Statistica Sinica **29** (2019), 895-916
doi:https://doi.org/10.5705/ss.202016.0516

ON PROFILE MM ALGORITHMS FOR GAMMA FRAILTY SURVIVAL MODELS

Xifen Huang, Jinfeng Xu and Guoliang Tian

Yunnan Normal University, The University of Hong Kong and Southern University of Science and Technology

Abstract: Gamma frailty survival models have been extensively used for the analysis of such multivariate failure time data as clustered failure times and recurrent events. Estimation and inference procedures in these models often center on the nonparametric maximum likelihood method and its numerical implementation via the EM algorithm. Despite its success in dealing with incomplete data problems, the algorithm may not fare well in high-dimensional situations. To address this problem, we propose a class of profile MM algorithms with good convergence properties. As a key step in constructing minorizing functions, the high-dimensional objective function is decomposed as a sum of separable low-dimensional functions. This allows the algorithm to bypass the difficulty of inverting large matrix and facilitates its pertinent use in high-dimensional problems. Simulation studies show that the proposed algorithms perform well in various situations and converge reliably with practical sample sizes. The method is illustrated using data from a colorectal cancer study.

Key words and phrases: MM algorithm, nonparametric maximum likelihood, survival data.

1. Introduction

In many biomedical studies involving failure data, there may be more than one failure time on each study subject or study subjects having univariate failure times may be grouped in a manner that leads to dependencies within groups (Kalbfleisch and Prentice (2002)). This gives rise to multivariate failure time data or clustered failure time data. In such contexts, it is of interest to assess the strength and nature of dependencies among multiple failure times. Shared frailty or random effect models have been commonly used to account for the dependence of correlated failure times (Clayton (1978); Clayton and Cuzick (1985); Oakes (1989); Zeng, Chen and Ibrahim (2009)). In particular, the proportional hazards model (Cox (1972)) with gamma frailty was used to incorporate covariates by Nielsen et al. (1992), Klein (1992), and Andersen et al. (1997). Shared frailty or random-effects have also been used to jointly model both recurrent events and the terminal event (Liu, Wolfe and Huang (2004); Zeng and Lin (2007); Zeng and Lin (2009) and Zeng, Lin and Lin (2008)). The computation involved in frailty models with survival data is usually intensive since the unknown parameters characterizing the nonparametric baseline cumulative hazard function is of the same magnitude as the sample size and hence large. The existing approaches rely on the EM algorithms which uses Newton's method and involves matrix inversion and may not perform well in such settings.

As a generalization of the EM algorithm (Dempster, Laird and Rubin (1977)), the *minorization–maximization* (MM) algorithm (Becker, Yang and Lange (1997); Lange, Hunter and Yang (2000)) increases the likelihood at each iteration and reliably converges to the maximum from well-chosen initial values (Hunter and Lange (2004)). The MM principle is an important and useful tool for optimization problems and has a broad range of applications in statistics because of its conceptual simplicity, ease of implementation and numerical stability. The MM principle has been applied in quantile regressions (Hunter and Lange (2000)), the Bradley–Terry model (Hunter (2004)), variable selection (Hunter and Li (2005); Yen (2011)), constrained estimation (Mkhadri, N'Guessan and Hafidi (2010)), sparse logistic PCA (Lee and Huang (2013)), distance majorization (Chi, Zhou and Lange (2014)), and the generalized heron problem (Chi and Lange (2014)). For a more detailed review, we refer to a recent discussion paper (Lange, Chi and Zhou (2014)). In this paper, we propose a class of profile MM algorithms for gamma frailty models with survival data. As a key step in constructing minorizing functions, the high-dimensional objective function is decomposed into separable low-dimensional functions. This allows the algorithms to bypass the difficulty of inverting large matrix and facilitate their pertinent use in high-dimensional situations. Furthermore, as pointed out by a referee, the decomposition meshes well with the regularized estimation in sparse high-dimensional models, as demonstrated in our numerical studies in Section 5.

The rest of the paper is organized as follows. In Section 2, we introduce gamma frailty survival models. Section 3 presents three profile MM algorithms. In Section 4, we establish the convergence properties of these algorithms under mild regularity conditions. Section 5 provides simulation studies to assess their practical performance. Section 6 illustrates the method using data from a colorectal cancer study. Some concluding remarks and discussions are given in Section 7.

