aligned}
& -\log\left[\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)} \right] \\
\geqslant \;& 1 - \log\left[\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)} \right] - \frac{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(T_i)}}{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)}} && (3.4) \\
\overset{(3.2)}{\geqslant} \;& 1 - \log\left[\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)} \right] \\
& - \frac{\sum_{l=1}^{n} \sum_{k=1}^{p} Y_l(T_i) \lambda_{lk} \exp\left[\lambda_{lk}^{-1} Z_{lk}(T_i)(\beta_k - \beta_k^{(m)}) + \boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i) \right]}{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top} \mathbf{Z}_l(T_i)}}. && (3.5)
\end{aligned}
$$

Based on the inequality (3.5), for a given vector $\boldsymbol{\beta}^{(m)} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$, we define a surrogate function

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = c_0 + \sum_{i=1}^{n} \Delta_i \left[\boldsymbol{\beta}^\top \mathbf{Z}_i(T_i) - \frac{\sum_{l=1}^{n} \sum_{k=1}^{p} Y_l(T_i) \lambda_{lk}\, g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)})}{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}} \right], \quad \boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}), \tag{3.6}$$

where

$$c_0 = \sum_{i=1}^{n} \Delta_i \{1 - \log[\textstyle\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}]\} \tag{3.7}$$

is a constant independent of $\boldsymbol{\beta}$ and

$$g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)}) = \exp[\lambda_{lk}^{-1} Z_{lk}(T_i)(\beta_k - \beta_k^{(m)}) + \boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)]. \tag{3.8}$$

By combining (2.3), (3.5) and (3.6), we immediately obtain

$$\ell(\boldsymbol{\beta}) \geqslant Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}), \quad \forall\, \boldsymbol{\beta},\, \boldsymbol{\beta}^{(m)} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}). \tag{3.9}$$

In addition, it is noted that the equality case in the inequalities (3.2) and (3.5) holds if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$, resulting in $\ell(\boldsymbol{\beta}^{(m)}) = Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})$. Therefore, we can establish the following MM algorithm for the case with constraints:

$$\boldsymbol{\beta}^{(m+1)} = \arg \max_{\boldsymbol{\beta} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}). \tag{3.10}$$

The first partial derivatives of $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ with respect to the k-th component of $\boldsymbol{\beta}$ are

$$\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k} = \sum_{i=1}^{n} \Delta_i \left[Z_{ik}(T_i) - \frac{\sum_{l=1}^{n} Y_l(T_i) Z_{lk}(T_i)\, g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)})}{\sum_{l=1}^{n} Y_l(T_i) \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}} \right], \quad k = 1, \ldots, p.$$

Note that $g_{lk}(\beta_k^{(m)}|\boldsymbol{\beta}^{(m)}) = \exp[\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)]$, we have

$$\left.\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} = \sum_{i=1}^{n} \Delta_i \left[Z_{ik}(T_i) - \frac{\sum_{l=1}^{n} Y_l(T_i) Z_{lk}(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}{\sum_{l=1}^{n} Y_l(T_i) \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}} \right], \quad k = 1, \ldots, p.$$

In the vector form, we obtain the score vector of $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ as

$$\frac{\partial Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \Delta_i \left[\mathbf{Z}_i(T_i) - \frac{\sum_{l=1}^{n} Y_l(T_i) \mathbf{Z}_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}{\sum_{l=1}^{n} Y_l(T_i) \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}} \right]. \tag{3.11}$$

Similarly, the second partial derivatives of $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ are given by

$$\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k^2} &= -\sum_{i=1}^{n} \Delta_i \frac{\sum_{l=1}^{n} Y_l(T_i)[Z_{lk}^2(T_i)/\lambda_{lk}]\, g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)})}{\sum_{l=1}^{n} Y_l(T_i) \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}, \quad k = 1, \ldots, p, \\
\frac{\partial^2 Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k \partial \beta_{k'}} &= 0, \quad k \neq k'; \quad k,\, k' = 1, \ldots, p.
\end{aligned}$$

The negative Hessian matrix of $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ is a diagonal matrix, i.e.,

$$-\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \mathrm{diag}\left(-\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial\beta_1^2}, \ldots, -\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial\beta_p^2}\right), \tag{3.12}$$

where

$$-\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial\beta_k^2} = \sum_{i=1}^n \Delta_i \frac{\sum_{l=1}^n Y_l(T_i)[Z_{lk}^2(T_i)/\lambda_{lk}]\,\mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}{\sum_{l=1}^n Y_l(T_i)\,\mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}, \quad k = 1, \ldots, p.$$

3.2 Derivation of the MLE for a class of linear inequality constraints

Consider a class of linear inequality constraints of the form $\boldsymbol{a} \leqslant \mathbf{A}\boldsymbol{\beta} \leqslant \boldsymbol{b}$, where $\mathbf{A}$ is a known $r \times p$ matrix, $\boldsymbol{a}$ and $\boldsymbol{b}$ are two known $r \times 1$ vectors. Several typical order restrictions in isotonic regression are special cases of such linear inequality constraints including the simple ordering $\beta_1 \leqslant \cdots \leqslant \beta_p$, the tree ordering $\beta_k \leqslant \beta_p$ $(k = 1, \ldots, p-1)$, the umbrella ordering $\beta_1 \leqslant \cdots \leqslant \beta_h \geqslant \beta_{h+1} \geqslant \cdots \geqslant \beta_p$ and the increasing convex ordering

$$0 \leqslant \frac{\beta_2 - \beta_1}{d_2 - d_1} \leqslant \frac{\beta_3 - \beta_2}{d_3 - d_2} \leqslant \cdots \leqslant \frac{\beta_p - \beta_{p-1}}{d_p - d_{p-1}},$$

where $\{d_k\}_{k=1}^p$ are known and $d_1 < \cdots < d_p$. For example, the simple ordering $\beta_1 \leqslant \cdots \leqslant \beta_p$ can be converted into the box constraint of the form $\boldsymbol{a} \leqslant \boldsymbol{\mu} \leqslant \boldsymbol{b}$, where

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top = \mathbf{A}_1\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$\boldsymbol{a}_{p\times 1} = (-\infty, 0, \ldots, 0)^\top$ and $\boldsymbol{b}_{p\times 1} = (+\infty, \ldots, +\infty)^\top$. Similarly, we can convert the tree ordering (corresponding to $s = 2$), the umbrella ordering ($s = 3$), and the increasing convex ordering ($s = 4$) into box constraints of the form $\boldsymbol{a} \leqslant \boldsymbol{\mu} \leqslant \boldsymbol{b}$, where $\boldsymbol{\mu} = \mathbf{A}_s\boldsymbol{\beta}$ and $\{\mathbf{A}_s\}_{s=2}^4$ are respectively given by (4.2), (4.3) and (4.4) of Tian *et al.* (2008). Furthermore, Tian *et al.* (2008) obtained the following results.

Proposition 1 (Tian et al., 2008). *Let $\boldsymbol{a} \leqslant \mathbf{A}\boldsymbol{\beta} \leqslant \boldsymbol{b}$ where $\mathbf{A}$ is an $r \times p$ matrix. (i) If $r = p$ and $\mathbf{A}^{-1}$ exists, then $\boldsymbol{\beta} = \mathbf{A}^{-1}\boldsymbol{\mu}$ with $\boldsymbol{\mu} \in [\boldsymbol{a}, \boldsymbol{b}]$; (ii) If $\mathbf{A}$ is a full row-rank matrix, then there exist two $p \times 1$ vectors $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and a $p \times p$ nonsingular matrix $\mathbf{A}^*$ such that $\boldsymbol{\beta} = (\mathbf{A}^*)^{-1}\boldsymbol{\mu}$ with $\boldsymbol{\mu} \in [\boldsymbol{a}^*, \boldsymbol{b}^*]$; (iii) If $\mathbf{A}$ is a full column-rank matrix, then $\boldsymbol{\beta} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\boldsymbol{\mu}$ with $\boldsymbol{\mu} \in [\boldsymbol{a}, \boldsymbol{b}]$.*