2. Gamma Frailty Survival Models

For ease of exposition, we illustrate the proposed method with clustered failure time data although a parallel approach can be similarly developed for other types of data. Let T_{ij}, C_{ij} and $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijq})^T$ denote the survival time, the censoring time, and a vector of covariates, respectively, for the j-th individual in the i-th cluster, for $j = 1, \ldots, M_i$, and $i = 1, \ldots, B$. We assume that the right-censoring is noninformative so that C_{ij} is independent of T_{ij} given $\mathbf{X}_{ij}$. Data consist of $Y_{\text{obs}} = \{(Y_{ij} = T_{ij} \wedge C_{ij}, I_{ij}, \mathbf{X}_{ij}), i = 1, \ldots, B, j = 1, \ldots, M_i\}$, where Y_{ij} is the observed time and $I_{ij} = I(T_{ij} \le C_{ij})$ is the censoring indicator. Conditional on a cluster-specific frailty ω_i, the frailty model postulates that the instantaneous hazard rate function of T_{ij} is

$$\lambda(t|X_{ij}, \omega_i) = \lim_{\Delta t \to 0} \frac{P(t \le T_{ij} < t + \Delta t | T_{ij} \ge t, \mathbf{X}_{ij}, \omega_i)}{\Delta t} = \lambda_0(t) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}) \omega_i, \tag{2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard rate and $\boldsymbol{\beta}$ is a vector of unknown regression parameters. We assume that the frailty ω has a gamma distribution with mean 1, variance θ and density

$$g(\omega) = \frac{\omega^{1/(\theta-1)} \exp\left(-\omega/\theta\right)}{\Gamma\left(1/\theta\right) \theta^{1/\theta}}, \quad \theta > 0.$$

Here θ measures the heterogeneity between clusters and a larger θ indicates a stronger intra-cluster dependence.

The model parameters consist of $\theta, \boldsymbol{\beta}$, and the nonparametric component $\lambda_0(\cdot)$. The estimation and inferences in this model center on the nonparametric maximum likelihood method. For the asymptotic properties of the nonparametric maximum likelihood estimator, see for example, Murphy (1995), Parner (1998), and Zeng and Lin (2007).

3. A Class of Profile MM Algorithms

The MM principle provides a powerful tool for developing optimization algorithms. Due to its flexibility in constructing minorizing functions, a high-dimensional objective function can be decomposed into separable low-dimensional functions which leads to numerically convenient solutions in the maximization step. This motivates us to develop a class of MM algorithms for gamma frailty survival models.

3.1. The minorization–maximization principle

We first briefly review the minorization–maximization (MM) principle. Let $\ell(\boldsymbol{\alpha})$ be the objective function to be maximized, where $\boldsymbol{\alpha}$ denotes the unknown vector of parameters, $\boldsymbol{\alpha} \in \Theta$, and Θ the parameter space. The MM method iterates between the minorization step and the maximization step until convergence. The minorization step first constructs a surrogate function $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)})$ such that

$$Q\left(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)}\right) \leq \ell(\boldsymbol{\alpha}), \quad \forall\, \boldsymbol{\alpha}, \boldsymbol{\alpha}^{(t)} \in \Theta \quad \text{and} \quad Q\left(\boldsymbol{\alpha}^{(t)}|\boldsymbol{\alpha}^{(t)}\right) = \ell\left(\boldsymbol{\alpha}^{(t)}\right), \tag{3.1}$$

where $\boldsymbol{\alpha}^{(t)}$ denotes the current estimate of $\hat{\boldsymbol{\alpha}}$ in the t-th iteration. Here the $Q(\cdot|\boldsymbol{\alpha}^{(t)})$ function always lies under $\ell(\cdot)$ and is tangent to it at the point $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$. The maximization step then updates $\boldsymbol{\alpha}^{(t)}$ by $\boldsymbol{\alpha}^{(t+1)}$ which maximizes the surrogate function $Q(\cdot|\boldsymbol{\alpha}^{(t)})$ instead of $\ell(\boldsymbol{\alpha})$. Thus

$$\ell\left(\boldsymbol{\alpha}^{(t+1)}\right) \geq Q\left(\boldsymbol{\alpha}^{(t+1)}|\boldsymbol{\alpha}^{(t)}\right) \geq Q\left(\boldsymbol{\alpha}^{(t)}|\boldsymbol{\alpha}^{(t)}\right) = \ell\left(\boldsymbol{\alpha}^{(t)}\right).$$

The MM algorithm increases the objective function at each iteration and possesses the *ascent property* driving the target function $\ell(\boldsymbol{\alpha})$ uphill.

3.2. A profile MM algorithm

We propose MM algorithms for the nonparametric maximum likelihood estimation in the gamma frailty survival models. Under the assumption that the censoring time is independent of the failure time and the frailty given the covariates, the log likelihood function is

$$\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}}) = \sum_{i=1}^{B} \log \int_0^{+\infty} \tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0)\, \mathrm{d}\omega_i,$$

where

$$\begin{aligned}\tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0) = {} & \frac{\omega_i^{1/(\theta-1)} \exp\left(-\omega_i/\theta\right)}{\Gamma\left(1/\theta\right) \theta^{1/\theta}} \\ & \times \prod_{j=1}^{M_i} \left\{\lambda_0(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right\}^{I_{ij}} \exp\left(-\Lambda_0(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right).\end{aligned}$$

If

$$v_i\left(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right) = \frac{\tau_i\left(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right)}{\int_0^{+\infty} \tau_i\left(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right)\, \mathrm{d}\omega_i}.$$

Then

$$\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}})$$

$$= \sum_{i=1}^{B} \log \int_0^{+\infty} \left\{ v_i\left(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right) \frac{\tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0)}{v_i\left(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right)} \right\} \mathrm{d}\omega_i. \tag{3.2}$$

By Jensen's inequality,

$$\varphi\Big(\int_{\mathbb{X}} f(x)\cdot g(x)\,\mathrm{d}x\Big) \geq \int_{\mathbb{X}} \varphi\big(f(x)\big)\cdot g(x)\,\mathrm{d}x,$$

where $\mathbb{X}$ is a subset of the real line $\mathbb{R}$, $\varphi()$ is a concave function, $f(\cdot)$ is an arbitrary real-valued function defined on $\mathbb{X}$ and $g(\cdot)$ is a density function defined on $\mathbb{X}$. Noticing that $v_i(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ is a density function, we apply Jensen's inequality to (3.2). By calculation, we construct the surrogate function for $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}})$,

$$Q_1\Big(\theta, \boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\Big) = Q_{11}\Big(\theta|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\Big) + Q_{12}\Big(\boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\Big),$$

where

$$\begin{aligned} &Q_{11}\left(\theta|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right) \\ &= \sum_{i=1}^{B} \left[\frac{1}{\theta}\left\{\psi\left(A_i^{(k)}\right) - \log\left(\Pi_i^{(k)}\right)\right\} - \frac{A_i^{(k)}}{\Pi_i^{(k)}\theta} - \log\Gamma\left(\frac{1}{\theta}\right) - \frac{\log(\theta)}{\theta}\right], \end{aligned} \tag{3.3}$$

$$\begin{aligned} &Q_{12}\left(\boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right) \\ &= \sum_{i=1}^{B}\sum_{j=1}^{M_i} \left\{ I_{ij}\left(\log(\lambda_0(t_{ij})) + \mathbf{X}_{ij}^T\boldsymbol{\beta}\right) - \frac{A_i^{(k)}}{\Pi_i^{(k)}}\Lambda_0(t_{ij})\exp(\mathbf{X}_{ij}^T\boldsymbol{\beta}) \right\}, \end{aligned} \tag{3.4}$$

with

$$A_i^{(k)} = D_i + \frac{1}{\theta^{(k)}}, \quad \Pi_i^{(k)} = \frac{1}{\theta^{(k)}} + \sum_{j=1}^{M_i} \Lambda_0^{(k)}(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}^{(k)}\right).$$