Proposition 1 indicates that we could focus on the box constraint. Since (3.12) is a diagonal matrix, the surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ is a separable function in the form $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = \sum_{k=1}^{p} Q_k(\beta_k|\boldsymbol{\beta}^{(m)})$, where $Q_k(\beta_k|\boldsymbol{\beta}^{(m)})$ is a one-dimensional concave function of β_k given $\boldsymbol{\beta}^{(m)}$ for $k = 1, \ldots, p$. Thus, solving the constrained optimization problem (3.10) with $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{a}, \boldsymbol{b}] = \prod_{k=1}^{p}[a_k, b_k]$ is equivalent to separately maximizing $Q_k(\beta_k|\boldsymbol{\beta}^{(m)})$ with respect to β_k subject to a lower bound a_k and an upper bound b_k, resulting in an explicit solution via a median function. In other words, we have the following MM algorithm:

$$\beta_k^{(m+1)} = \arg \max_{\beta_k \in [a_k, b_k]} Q_k(\beta_k|\boldsymbol{\beta}^{(m)}), \quad k = 1, \ldots, p, \tag{3.13}$$

or

$$\begin{aligned} \theta_k^{(m+1)} &= \arg \max_{\beta_k \in \mathbb{R}} Q_k(\beta_k|\boldsymbol{\beta}^{(m)}), \qquad (3.14) \\ \beta_k^{(m+1)} &= \text{median}(a_k,\, \theta_k^{(m+1)},\, b_k), \quad k = 1, \ldots, p, \end{aligned}$$

where $\mathbb{R}$ is the real line. Note that (3.14) is an unconstrained optimization problem of one dimension, the Newton–Raphson algorithm can be applied to obtain $\theta_k^{(m+1)}$. In fact, two built in `R` functions "`optimize`" (*one dimensional optimization*; this function searches the interval from lower to upper for a minimum or maximum of a function with respect to its first argument) and "`nlm`" (*non-linear minimization*; this function carries out a minimization of a function using a Newton-type algorithm) can facilitate the programming of `R`. We summarize the MM algorithm in the matrix form as follows and postpone the discussion of its convergence in the next subsection.

THE MM ALGORITHM FOR CONSTRAINED ESTIMATION IN COX'S MODEL:

Step 1: Given $\boldsymbol{\beta}^{(m)} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{a}, \boldsymbol{b}]$, calculate the score vector (3.11) and the negative Hessian matrix (3.12) for the $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ function;

Step 2: Update $\boldsymbol{\beta}^{(m+1)}$ via the following iterations:

$$\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}), \tag{3.15}$$

$$\boldsymbol{\beta}^{(m+1)} = \text{median}\left(\boldsymbol{a},\ \boldsymbol{\theta}^{(m+1)},\ \boldsymbol{b}\right), \tag{3.16}$$

where $\mathbb{R}^p$ is the p-dimensional Euclidean space.

Remark 1: In practice, alternatively, we suggest using a simple one-step Newton–Raphson method to calculate $\theta_k^{(m+1)}$ in (3.14) as follows:

$$\theta_k^{(m+1)} = \beta_k^{(m)} + \left[-\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k^2} \right]^{-1} \frac{\partial Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial \beta_k}, \qquad k = 1, \ldots, p, \tag{3.17}$$

so that (3.15) can be replaced by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[-\frac{\partial^2 Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} \right]^{-1} \frac{\partial Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}}. \tag{3.18}$$

The idea of the one-step Newton–Raphson method was also suggested by Becker *et al.* (1997, p.49, p.51) in reformulating the De Pierro (1995) algorithm. It is noted that the one-step Newton–Raphson update cannot guarantee the increase of the surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ (or the log-likelihood function) at each iteration. Thus, like Newton-type methods, the proposed MM algorithm based on (3.18) does not possess the monotone convergence. In this sense, it is not a real MM algorithm. However, based on our limited experience, the one-step Newton–Raphson method does not affect the final convergence of the above MM algorithm.

3.3 The ascent property of the new MM algorithm

For the proposed MM algorithm for constrained estimation in Cox's model, we have the following result.

Theorem 3 (The ascent property). *Let the partial log-likelihood function $\ell(\boldsymbol{\beta})$ and the surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ be given by (2.3) and (3.6), respectively. We have (i) $\ell(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) \geqslant 0$ for all $\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)} \in \mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{a}, \boldsymbol{b}]$, where the equality holds if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$; (ii) increasing $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ results in an increase in $\ell(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in [\boldsymbol{a}, \boldsymbol{b}]$.*

Proof. The assertion (i) is a special case of (3.9) when $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{a}, \boldsymbol{b}]$. Thus $\ell(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) \geqslant 0$ for all $\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)} \in [\boldsymbol{a}, \boldsymbol{b}]$, and $\ell(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ achieves its minimum zero at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$. On the other hand, from (3.15) and (3.16), we have $Q(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)}) \geqslant Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})$ for $\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)} \in [\boldsymbol{a}, \boldsymbol{b}]$. Combining the two facts, we obtain the following ascent property:

$$\begin{aligned} \ell(\boldsymbol{\beta}^{(m+1)}) &= [\ell(\boldsymbol{\beta}^{(m+1)}) - Q(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)})] + Q(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)}) \\ &\geqslant [\ell(\boldsymbol{\beta}^{(m)}) - Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})] + Q(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}) = \ell(\boldsymbol{\beta}^{(m)}) \end{aligned}$$

where the inequality is strict if $\boldsymbol{\beta}^{(m+1)} \neq \boldsymbol{\beta}^{(m)}$. $\square$

The ascent property indicates that finding (2.4) with $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{a}, \boldsymbol{b}]$ is equivalent to iteratively finding (3.10) or (3.15) and (3.16). Moreover, this ascent property guarantees not only the convergence of $\boldsymbol{\beta}^{(m+1)} \to \hat{\boldsymbol{\beta}}$ as $m \to \infty$ but also the monotone convergence, namely, each iteration must increase the partial log-likelihood $\ell(\boldsymbol{\beta})$.

3.4 Standard error estimation via a nonparametric bootstrapping

In the absence of constraints, the asymptotic variance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}$ takes the form $[\boldsymbol{\Sigma}(\boldsymbol{\beta})]^{-1}$, defined in Assumption (A4) (Andersen and Gill, 1982; Kalbfleisch and Prentice, 2002). Thus, the estimation of standard error can be calculated by replacing the large-sample quantities in $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ with their small-sample quantities. However, in the presence of constraints, due to the complicated derivation of $\mathbf{U}(\boldsymbol{\beta})$ in the asymptotic variance of constrained estimator $\hat{\boldsymbol{\beta}}$ in Theorem 2, there are no procedures available. We adopt the nonparametric bootstrap approach (Hjort, 1985; Efron and Tibshirani, 1993; Burr, 1994) to estimate the standard error of $\hat{\boldsymbol{\beta}}$.

The basic idea of the nonparametric bootstrap approach is to construct an empirical distribution function by repeatedly sampling from the observed data. As a computer-based method, it is widely used to estimate the standard error of an estimator, especially when the underlying distribution is unknown.

Let $Y_{\rm obs} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ denote the observed data for n subjects, where $\mathbf{X}_i \hat{=} (T_i, \Delta_i, \mathbf{Z}_i)$ denote the observed time, the right censoring indicator and the covariate vector for the i-th subject. We randomly draw from $Y_{\rm obs} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with replacement to obtain a bootstrap sample $Y_{\rm obs}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$, where each $\mathbf{X}_i^*$ equals any one of the n values $\mathbf{X}_i$ with probability $1/n$. In fact, we can use the built-in `R` function, `sample`$(\mathbf{X}, n, \text{prob} = \text{rep}(1/n, n), \text{replace} = \text{T})$, to produce a vector of length n randomly chosen from $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with equal probabilities $\{1/n, \ldots, 1/n\}$ with replacement. A bootstrap replication $\hat{\boldsymbol{\beta}}^*$ can be obtained by using the proposed MM algorithm based on $Y_{\rm obs}^*$. Independently repeating this process B times, we can obtain B bootstrap replications $\{\hat{\boldsymbol{\beta}}^*(b)\}_{b=1}^B$, where $\hat{\boldsymbol{\beta}}^*(b) = (\hat{\beta}_1^*(b), \ldots, \hat{\beta}_p^*(b))^\top$. Therefore, the standard error of the k-th component of the constrained MLEs $\hat{\boldsymbol{\beta}}$ can be estimated by the following sample standard derivation

$$\widehat{\mathrm{se}}(\hat{\beta}_k) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\beta}_k^*(b) - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_k^*(b)\right]^2}, \quad k = 1, \ldots, p. \tag{3.19}$$

When the constrained MLEs $\hat{\boldsymbol{\beta}}$ are located in the interior of the box constraints $[\boldsymbol{a}, \boldsymbol{b}]$, based on the result in Theorem 2 and (3.19), we can construct a Wald-type bootstrap *confidence interval* (CI) of β_k. In other words, if $\{\hat{\beta}_k^*(b)\}_{b=1}^B$ is approximately normally distributed, the $(1-\alpha)100\%$ Wald-type bootstrap CI of β_k is given by

$$[\hat{\beta}_k - z_{\alpha/2} \cdot \widehat{\mathrm{se}}(\hat{\beta}_k),\ \hat{\beta}_k + z_{\alpha/2} \cdot \widehat{\mathrm{se}}(\hat{\beta}_k)], \tag{3.20}$$

where z_α denotes the upper α-th quantile of the standard normal distribution. When the constrained MLEs $\hat{\boldsymbol{\beta}}$ are located on the boundaries of the box constraints $[\boldsymbol{a}, \boldsymbol{b}]$, generally speaking, the bootstrap replications $\{\hat{\beta}_k^*(b)\}_{b=1}^B$ are non-normally distributed. In this case, the $(1-\alpha)100\%$ bootstrap CI for β_k can be constructed as

$$[\hat{\beta}_{k,\mathrm{L}},\ \hat{\beta}_{k,\mathrm{U}}], \tag{3.21}$$

where $\hat{\beta}_{k,\mathrm{L}}$ and $\hat{\beta}_{k,\mathrm{U}}$ are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\hat{\beta}_k^*(b)\}_{b=1}^B$, respectively.

4. Simulation studies

As we mentioned before, evidences in some real-data analyses show that ignoring the restriction on parameters may result in misleading inference (Tan *et al.*, 2005; Fang *et al.*, 2006). We will conduct several simulation studies to compare the bias, MSE and variance for three estimators (denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$, respectively) of regression coefficients $\boldsymbol{\beta}$ in Cox's model with and without constraints, where $\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$ denotes the unconstrained estimator calculated by Newton–Raphson algorithm based on the partial likelihood function $\ell(\boldsymbol{\beta})$ in (2.3), $\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$ denotes the unconstrained estimator calculated by the MM algorithm based on the surrogate function $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ in (3.6), and $\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$ denotes the constrained estimator calculated by the proposed MM algorithm.

4.1 Experiments 1 and 2

We consider a proportional hazards model, where the hazard function of the failure time $\widetilde{T}$ given covariates (Z_1, Z_2) is assumed to be

$$\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2).$$

Furthermore, the regression coefficients β_1 and β_2 are assumed to be restricted by Case I: $\beta_1 \leqslant 0$ (box constraint) and Case Ⅱ: $\beta_1 \leqslant \beta_2$ (simple ordering constraint).

For Case I, let $\beta_1 = -0.5$ and $\beta_2 = 0.693$. We generate $Z_1 \sim$ Bernoulli$(1, 0.5)$ and independently generate $Z_2 \sim N(0,1)$. For Case Ⅱ, let $\beta_1 = 0.25$ and $\beta_2 = 0.5$. We generate $Z_1 \sim N(0,1)$ and independently generate $Z_2 \sim N(0.5, 1)$. For both cases, the baseline hazard function $\lambda_0(t)$ is set to be 1 and $2t$, respectively. Thus, the marginal distribution of failure time $\widetilde{T}$ is exponential with failure rate $\exp(\beta_1 Z_1 + \beta_2 Z_2)$ and Weibull distribution with shape parameter 2 and scale parameter $[\exp(\beta_1 Z_1 + \beta_2 Z_2)]^{-1/2}$, respectively. The censoring time C is generated from uniform distribution $U(0, c)$ with c chosen to depend on desired percentage of censoring. Approximately 30%, 50% and 80% censoring rates ρ are considered. Sample size n is set to be 50 and 100, respectively.

For each setting, we compare the estimators $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ and $\hat{\boldsymbol{\beta}}_{\text{UMM}}$ (without constraints) with the estimator $\hat{\boldsymbol{\beta}}_{\text{CMM}}$ (with constraints). Sample bias, sample MSE and sample variance of each estimator are calculated based on 1000 independent simulated data sets. The stopping criteria are specified by $|\ell(\boldsymbol{\beta}^{(m+1)}) - \ell(\boldsymbol{\beta}^{(m)})| < 10^{-4}$ for all three estimation methods. The corresponding simulation results are displayed in Table 1 and 2.

[Insert Tables 1 and 2 here]

Under all considered cases, estimators without constraints $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ and $\hat{\boldsymbol{\beta}}_{\text{UMM}}$ show almost identical results. As expected, the estimator with constraints performs better than estimators without constraints, since the sample MSE and sample variance of $\hat{\boldsymbol{\beta}}_{\text{CMM}}$ are notably smaller than the estimators $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ and $\hat{\boldsymbol{\beta}}_{\text{UMM}}$. This fact confirms that ignoring the constraints may reduce the efficiency of estimation. In terms of bias, the three estimators are all biased under all settings. However, the estimator with constraints sometimes may cause a larger bias than estimators without constraints. For example, the estimator for β_1 in Case I, for $n = 50$, $\rho = 0.30$, bias with and without constraints are -0.0302 and -0.0158, respectively.

4.2 Experiments 3 and 4

In this subsection, we would like to evaluate the performance of the proposed MM algorithm when the constraints are wrongly chosen via simulations. For Case I in Table 1, we set the constraint to be $\beta_1 \geqslant 0$ when the true value of β_1 is -0.5. For Case Ⅱ in Table 2, we choose the constraint to be $\beta_1 \geqslant \beta_2$ when the true values $\beta_1 = 0.25$, $\beta_2 = 0.5$. We consider $\lambda_0(t) = 1$,

$\rho = 30\%,\ 50\%,\ 80\%$, and $n = 100$, respectively. The simulation results are reported in Tables 3 and 4.

[Insert Tables 3 and 4 here]

Under all the cases here, we note that the simulation results of the estimators without constraints (i.e., $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ and $\hat{\boldsymbol{\beta}}_{\text{UMM}}$) remain to be consistent with those in Tables 1 and 2. However, non-ignorable biases arise in the simulation results of the estimate with constraints (i.e., $\hat{\boldsymbol{\beta}}_{\text{CMM}}$) when the constraints are poorly imposed. For example, in Table 3, the estimates of β_1 are close to 0, which is on the border of the box constraint. While the results of the estimators of β_2 are consistent with those in Table 1 since the constraint is only imposed on β_1. In Table 4, the estimators of both β_1 and β_2 are biased.

5. Numerical illustrations

In this section, we analyze two real data sets to illustrate the proposed method by comparing the estimations with and without restrictions.

5.1 Breast cancer trial

The most important discriminant in staging breast carcinoma is the presence of positive axillary lymph nodes. Sedmak *et al.* (1989) designed a study to determine if female breast cancer patients, originally classified as lymph node-negative by standard light microscopy (SLM), could be more accurately classified by immunohistochemical (IH) examination of their lymph nodes with an anticytokeratin monoclonal antibody cocktail. Identical section of lymph nodes were sequentially examined by SLM and IH. Forty five breast cancer patients with negative axillary lymph nodes by SLM examination and a minimum ten-year follow-up were selected from The Ohio State University Hospitals Cancer Registry. Of these 45 patients, 9 were immunoperoxidase positive and the remaining 36 were negative. Survival times in months for the patients are given in Table 5, which are obtained from Table 1.3 in Klein and Moeschberger (2003, p.7).