The surrogate function $Q_1(\theta, \boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ separates the parameters θ and $(\boldsymbol{\beta}, \Lambda_0)$ into (3.3) and (3.4), respectively. In the maximization step, updating θ is straightforward while it is more challenging to update $(\boldsymbol{\beta}, \Lambda_0)$ due to the presence of the nonparametric component Λ_0. Following Johansen (1983), and as in Klein (1992), we consider the profile estimation approach and first profile out Λ_0 in $Q_{12}(\boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ for any given $\boldsymbol{\beta}$. This gives the estimate of Λ_0 given $\boldsymbol{\beta}$ as

$$\mathrm{d}\hat{\Lambda}_0(t_{ij}) = \frac{I_{ij}}{\sum_{r=1}^{B}\left(A_r^{(k)}/\Pi_r^{(k)}\right)\sum_{s=1}^{M_r} I(t_{rs} \geq t_{ij})\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})}. \tag{3.5}$$

Substituting (3.5) into $Q_{12}(\boldsymbol{\beta}, \Lambda_0|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ yields the function

$$Q_{13}\left(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)}\right)$$
$$=\sum_{i=1}^{B}\sum_{j=1}^{M_i}\left\{I_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta}-I_{ij}\log\left(\sum_{r=1}^{B}\frac{A_r^{(k)}}{\Pi_r^{(k)}}\sum_{s=1}^{M_r}I(t_{rs}\geq t_{ij})\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})\right)\right\}, \tag{3.6}$$

which involves only $\boldsymbol{\beta}$. It is easy to see that $Q_{13}(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$ is a concave function of $\boldsymbol{\beta}$ and takes the form of the log partial likelihood in the Cox model. Standard Cox regression programs can be used to solve it. This MM algorithm much resembles its EM counterpart (Klein (1992)) as they utilize similar minorizing and profiling steps. We refer to this MM algorithm as MM1. The algorithm is stated as follows.

Step 1. Let $(\theta^{(0)},\boldsymbol{\beta}^{(0)},\Lambda_0^{(0)})$ be initial values of $(\theta,\boldsymbol{\beta},\Lambda_0)$.

Step 2. Update the estimate of θ via maximizing (3.3). Update the estimate of $\boldsymbol{\beta}$ using a standard Cox regression program to maximize (3.6).

Step 3. Using the updated estimate of $\boldsymbol{\beta}$, compute the estimate of $\Lambda_0(t_{ij})$ via (3.5).

Step 4. Iterate steps 2 and 3 until convergence.

3.3. A second profile MM algorithm

The MM1 or its EM counterpart relies on the fact that, after profiling out Λ_0, the resulting function such as (3.6) is concave. When this does not hold, directly using Newton's method to maximize $Q_{13}(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$ is difficult especially when there exist a large number of covariates. In such situations, it is of interest to develop MM algorithms that can avoid the concavity requirement and bypass Newton's method and matrix inversion. This is where the MM principle best exhibits its advantages. To maximize $Q_{13}(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$, we further construct minorizing functions to decompose the high-dimensional maximization into separate low-dimensional ones. We first utilize the supporting hyperplane inequality

$$-\log(x)\geq-\log(x_0)-\frac{x-x_0}{x_0} \tag{3.7}$$

to minorize $Q_{13}(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$ by the surrogate function

$$Q_{14}(\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$$
$$=\sum_{i=1}^{B}\sum_{j=1}^{M_i}\left\{I_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta}-\frac{I_{ij}\sum_{r=1}^{B}\left(A_r^{(k)}/\Pi_r^{(k)}\right)\sum_{s=1}^{M_r}I(t_{rs}\geq t_{ij})\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})}{\sum_{r=1}^{B}\left(A_r^{(k)}/\Pi_r^{(k)}\right)\sum_{s=1}^{M_r}I(t_{rs}\geq t_{ij})\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)})}\right\}+c,$$

where c is a constant not depending on $\boldsymbol{\beta}$. As in Ding, Tian and Yuen (2015), we next apply Jensen's inequality to the concave function $-\exp(\cdot)$ in $Q_{14}(\boldsymbol{\beta}|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ by rewriting

$$\mathbf{X}_{rs}^T\boldsymbol{\beta} = \sum_{p=1}^{q} \delta_{prs}\big(\delta_{prs}^{-1} X_{prs}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^{(k)}) + \mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)}\big),$$

where $\delta_{prs} = |X_{prs}|/\sum_{p=1}^{q}|X_{prs}|$. In the end, the minorizing function for $Q_{13}(\boldsymbol{\beta}|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ is