[Insert Table 5 here]

We use a proportional hazards model with the following hazard function to model the above data set:

$$\lambda(t|Z_1) = \lambda_0(t)\exp(\beta_1 Z_1), \tag{5.1}$$

where Z_1, as the unique covariate, denotes the immunoperoxidase status of these patients, and $Z_1 = 1$ if a patient's immunoperoxidase status is positive and $Z_1 = 0$ otherwise. Since the results in Sedmak *et al.* (1989) showed that patients with IH-detected metastases had significantly higher risk of death than that without IH-detected metastases, it is reasonable to impose a non-negativity restriction on β_1:

$$\beta_1 \geqslant 0. \tag{5.2}$$

To analyze the data, we apply the proposed MM algorithm to estimate the regression coefficient β_1 in model (5.1) with constraint specified by (5.2). The corresponding standard error of $\hat{\beta}_1$ can be estimated by the non-parametric bootstrap approach as shown in (3.19). For the purpose of comparison, we also compute the unconstrained MLE of β_1 by using the Newton–Raphson algorithm based on the partial likelihood function. The results are listed in Table 6 and show that immunoperoxidase positive significantly increases the risk of death under both methods. In addition, the MLE $\hat{\beta}_1 = 0.9802$ indicates that a patient with positive immunoperoxidase could be $\exp(0.9802) = 2.67$ times more likely to die than a patient with negative immunoperoxidase. Table 6 also shows that the proposed MM algorithm with constraints has a smaller standard error, resulting in a shorter confidential interval. This demonstrates that ignoring the constraint would lower the analysis efficacy.

[Insert Table 6 here]

5.2 Bone marrow transplant study

Bone marrow transplant is a standard treatment for acute leukemia. Recovery following bone marrow transplantation is a complex process. Prognosis for recovery may depend on risk factors known at the time of transplantation, such as patient and donor's age and gender, the stage of initial disease, and so on. The final prognosis may change, such as development of acute or chronic graft-versus-host disease (GVHD), return of the platelet count to normal levels, return of granulocytes to normal levels, or development of infections, etc.. Copelan *et al.* (1991) studied a multi-center trial of patients prepared for transplantation with a radiation-free conditioning regimen to illustrate the recovery process. The preparative regimen used in this study of allogeneic marrow transplants for patients with acute myelocytic leukemia (AML) and acute lymphoblastic leukemia (ALL) was a combination of 16 mg/kg of oral Busulfan (BU) and 120 mg/kg of intravenous cyclophosphamide. A total of 137 patients were treated at one of four hospitals: The Ohio State University Hospitals, Hahnemann University, St. Vincent's Hospital and Alfred Hospital. The study consists of transplants conducted at these institutions from March, 1984 to June, 1989. The maximum follow-up was 7 years.

Transplantation can be considered as a failure when a patient's leukemia relapse or he/she dies while in remission. Several potential risk factors were measured at the time of transplantation. Disease groups were categorized based on the patients risk status as ALL, AML low-risk first remission, and AML high-risk second remission or untreated first relapse or second or greater relapse or never in remission. For AML patients, their French–American–British (FAB) classification were based on standard morphological criteria. AML patients with an FAB classification of M4 or M5 are considered to have a possible elevated risk of relapse or treatment-related death. Patients at the two hospitals (St. Vincent's Hospital and Alfred Hospital) were given an GVHD prophylactic combining methotrexate (MTX) with cyclosporine and possibly methylprednisolone. Patients at the other two hospitals were not given methotrexate but rather a combination of cyclosporine and methylprednisolone. Other risk factors include patient and donor's gender, age, cytomegalovirus immune status (CMV) status, waiting time from diagnosis to transplantation and so on.

Table 7 lists the demographic characteristics for all 137 patients, which are obtained from Klein and Moeschberger (2003, p.483–487). There are 45 AML patients with FAB classification of M4 or M5. The average ages of patient and donor are 28.36 and 28.33, respectively. CMV status is positive in 49.64% patients and 42.34% donors, respectively.

[Insert Table 7 here]

Following the analysis of Klein and Moeschberger (2003), we build the proportional hazards model as follows:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)$$

$$\times \exp(\beta_1\,\mathrm{FAB} + \beta_2\,\mathrm{AMLlow} + \beta_3\,\mathrm{AMLhigh} + \beta_4\,\mathrm{DonAge} + \beta_5\,\mathrm{RecAge} + \beta_6\,\mathrm{DRAge}), \quad (5.3)$$

where FAB $= 1$ if FAB classification is M4 or M5 for AML patients, AMLlow $= 1$ if AML low-risk, AMLhigh $= 1$ if AML high-risk, DonAge $=$ donor age -28, RecAge $=$ patient age -28, and DRAge $=$ DonAge $\times$ RecAge. Furthermore, as we mentioned before, AML patients with an

FAB classification of M4 or M5 tend to have an increased risk of relapse or treatment-related death, we focus on the effect of FAB and impose non-negativity restriction on β_1:

$$\beta_1 \geqslant 0. \tag{5.4}$$

We calculate the constrained MLEs of the regression coefficients in model (5.3) with constraint specified by (5.4) by the proposed MM algorithm, and estimate the corresponding standard error via the non-parametric bootstrap method. The results are listed in the third column of Table 8, which suggests that FAB classification, risk groups and age have significant effects on risk of relapse or treatment-related death. Unsurprisingly, patients whose FAB classification of M4 or M5 have a higher risk after the transplantation. In addition, we also calculate the unrestricted MLEs of the regression coefficients by using the Newton–Raphson algorithm based on the partial likelihood function, and list the results in the second column of Table 8. By comparing the two columns, the evidence strongly supports that the estimation with constraint performs better than the estimation without restrictions by providing a smaller standard error and a shorter confidential interval. Figure 1 is a plot of the estimated survival function for two FAB classifications. Again, this plot shows that patients with FAB classification of M4 or M5 have a worse survival curve. All analyses suggest that considering FAB effect with restriction improves the efficiency of the study.

[Insert Table 8 and Figure 1 here]

6. Discussion

We studied the constrained estimation in the proportional hazards model with right-censored survival data and derived two asymptotic properties (i.e., consistency and asymptotic normality) by using the Lagrangian method based on KKT conditions. We developed a new MM algorithm for calculating the MLE of the regression coefficients with box or linear inequality restrictions in the proportional hazards model, where a surrogate function with a diagonal Hessian matrix is established, resulting in an explicit solution in the second M-step via a median function. Standard error estimation is also introduced through a non-parametric bootstrap approach. Simulation studies suggested that the proposed estimation provides a feasible and efficient method for the inference of the regression coefficients with box constraints in Cox's model. We applied the proposed method to analyze data sets from a breast cancer trial and a bone marrow transplant study, respectively. The gain in these analyses suggests that the consideration of restrictions improves the efficiency of the study.

Merely based on (3.2), we can construct the second surrogate function, denoted by

$$Q_2(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = \sum_{i=1}^{n} \Delta_i \left\{ \boldsymbol{\beta}^{\top}\mathbf{Z}_i(T_i) - \log\left[\sum_{l=1}^{n}\sum_{k=1}^{p} Y_l(T_i)\lambda_{lk}\, g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)})\right]\right\},$$

where $g_{lk}(\beta_k|\boldsymbol{\beta}^{(m)})$ is given by (3.8). However, the Hessian matrix of $Q_2(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ is not diagonal. In other words, the MM algorithm with $Q_2(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ as the surrogate function cannot be applied to the proportional hazards model with box constraints but can be applied to the proportional hazards model without constraints. Similarly, only based on (3.4), although we can establish the third surrogate function

$$Q_3(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = c_0 + \sum_{i=1}^{n} \Delta_i \left[\boldsymbol{\beta}^{\top}\mathbf{Z}_i(T_i) - \frac{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{\top}\mathbf{Z}_l(T_i)}}{\sum_{l=1}^{n} Y_l(T_i)\, \mathrm{e}^{\boldsymbol{\beta}^{(m)\top}\mathbf{Z}_l(T_i)}}\right],$$

where c_0 is given by (3.7), the Hessian matrix of $Q_3(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ is also not diagonal. Up to now, there are at least three algorithms (i.e., two MM algorithms based on $Q_2(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ and $Q_3(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ and the QLB algorithm of Böhning and Lindsay, 1988) with monotone convergence, which can be applied to Cox's model without constraints. The comparison among the convergence rates of the three algorithms is one of our research topics in the future.

As a general resampling statistical inference tool, the bootstrap method has many advantages in practice. Many authors studied the bootstrap theory for parametric models (Bickel and Freedman, 1981; Singh, 1981; Mason and Newton, 1992). However, theoretical studies on the bootstrap inference in semi-parametric models are quite challenging. Hjort (1985) studied the asymptotic properties of a bootstrap procedure under the Cox's model. Burr (1992) assessed bootstrap confidence intervals under the Cox's model. Cheng and Huang (2010) showed the consistency of the bootstrap method in semi-parametric models. However, all these papers only considered the asymptotic theories of the bootstrap without parameter constraints. One of our future researches includes the study of the asymptotic properties of the bootstrap procedure for the Cox model under constraints.

Appendix: Proofs of asymptotic properties

Proof of Theorem 1. Let $A(\boldsymbol{\beta}, t)$ be the logarithm of the Cox partial-likelihood function evaluated at time t, i.e.

$$A(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \boldsymbol{\beta}^\top \mathbf{Z}_i(s)\,\mathrm{d}N_i(s) - \sum_{i=1}^{n} \int_0^t \log\left(\sum_{l=1}^{n} Y_l(s)\,\mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{Z}_l(s)}\right) \mathrm{d}N_i(s). \tag{A.1}$$

Then, we have $\ell(\boldsymbol{\beta}) = A(\boldsymbol{\beta}, \tau)$. Based on the discussions in Anderson and Gill (1982) regarding asymptotic properties of Cox models, we observe that

$$C(\boldsymbol{\beta}, \tau) = \frac{A(\boldsymbol{\beta}, \tau) - A(\boldsymbol{\beta}^0, \tau)}{n}$$

uniformly converges to

$$D(\boldsymbol{\beta}) = \int_0^\tau \left\{ (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top \boldsymbol{s}^{(1)}(\boldsymbol{\beta}^0, t) - \log\left[\frac{\boldsymbol{s}^{(0)}(\boldsymbol{\beta}, t)}{\boldsymbol{s}^{(0)}(\boldsymbol{\beta}^0, t)}\right] \boldsymbol{s}^{(0)}(\boldsymbol{\beta}^0, t) \right\} \lambda_0(t)\,\mathrm{d}t,$$

which is a continuous and concave function of $\boldsymbol{\beta}$ and has a unique maximum at $\boldsymbol{\beta}^0$. That is,

$$D(\boldsymbol{\beta}) \leqslant D(\boldsymbol{\beta}^0)$$

with equality if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}^0$.

Using the method of the contradiction, we assume that there is a set of positive probability such that $\hat{\boldsymbol{\beta}}_n$ does not converge to $\boldsymbol{\beta}^0$. Then there exists a subsequence $\{\hat{\boldsymbol{\beta}}_{j_n}\}$ of $\{\hat{\boldsymbol{\beta}}_n\}$ which converges to $\tilde{\boldsymbol{\beta}}$ not equal to $\boldsymbol{\beta}^0$. Since $\hat{\boldsymbol{\beta}}_{j_n}$ is the maximum, we have $C(\hat{\boldsymbol{\beta}}_{j_n}, \tau) \geqslant C(\boldsymbol{\beta}^0, \tau)$. By the uniform convergency and continuity of limit, we obtain

$$D(\tilde{\boldsymbol{\beta}}) \geqslant D(\boldsymbol{\beta}^0), \qquad \text{for} \quad \tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^0.$$

By this contradiction, we have the convergency of $\hat{\boldsymbol{\beta}}_n$ to $\boldsymbol{\beta}^0$ in probability. □

By combining the method developed by Moore and Sadler (2006) and Moore *et al.* (2008) in the proof of the asymptotic normality of constrained MLE in parametric model with the asymptotic theory of Cox's model without constraints, we can obtain the asymptotic normality of the constrained estimator.

Proof of Theorem 2. Since $\hat{\boldsymbol{\beta}}_n$ is a stationary point of the $\ell(\boldsymbol{\beta})$, by Eq. (2.8) and the definition of $\mathbf{U}(\boldsymbol{\beta})$, we have

$$[\nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}}_n)]^\top \mathbf{U}(\hat{\boldsymbol{\beta}}_n) = 0.$$

Define $\mathbf{U}(\boldsymbol{\beta}) = (\boldsymbol{u}_1(\boldsymbol{\beta}), \ldots, \boldsymbol{u}_d(\boldsymbol{\beta}))$, where $\boldsymbol{u}_j(\boldsymbol{\beta})$ is the j-th column of $\mathbf{U}(\boldsymbol{\beta})$ and $d = p - q$, we have

$$[\nabla_{\boldsymbol{\beta}} \ell(\hat{\boldsymbol{\beta}}_n)]^\top \boldsymbol{u}_j(\hat{\boldsymbol{\beta}}_n) = 0, \quad j = 1, \ldots, d. \tag{A.2}$$

The first-order Taylor expansion of the left-hand side of Eq. (A.2) around $\boldsymbol{\beta}$ yields

$$\begin{aligned} 0 &= [\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})]^\top \boldsymbol{u}_j(\boldsymbol{\beta}) + [\boldsymbol{u}_j(\boldsymbol{\beta})]^\top \nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + [\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})]^\top \nabla_{\boldsymbol{\beta}} \boldsymbol{u}_j(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + o_{\mathrm{P}}(1) \\ &= [\boldsymbol{u}_j(\hat{\boldsymbol{\beta}}_n)]^\top \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + [\boldsymbol{u}_j(\boldsymbol{\beta})]^\top \nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + o_{\mathrm{P}}(1), \end{aligned}$$

where $o_{\mathrm{P}}(1)$ vanishes as $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| \xrightarrow{\mathrm{P}} 0$. Thus, we have

$$[\boldsymbol{u}_j(\hat{\boldsymbol{\beta}}_n)]^\top \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0) + [\boldsymbol{u}_j(\boldsymbol{\beta}^0)]^\top \nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) + o_{\mathrm{P}}(1) = 0. \tag{A.3}$$

Since $\mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ is connected, there exists a path-connected curve on the surface of $\boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}$, including the constrained estimators $\{\hat{\boldsymbol{\beta}}_n\}$ and $\boldsymbol{\beta}^0$. Let $\boldsymbol{\varphi}(w)$: $\mathbb{R} \to \mathbb{S}(\boldsymbol{f}, \boldsymbol{g})$ be a continuously differentiable map such that $\boldsymbol{\varphi}(0) = \boldsymbol{\beta}^0$ and $\boldsymbol{\varphi}(1/n) = \hat{\boldsymbol{\beta}}_n$. Thus, we have

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0 = \nabla_w \boldsymbol{\varphi}(a_n)\frac{1}{n}, \quad \text{where} \quad 0 < a_n < \frac{1}{n}.$$

Since $\boldsymbol{h}(\boldsymbol{\varphi}(a_n)) = 0$, we obtain

$$0 = \nabla_w \boldsymbol{h}(\boldsymbol{\varphi}(a_n)) = \mathbf{F}(\boldsymbol{\varphi}(a_n)) \nabla_w \boldsymbol{\varphi}(a_n). \tag{A.4}$$

Thus, there exists $\boldsymbol{b}_n \in \mathbb{R}^d$ such that

$$\nabla_w \boldsymbol{\varphi}(a_n) = \mathbf{U}(\boldsymbol{\varphi}(a_n)) \cdot \boldsymbol{b}_n. \tag{A.5}$$

Due to Eq.'s (A.2)–(A.5), we have

$$\boldsymbol{b}_n = \left[\mathbf{U}(\boldsymbol{\beta}^0)^\top \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0)\right) \mathbf{U}(\boldsymbol{\varphi}(a_n))\right]^{-1} \mathbf{U}(\hat{\boldsymbol{\beta}}_n)^\top \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0) + o_{\mathrm{P}}(1).$$