$$Q_{15}\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right) \hat{=} \sum_{p=1}^{q} Q_{15p}\left(\boldsymbol{\beta}_p|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right), \qquad (3.8)$$

where

$$Q_{15p}(\boldsymbol{\beta}_p|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) = \sum_{i=1}^{B}\sum_{j=1}^{M_i}\Bigg[I_{ij}\boldsymbol{\beta}_p X_{pij} \qquad (3.9)$$
$$-\frac{I_{ij}\sum_{r=1}^{B}\left(A_r^{(k)}/\Pi_r^{(k)}\right)\sum_{s=1}^{M_r} I(t_{rs} \geq t_{ij})\delta_{prs}\exp(\delta_{prs}^{-1}(\boldsymbol{\beta}_p - \boldsymbol{\beta}_p^{(k)})X_{prs} + \mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)})}{\sum_{r=1}^{B}\left(A_r^{(k)}/\Pi_r^{(k)}\right)\sum_{s=1}^{M_r} I(t_{rs} \geq t_{ij})\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)})}\Bigg].$$

From (3.9), it can be seen that the objective function to be maximized is decomposed into a sum of q univariate functions. The resulting MM algorithm only involves $q+1$ separate univariate optimizations in its maximization step and matrix inversion is not needed. We refer to this algorithm as MM2. The algorithm is stated as follows.

Step 1. Let $(\theta^{(0)}, \boldsymbol{\beta}^{(0)}, \Lambda_0^{(0)})$ be initial values of $(\theta, \boldsymbol{\beta}, \Lambda_0)$.

Step 2. Update the estimate of θ via (3.3). Update the estimate of $\boldsymbol{\beta}_p$ based on (3.9) for $p = 1, \ldots, q$.

Step 3. Using the updated estimate of $\boldsymbol{\beta}$, compute the estimate of $\Lambda_0(t_{ij})$ via (3.5).

Step 4. Iterate steps 2 and 3 until convergence.

3.4. A third profile MM algorithm

MM1 and MM2 are developed regardless of whether an analytic form of $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0\,|Y_{\text{obs}})$ is available or not. the integral is tractable and $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}})$ can be explicitly written as

$$\ell_2(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}})$$

$$= \sum_{i=1}^{B} \left[\sum_{j=1}^{M_i} \left\{ I_{ij} \log(\lambda_0(t_{ij})) + I_{ij} \mathbf{X}_{ij}^T \boldsymbol{\beta} \right\} + \log \Gamma\left(D_i + \frac{1}{\theta} \right) - \log \Gamma\left(\frac{1}{\theta} \right) - \frac{\log(\theta)}{\theta} \right.$$
$$\left. - \left(D_i + \frac{1}{\theta} \right) \log \left(\frac{1}{\theta} + \sum_{j=1}^{M_i} \Lambda_0(t_{ij}) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}) \right) \right],$$

where $D_i = \sum_{j=1}^{M_i} I_{ij}$ is the observed number of deaths in the i-th cluster. We develop an MM algorithm based on $\ell_2(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}})$ for the nonparametric maximum likelihood estimation of $(\theta, \boldsymbol{\beta}, \Lambda_0)$. Use the inequality (3.7) to minorize the last term of $\ell_2(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}})$ and obtain the surrogate function

$$Q^*(\theta, \boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) \tag{3.10}$$
$$= \sum_{i=1}^{B} \sum_{j=1}^{M_i} \left[I_{ij} \left(\log(\lambda_0(t_{ij})) + \mathbf{X}_{ij}^T \boldsymbol{\beta} \right) - \left(\frac{D_i}{\Pi_i^{(k)}} + \frac{1}{\Pi_i^{(k)} \theta} \right) \Lambda_0(t_{ij}) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}) \right]$$
$$+ \sum_{i=1}^{B} \left[\log \Gamma\left(D_i + \frac{1}{\theta} \right) - \log \Gamma\left(\frac{1}{\theta} \right) - \frac{\log(\theta)}{\theta} + \frac{1}{\theta} \left\{ 1 - \log\left(\Pi_i^{(k)} \right) - \frac{D_i}{\Pi_i^{(k)}} \right\} - \frac{1}{\Pi_i^{(k)} \theta^2} \right],$$