Hence, we obtain

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) &= \mathbf{U}(\boldsymbol{\varphi}(a_n)) \left[\mathbf{U}(\boldsymbol{\beta}^0)^\top \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0)\right) \mathbf{U}(\boldsymbol{\varphi}(a_n))\right]^{-1} \mathbf{U}(\hat{\boldsymbol{\beta}}_n)^\top \left(\frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0)\right) \\ &\quad + o_{\mathrm{P}}(1). \end{aligned}$$

Under Assumptions (A1)–(A4) and the discussions in Anderson and Gill (1982), Kalbfleisch and Prentice (2002), we have

$$\begin{aligned} -\frac{1}{n}\nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0) &\xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}(\boldsymbol{\beta}^0), \quad \text{and} \\ \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^0) &\xrightarrow{\mathrm{D}} N_p(\mathbf{0}, \ \boldsymbol{\Sigma}(\boldsymbol{\beta}^0)). \end{aligned}$$

By continuity, we have $\mathbf{U}(\hat{\boldsymbol{\beta}}_n) \to \mathbf{U}(\boldsymbol{\beta}^0)$ and $\mathbf{U}(\boldsymbol{\varphi}(a_n)) \to \mathbf{U}(\boldsymbol{\beta}^0)$, as $n \to \infty$. Due to Slutsky's theorem, we can obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) \xrightarrow{\mathrm{D}} N_p(\mathbf{0},\ \boldsymbol{\Omega}(\boldsymbol{\beta}^0)),$$

where

$$\boldsymbol{\Omega}(\boldsymbol{\beta}^0) = \mathbf{U}(\boldsymbol{\beta}^0)\left[\mathbf{U}(\boldsymbol{\beta}^0)^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}^0)\mathbf{U}(\boldsymbol{\beta}^0)\right]^{-1}\mathbf{U}(\boldsymbol{\beta}^0)^\top,$$

which indicates the asymptotic normality. □

Acknowledgments

The authors thank the Co-editor, Professor Jae Chang Lee, an Associate Editor and two referees for insightful comments which led to important improvements over an earlier draft.

References

[1] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100–1120.

[2] Becker, M. P., Yang, I. and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research* **6**, 38–54.

[3] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**, 1196–1217.

[4] Bickel, P. J., Klassen, C. X., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

[5] Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic approximation algorithms. *The Annals of the Institute of Statistical Mathematics* **40**, 641–663.

[6] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[7] Burr, D. (1994). A comparison of certain bootstrap confidence intervals in the Cox model. *Journal of the American Statistical Association* **89**, 1290–1320.

[8] Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **94**, 896–908.

[9] Chen, K. and Lo, S. H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755–764.

[10] Cheng, G. and Huang, J. Z. (2010). Bootstrap consistencyy for general semiparametric M-estimation. *The Annals of Statistics* **38**, 2884–2915.

[11] Copelan, E. A., Biggs, J. C., Thompson, J. M., Crilley, P., Szer, J., Klein, J. P., Kapoor, N., Avalos, B. R., Cunningham, I., Atkinson, K., Downs, K., Harmon, G. S., Daly, M. B., Brodsky, I., Bulova, S. I. and Tutschka, P. J. (1991). Treatment for Acute Myelocytic Leukemia with Allogeneic Bone Marrow Transplantation Following Preparation with Bu/Cy. *Blood* **78**, 838–843.

[12] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, B* **34**, 187–220.

[13] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

[14] De Pierro, A. R. (1995). A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on Medical Imaging* **14**, 132–137.

[15] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton.

[16] Fang, H. B., Tian, G. L., Xiong, X. P. and Tan, M. (2006). A multivariate random-effects model with restricted parameters: Application to assessing radiation therapy for brain tumours. *Statistics in Medicine* **25**, 1948–1959.

[17] Hjort, N. L. (1985). Bootstrapping Cox's regression model. Technical report NSF–241, Department of Statistics, Stanford University.

[18] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**, 540–568.

[19] Hunter, D. R. and Lange, K. (2002). Computing estimates in the proportional odds model. *Annals of the Institute of Statistical Mathematics* **54**(1), 155–168.

[20] Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58**, 30–37.

[21] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

[22] Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer, New York.

[23] Lange, K. (2004). *Optimization*. Springer, New York.

[24] Lange, K. (2010). *Numerical Analysis for Statisticians* (2-nd Edition). Springer, New York.

[25] Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* **9**, 1–20.

[26] Liew, C. K. (1976). Inequality constrained least-squares estimation. *Journal of the American Statistical Association* **71**, 746–751.

[27] Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* **88**, 1341–1349.

[28] Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine* **13**, 2233–2247.

[29] Madsen, K., Nielsen, H. B. and Tingleff, O. (2004). Optimization with constraints. Informatics and Mathematical Modelling. Technical University of Denmark.

[30] Mason, D. and Newton, M. (1992). A rank statistic approach to the consistency of a general bootstrap. *The American Statistician* **20**, 1611–1624.

[31] Moore, T. J. and Sadler, B. M. (2006). Maximum-likelihood estimation and scoring under parametric constrains. Army Research Lab, Aldelphi, MD, Tech. Rep. ARL-TR-3805.

[32] Moore, T. J., Sadler, B. M. and Kozick, R. J. (2008). Maximum-likelihood estimation, the Cramér–Rao bound, and the method of scoring with parameter constraints. *IEEE Transactions on Signal Processing* **56**, 895–908.

[33] Nyquist, H. (1991). Restricted estimation of generalized linear models. *Journal of the Royal Statistical Society, C (Applied Statistics)* **40**, 133–141.

[34] Sedmak, D. D., Meineke, T. A., Knechtges, D. S. and Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metastases. *Modern Pathology* **2**, 516–520.

[35] Silvapulle, M. J. (1997). On order restricted inference in some mixed linear models. *Statistics and Probability Letters* **36**, 23–27.

[36] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics* **9**, 1187–1195.

[37] Tan, M., Fang, H. B., Tian, G. L. and Houghton, P. J. (2005). Repeated-measures models with constrained parameters for incomplete data in tumour xenograft experiments. *Statistics in Medicine* **24**, 109–119.

[38] Tian, G. L., Ng, K. W and Tan, M. (2008). EM-type algorithms for conputing restricted MLEs in multivariate normal distributions and multivariate t-distributions. *Computational Statistics and Data Analysis* **52**, 4768–4778.

[39] Wang, J. (1996). Asymptotics of least-squares estimation for constrained nonlinear regression. *The Annals of Statistics* **24**, 1316–1326.

[40] Wang, J. (2000). Approximate representation of estimators in constrained regression problems. *Scandinavian Journal of Statistics* **27**, 21–33.

[41] Xu, J. and Wang, J. (2008). Maximum likelihood estimation of linear models for longitudinal data with inequality constraints. *Communications in Statistics — Theory and Methods* **37**, 931–946.

Table 1: Simulation results based on the model $\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$ with box constraint $\beta_1 \leqslant 0$, where $Z_1 \sim$ Bernoulli$(1, 0.5)$ is independent of $Z_2 \sim N(0, 1)$

			$\lambda_0(t) = 1$					
n	ρ	Method	$\beta_1 = -0.5$			$\beta_2 = 0.693$		
			Bias	MSE	Variance	Bias	MSE	Variance
50	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0159	0.1492	0.1491	0.0355	0.0463	0.0451
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0157	0.1491	0.1490	0.0355	0.0463	0.0451
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0301	0.1310	0.1302	0.0348	0.0462	0.0450
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0341	0.1926	0.1916	0.0462	0.0706	0.0685
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0339	0.1925	0.1916	0.0462	0.0706	0.0685
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0570	0.1610	0.1579	0.0460	0.0699	0.0678
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0226	0.5929	0.5930	0.1147	0.2545	0.2416
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0225	0.5926	0.5927	0.1146	0.2539	0.2410
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.1299	0.3913	0.3748	0.1079	0.2377	0.2263
100	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0105	0.0623	0.0623	0.0242	0.0222	0.0216
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0103	0.0623	0.0623	0.0242	0.0222	0.0216
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0125	0.0598	0.0597	0.0242	0.0222	0.0216
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0097	0.0905	0.0905	0.0110	0.0291	0.0290
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0095	0.0905	0.0905	0.0110	0.0291	0.0290
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0150	0.0835	0.0834	0.0109	0.0290	0.0289
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0457	0.2754	0.2736	0.0343	0.0726	0.0715
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0456	0.2753	0.2735	0.0343	0.0726	0.0715
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0834	0.2186	0.2118	0.0338	0.0723	0.0712