where $\Pi_i^{(k)} = 1/\theta^k + \sum_{j=1}^{M_i} \Lambda_0^{(k)}(t_{ij}) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}^{(k)})$. Next, profile out Λ_0 for any given $(\boldsymbol{\beta}, \theta)$ and estimate $\Lambda_0(t_{ij})$ by

$$\mathrm{d}\hat{\Lambda}_0(t_i j) = \frac{I_{ij}}{\sum_{r=1}^{B} \sum_{s=1}^{M_r} I(t_{rs} \geq t_{ij}) \left(D_r / \Pi_r^{(k)} + 1/\Pi_r^{(k)} \theta \right) \exp(\mathbf{X}_{rs}^T \boldsymbol{\beta})}. \tag{3.11}$$

Substituting (3.11) into (3.10) yields the surrogate function

$$Q_1^* \left(\theta, \boldsymbol{\beta} | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)} \right) \tag{3.12}$$
$$= \sum_{i=1}^{B} \left[\log\left(\Gamma\left(D_i + \frac{1}{\theta} \right) \right) - \log\left(\Gamma\left(\frac{1}{\theta} \right) \right) - \frac{\log(\theta)}{\theta} + \frac{1}{\theta} \left\{ 1 - \log\left(\Pi_i^{(k)} \right) - \frac{D_i}{\Pi_i^{(k)}} \right\} - \frac{1}{\Pi_i^{(k)} \theta^2} \right]$$
$$+ \sum_{i=1}^{B} \sum_{j=1}^{M_i} I_{ij} \left(\mathbf{X}_{ij}^T \boldsymbol{\beta} - \log\left(\sum_{r=1}^{B} \sum_{s=1}^{M_r} I(t_{rs} \geq t_{ij}) \left(\frac{D_r}{\Pi_r^{(k)}} + \frac{1}{\Pi_r^{(k)} \theta} \right) \exp(\mathbf{X}_{rs}^T \boldsymbol{\beta}) \right) \right),$$

that only involves paramters $\boldsymbol{\beta}$ and θ. From (3.12), the surrogate function $Q_1^*(\theta, \boldsymbol{\beta} | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ resulting from the profiling step does not take the form of the log-partial likelihood in the Cox model and hence standard Cox regression programs not can be used to solve the function.

To construct a minorizing function for $Q_1^*(\theta, \boldsymbol{\beta} | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$, we first apply the supporting hyperplane inequality (3.7) to the last term. This gives the surrogate function

$$
\begin{aligned}
&Q_2^*(\theta, \boldsymbol{\beta}|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})\\
&= \sum_{i=1}^{B}\left[\log\left(\Gamma\left(D_i+\frac{1}{\theta}\right)\right)-\left(\log\Gamma\left(\frac{1}{\theta}\right)\right)-\frac{\log(\theta)}{\theta}+\frac{1}{\theta}\left\{1-\log\left(\Pi_i^{(k)}\right)-\frac{D_i}{\Pi_i^{(k)}}\right\}-\frac{1}{\Pi_i^{(k)}\theta^2}\right]\\
&\quad+\sum_{i=1}^{B}\sum_{j=1}^{M_i} I_{ij}\left((\mathbf{X}_{ij}^T\boldsymbol{\beta}-\frac{1}{u_{ij}^{(k)}}\sum_{r=1}^{B}\sum_{s=1}^{M_r} I(t_{rs}\geq t_{ij})\left(\frac{D_r}{\Pi_r^{(k)}}+\frac{1}{\Pi_r^{(k)}\theta}\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\right),
\end{aligned}
$$

where

$$
u_{ij}^{(k)} = \sum_{r=1}^{B}\sum_{s=1}^{M_r} I(t_{rs}\geq t_{ij})\left(\frac{D_r}{\Pi_r^{(k)}}+\frac{1}{\Pi_r^{(k)}\theta^{(k)}}\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)}\right).
$$