			$\lambda_0(t) = 2t$					
n	ρ	Method	$\beta_1 = -0.5$			$\beta_2 = 0.693$		
			Bias	MSE	Variance	Bias	MSE	Variance
50	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0060	0.1610	0.1611	0.0478	0.0545	0.0523
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0057	0.1609	0.1610	0.0478	0.0545	0.0522
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0231	0.1372	0.1368	0.0472	0.0540	0.0518
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0255	0.2244	0.2240	0.0645	0.0851	0.0811
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0253	0.2243	0.2239	0.0645	0.0851	0.0810
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0575	0.1787	0.1756	0.0625	0.0826	0.0787
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0638	0.5937	0.5903	0.0827	0.2271	0.2205
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0637	0.5933	0.5899	0.0826	0.2270	0.2204
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.1606	0.4281	0.4027	0.0808	0.2213	0.2150
100	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0025	0.0655	0.0655	0.0220	0.0235	0.0230
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0022	0.0654	0.0655	0.0220	0.0235	0.0230
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0038	0.0636	0.0636	0.0220	0.0234	0.0230
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0158	0.0988	0.0987	0.0236	0.0323	0.0318
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0155	0.0988	0.0986	0.0236	0.0323	0.0318
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0208	0.0922	0.0918	0.0236	0.0322	0.0316
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0235	0.2564	0.2562	0.0507	0.0854	0.0829
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0233	0.2563	0.2560	0.0507	0.0854	0.0829
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	-0.0626	0.1982	0.1944	0.0485	0.0834	0.0812

NOTE: $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ denotes the unconstrained estimator calculated by Newton-Raphson algorithm based on the partial likelihood function. $\hat{\boldsymbol{\beta}}_{\text{UMM}}$ denotes the unconstrained estimator calculated by MM algorithm based on the surrogate function. $\hat{\boldsymbol{\beta}}_{\text{CMM}}$ denotes the constrained estimator calculated by the proposed MM algorithm.

Table 2: Simulation results based on the model $\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$ with inequality constraint $\beta_1 \leqslant \beta_2$, where $Z_1 \sim N(0,1)$ is independent of $Z_2 \sim N(0.5,1)$

n	ρ	Method	$\lambda_0(t)=1$					
			$\beta_1=0.25$			$\beta_2=0.50$		
			Bias	MSE	Variance	Bias	MSE	Variance
50	0.30	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0187	0.0410	0.0407	0.0154	0.0403	0.0401
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0187	0.0409	0.0406	0.0153	0.0402	0.0400
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0033	0.0326	0.0326	0.0281	0.0358	0.0350
	0.50	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0221	0.0551	0.0547	0.0311	0.0670	0.0661
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0222	0.0551	0.0546	0.0310	0.0669	0.0660
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0013	0.0448	0.0448	0.0499	0.0599	0.0575
	0.80	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0283	0.1736	0.1729	0.0629	0.1870	0.1832
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0282	0.1733	0.1726	0.0626	0.1859	0.1822
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0274	0.1326	0.1320	0.1100	0.1614	0.1494
100	0.30	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0149	0.0185	0.0183	0.0113	0.0191	0.0190
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0150	0.0185	0.0183	0.0112	0.0190	0.0189
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0113	0.0172	0.0171	0.0146	0.0180	0.0178
	0.50	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0107	0.0254	0.0254	0.0201	0.0275	0.0271
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0107	0.0254	0.0253	0.0199	0.0274	0.0270
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0034	0.0225	0.0225	0.0265	0.0254	0.0247
	0.80	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0136	0.0663	0.0662	0.0246	0.0652	0.0647
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0136	0.0663	0.0662	0.0245	0.0651	0.0645
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0112	0.0528	0.0527	0.0476	0.0565	0.0542

n	ρ	Method	$\lambda_0(t)=2t$					
			$\beta_1=0.25$			$\beta_2=0.5$		
			Bias	MSE	Variance	Bias	MSE	Variance
50	0.30	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0096	0.0456	0.0456	0.0310	0.0463	0.0454
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0096	0.0456	0.0456	0.0309	0.0462	0.0453
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0050	0.0384	0.0384	0.0438	0.0415	0.0396
	0.50	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0240	0.0659	0.0654	0.0462	0.0719	0.0698
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0240	0.0658	0.0653	0.0461	0.0717	0.0697
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0022	0.0531	0.0531	0.0645	0.0658	0.0617
	0.80	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0210	0.1754	0.1751	0.0697	0.1894	0.1848
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0210	0.1753	0.1750	0.0694	0.1885	0.1839
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0355	0.1255	0.1243	0.1171	0.1628	0.1493
100	0.30	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	−0.0013	0.0175	0.0175	0.0256	0.0204	0.0198
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	−0.0012	0.0175	0.0175	0.0255	0.0204	0.0197
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0041	0.0164	0.0164	0.0281	0.0197	0.0189
	0.50	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0168	0.0256	0.0254	0.0299	0.0275	0.0266
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0169	0.0256	0.0254	0.0298	0.0274	0.0265
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	0.0095	0.0223	0.0223	0.0365	0.0255	0.0242
	0.80	$\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$	0.0064	0.0643	0.0643	0.0216	0.0706	0.0702
		$\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$	0.0064	0.0643	0.0643	0.0215	0.0705	0.0701
		$\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$	−0.0193	0.0506	0.0503	0.0437	0.0614	0.0596

NOTE: $\hat{\boldsymbol{\beta}}_{\mathrm{UNR}}$ denotes the unconstrained estimator calculated by Newton-Raphson algorithm based on the partial likelihood function. $\hat{\boldsymbol{\beta}}_{\mathrm{UMM}}$ denotes the unconstrained estimator calculated by MM algorithm based on the surrogate function. $\hat{\boldsymbol{\beta}}_{\mathrm{CMM}}$ denotes the constrained estimator calculated by the proposed MM algorithm.

Table 3: Simulation results based on the model $\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$ with box constraint $\beta_1 \geqslant 0$, where $\lambda_0(t) = 1$, and $Z_1 \sim$ Bernoulli$(1, 0.5)$ is independent of $Z_2 \sim N(0, 1)$

n	ρ	Method	$\beta_1 \geqslant 0$					
			$\beta_1 = -0.5$			$\beta_2 = 0.693$		
			Bias	MSE	Variance	Bias	MSE	Variance
100	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0105	0.0623	0.0623	0.0242	0.0222	0.0216
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0103	0.0623	0.0623	0.0242	0.0222	0.0216
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.5022	0.2525	0.0003	-0.0034	0.0211	0.0211
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0220	0.0875	0.0871	0.0299	0.0306	0.0297
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0218	0.0875	0.0871	0.0298	0.0306	0.0297
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.5050	0.2560	0.0010	0.0074	0.0288	0.0288
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	-0.0402	0.2501	0.2487	0.0462	0.0778	0.0758
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	-0.0401	0.2500	0.2486	0.0462	0.0778	0.0757
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.5286	0.2904	0.0110	0.0310	0.0727	0.0718

NOTE: $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ denotes the unconstrained estimator calculated by Newton-Raphson algorithm based on the partial likelihood function. $\hat{\boldsymbol{\beta}}_{\text{UMM}}$ denotes the unconstrained estimator calculated by MM algorithm based on the surrogate function. $\hat{\boldsymbol{\beta}}_{\text{CMM}}$ denotes the constrained estimator calculated by the proposed MM algorithm.