For the term $-\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})/\theta$ with negative coefficients, as in Lange and Zhou (2014), we use the arithmetic-geometric mean inequality

$$
-\prod_{i=1}^{n} x_i^{a_i} \geq -\sum_{i=1}^{n}\frac{a_i}{||\mathbf{a}||_1}x_i^{||\mathbf{a}||_1}, \tag{3.13}
$$

where x_i and a_i are nonnegative. Now choosing $x_1=\theta^{(k)}/\theta$ and $x_2=\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})/\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)})$ in (3.13), we obtain the surrogate function

$$
\begin{aligned}
&Q_3^*\left(\theta, \boldsymbol{\beta}|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\right)\\
&= \sum_{i=1}^{B}\left[\log\left(\Gamma\left(D_i+\frac{1}{\theta}\right)\right)-\log\left(\Gamma\left(\frac{1}{\theta}\right)\right)-\frac{\log(\theta)}{\theta}+\frac{1}{\theta}\left\{1-\log(\Pi_i^{(k)})-\frac{D_i}{\Pi_i^{(k)}}\right\}-\frac{1}{\Pi_i^{(k)}\theta^2}\right.\\
&\quad\left.-\sum_{j=1}^{M_i}\frac{I_{ij}}{u_{ij}^{(k)}}\sum_{r=1}^{B}\sum_{s=1}^{M_r}\frac{I(t_{rs}\geq t_{ij})\theta^{(k)}\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)}\right)}{2\Pi_r^{(k)}\theta^2}\right]\\
&\quad+\sum_{i=1}^{B}\sum_{j=1}^{M_i} I_{ij}\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}-\frac{1}{u_{ij}^{(k)}}\sum_{r=1}^{B}\sum_{s=1}^{M_r} I(t_{rs}\geq t_{ij})\frac{D_r}{\Pi_r^{(k)}}\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})\right.\\
&\quad\left.-\frac{1}{u_{ij}^{(k)}}\sum_{r=1}^{B}\sum_{s=1}^{M_r} I(t_{rs}\geq t_{ij})\frac{\exp\left(2\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)}{2\Pi_r^{(k)}\theta^{(k)}\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)}\right)}\right),
\end{aligned} \tag{3.14}
$$

where θ and $\boldsymbol{\beta}$ are separated. To separate the parameters $\beta_1,\ldots,\beta_q$, we further minorize the concave functions $-\exp(\mathbf{X}_{rs}^T\boldsymbol{\beta})$ and $-\exp(2\mathbf{X}_{rs}^T\boldsymbol{\beta})$ in (3.14) using Jensen's inequality by rewriting $\mathbf{X}_{rs}^T\boldsymbol{\beta}=\sum_{p=1}^{q}\delta_{prs}(\delta_{prs}^{-1}X_{prs}(\boldsymbol{\beta}_p-\boldsymbol{\beta}_p^{(k)})+\mathbf{X}_{rs}^T\boldsymbol{\beta}^{(k)})$, $\delta_{prs}=|X_{prs}|/\sum_{p=1}^{q}|X_{prs}|$. Denote the resulting minorizing function for $Q_1^*(\theta,\boldsymbol{\beta}|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$ by $Q(\theta,\beta_1,\ldots,\beta_q|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})$. By calculation,

$$
Q(\theta,\beta_1,\ldots,\beta_q|\theta^{(k)},\boldsymbol{\beta}^{(k)},\Lambda_0^{(k)})
$$

$$= Q_1(\theta|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) + \sum_{p=1}^{q} Q_{2p}(\beta_p|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}), \tag{3.15}$$

where

$$\begin{aligned} &Q_1(\theta|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) \\ &= \sum_{i=1}^{B} \Bigg[\log\Gamma\left(D_i + \frac{1}{\theta}\right) - \log\Gamma\left(\frac{1}{\theta}\right) - \frac{\log(\theta)}{\theta} + \frac{1}{\theta}\left\{1 - \log(\Pi_i^{(k)}) - \frac{D_i}{\Pi_i^{(k)}}\right\} \\ &\quad - \frac{1}{\Pi_i^{(k)}\theta^2} - \sum_{j=1}^{M_i} \frac{I_{ij}}{u_{ij}^{(k)}} \sum_{r=1}^{B}\sum_{s=1}^{M_r} \frac{I(t_{rs} \geq t_{ij})\theta^{(k)} \exp\left(\mathbf{X}_{rs}^T \boldsymbol{\beta}^{(k)}\right)}{2\Pi_r^{(k)}\theta^2} \Bigg], \end{aligned} \tag{3.16}$$