Table 4: Simulation results based on the model $\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$ with inequality constraint $\beta_1 \geqslant \beta_2$, where $\lambda_0(t) = 1$, and $Z_1 \sim N(0, 1)$ is independent of $Z_2 \sim N(0.5, 1)$

n	ρ	Method	$\beta_1 \geqslant \beta_2$					
			$\beta_1 = 0.25$			$\beta_2 = 0.50$		
			Bias	MSE	Variance	Bias	MSE	Variance
100	0.30	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	0.0149	0.0185	0.0183	0.0113	0.0191	0.0190
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	0.0150	0.0185	0.0183	0.0112	0.0190	0.0189
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.1325	0.0277	0.0102	-0.1246	0.0257	0.0102
	0.50	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	0.0131	0.0254	0.0253	0.0131	0.0263	0.0262
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	0.0132	0.0254	0.0253	0.0130	0.0262	0.0261
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.1364	0.0322	0.0136	-0.1264	0.0289	0.0130
	0.80	$\hat{\boldsymbol{\beta}}_{\text{UNR}}$	0.0177	0.0629	0.0627	0.0160	0.0653	0.0651
		$\hat{\boldsymbol{\beta}}_{\text{UMM}}$	0.0177	0.0629	0.0627	0.0159	0.0652	0.0650
		$\hat{\boldsymbol{\beta}}_{\text{CMM}}$	0.1596	0.0597	0.0342	-0.1382	0.0519	0.0328

NOTE: $\hat{\boldsymbol{\beta}}_{\text{UNR}}$ denotes the unconstrained estimator calculated by Newton-Raphson algorithm based on the partial likelihood function. $\hat{\boldsymbol{\beta}}_{\text{UMM}}$ denotes the unconstrained estimator calculated by MM algorithm based on the surrogate function. $\hat{\boldsymbol{\beta}}_{\text{CMM}}$ denotes the constrained estimator calculated by the proposed MM algorithm.

Table 5: Time to death for breast cancer patients with different immunohistochemical responses

Immunoperoxidase negative ($n_1 = 36$)						Immunoperoxidase positive ($n_2 = 9$)					
19	25	30	34	37	46	22	23	38	42	73	77
47	51	56	57	61	66	89	115	144^+			
67	74	78	86	122^+	123^+						
130^+	130^+	133^+	134^+	136^+	141^+						
143^+	148^+	151^+	152^+	153^+	154^+						
156^+	162^+	164^+	165^+	182^+	189^+						

SOURCE: Data in this table are obtained from Table 1.3 in Klein and Moeschberger (2003, p.7).
NOTE: + denotes censored observation.

Table 6: Data analysis for the breast cancer trial

Variable	UNR method			CMM method		
	Estimate	std	95% CI	Estimate	std	95% CI
Immunoperoxidase	0.9802^*	0.4367	$[0.1243, 1.8361]$	0.9802^*	0.4193	$[0.1583, 1.8021]$

NOTE: UNR denotes the unconstrained estimator calculated by Newton–Raphson algorithm based on the partial likelihood function. CMM denotes the constrained estimator calculated by the proposed MM algorithm. "*" indicates that the parameter estimate is significant at 5% level.

Table 7: Demographics and characteristics for the bone marrow transplant study

Variables	% or mean ± std
FAB	
FAB Grade 4 or 5 and AML	32.85 (45/137)
Otherwise	67.15 (92/137)
Disease Group	
ALL	27.74 (38/137)
AML Low-risk	39.42 (54/137)
AML High-risk	32.85 (45/137)
Patient Age	28.36 ± 9.56
Donor Age	28.33 ± 10.18
Patient Sex	
Male	58.39 (80/137)
Female	41.61 (57/137)
Donor Sex	
Male	64.23 (88/137)
Female	35.77 (49/137)
Patient CMV Status	
CMV Positive	49.64 (68/137)
CMV Negative	50.36 (69/137)
Donor CMV Status	
CMV Positive	42.34 (58/137)
CMV Negative	57.66 (79/137)
Waiting Time to Transplant	275.09 ± 364.66
Hospital	
The Ohio State University	55.47 (76/137)
Alfred Hospital	12.41 (17/137)
St. Vincent's Hospital	16.79 (23/137)
Hahnemann University	15.33 (21/137)
MTX used as a GVHP	
Yes	29.20 (40/137)
No	70.80 (97/137)

SOURCE: Data in this table are obtained from Klein and Moeschberger (2003, p.483–487).

Table 8: Data analysis for the bone marrow transplant study

Variables	UNR method			CMM method		
	Estimate	std	95% CI	Estimate	std	95% CI
FAB	0.8364^*	0.2721	[0.3031, 1.3698]	0.8374^*	0.1252	[0.5920, 1.0828]
AMLlow	-1.0898^*	0.3826	[−1.8396, −0.3400]	-1.0906^*	0.1814	[−1.4462, −0.7351]
AMLhigh	−0.4053	0.3781	[−1.1465, 0.3358]	-0.4039^*	0.0964	[−0.5928, −0.2150]
DonAge	0.0040	0.0201	[−0.0354, 0.0433]	0.0039	0.0184	[−0.0322, 0.0399]
RecAge	0.0069	0.0205	[−0.0333, 0.0470]	0.0068	0.0187	[−0.0299, 0.0435]
DRAge	0.0032^*	0.0011	[0.0011, 0.0053]	0.0032^*	0.0009	[0.0015, 0.0048]

NOTE: UNR denotes the unconstrained estimator calculated by Newton–Raphson algorithm based on the partial likelihood function. CMM denotes the constrained estimator calculated by the proposed MM algorithm. "*" indicates that the parameter estimate is significant at 5% level.

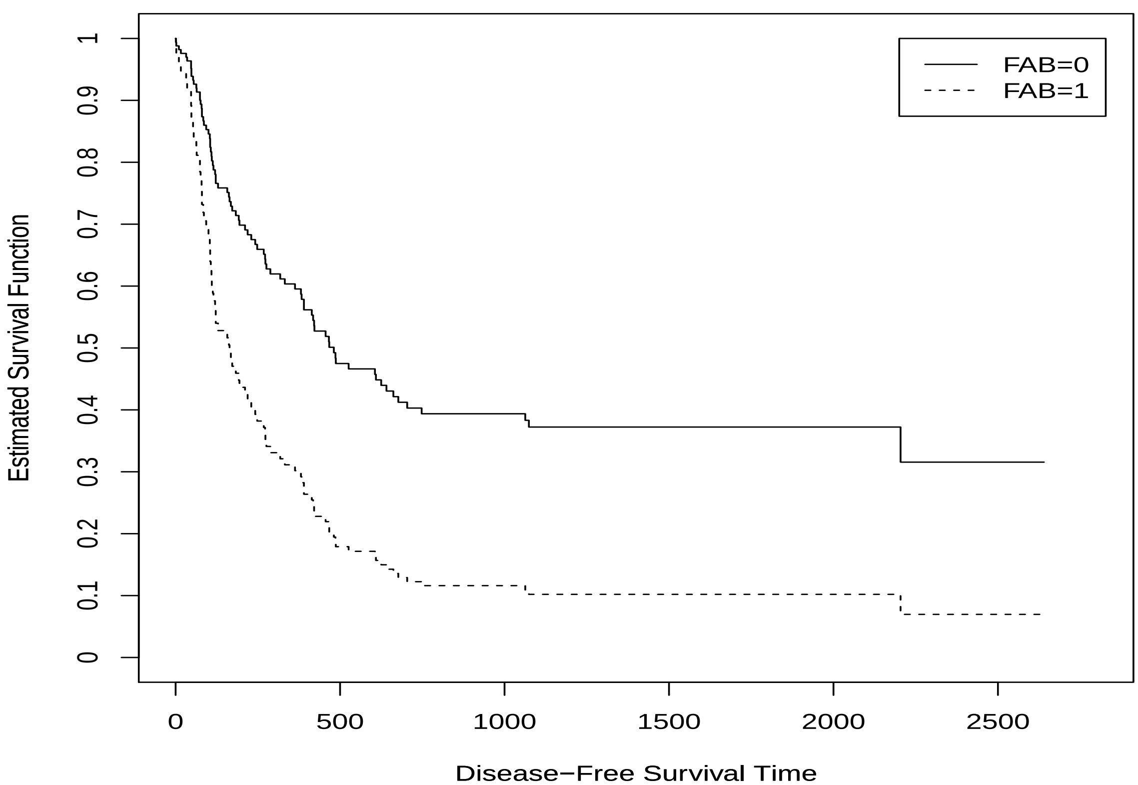

Figure 1: Estimated survival function for patients in the study of the born marrow transplant under different FAB levels.