$$\begin{aligned} Q_{2p}(\beta_p|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) &= \sum_{i=1}^{B}\sum_{j=1}^{M_i} \Bigg\{ I_{ij}\beta_p X_{pij} \\ &- \frac{I_{ij}}{u_{ij}^{(k)}} \sum_{r=1}^{B}\sum_{s=1}^{M_r} \frac{I(t_{rs} \geq t_{ij}) D_r \delta_{prs}}{\Pi_r^{(k)}} \exp\left(\delta_{prs}^{-1} X_{prs}\left(\beta_p - \beta_p^{(k)}\right) + \mathbf{X}_{rs}^T \boldsymbol{\beta}^{(k)}\right) \\ &- \frac{I_{ij}}{u_{ij}^{(k)}} \sum_{r=1}^{B}\sum_{s=1}^{M_r} \frac{I(t_{rs} \geq t_{ij}) \delta_{prs} \exp\left(\delta_{prs}^{-1} 2X_{prs}\left(\beta_p - \beta_p^{(k)}\right) + 2\mathbf{X}_{rs}^T \boldsymbol{\beta}^{(k)}\right)}{2\Pi_r^{(k)}\theta^{(k)} \exp(\mathbf{X}_{rs}^T \boldsymbol{\beta}^{(k)})} \Bigg\}, \end{aligned} \tag{3.17}$$

for $p = 1, \ldots, q$. By its construction, the frailty parameter θ and the regression parameters $\beta_1, \ldots, \beta_q$ are separated from each other in (3.15). Accordingly, the maximization step involves $q + 1$ separate univariate optimizations. We refer to this algorithm as MM3. The algorithm is stated as follows.

Step 1. Let $(\theta^{(0)}, \boldsymbol{\beta}^{(0)}, \Lambda_0^{(0)})$ be initial values of $(\theta, \boldsymbol{\beta}, \Lambda_0)$.

Step 2. Update the estimate of θ via (3.16). Update the estimate of β_p based on (3.17) for $p = 1, \ldots, q$.

Step 3. Using the updated estimate of θ and $\boldsymbol{\beta}$, compute the estimate of $\Lambda_0(t_{ij})$ via (3.11).

Step 4. Iterate steps 2 and 3 until convergence.

4. Convergence Properties of the Proposed MM Algorithms

In this section, we establish convergence properties of the three MM algorithms. We first present a lemma (Vaida (2005)) which gives general and verifiable conditions for proving the convergence of an MM sequence. Let $\ell(\cdot|Y_{\text{obs}})$ be the function to maximize and $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ be the minorizing function, where $\boldsymbol{\alpha}$ is the parameter vector and $\boldsymbol{\alpha}^{(k)}$ is its current estimate. Denote the maximizer of

$Q(\cdot|\boldsymbol{\alpha})$ by $M(\boldsymbol{\alpha})$. We need some regularity conditions.

C1. The parameter space Ω is an open set in $\mathbb{R}^d$.

C2. $\ell(\cdot|Y_{\text{obs}})$ is differentiable, with continuous derivative $\ell'(\cdot|Y_{\text{obs}})$.

C3. The level set $\Omega_c = \{\boldsymbol{\alpha} \in \Omega : \ell(\boldsymbol{\alpha}|Y_{\text{obs}}) \geq c\}$ is compact in $\mathbb{R}^d$.

C4. $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ is continuous in both $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{(k)}$, and differentiable in $\boldsymbol{\alpha}$.

C5. All the stationary points of $\ell(\cdot|Y_{\text{obs}})$ are isolated.

C6. There exists a unique global maximum of $Q(\cdot|\boldsymbol{\alpha}^{(k)})$.

Lemma 1. *(*Vaida (2005)*). Let $\boldsymbol{\alpha}^{(k)}, k = 0, 1, 2, ...$ denote an MM sequence.*
(i) If **C6** *holds, then $M(\cdot)$ is continuous at $\boldsymbol{\alpha}^{(k)}$.*
(ii) If **C1-C6** *hold, then for any starting value $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\alpha}^{(k)} \to \boldsymbol{\alpha}^*$ when $k \to \infty$, for some stationary point $\boldsymbol{\alpha}^*$. Moreover, $M(\boldsymbol{\alpha}^*) = \boldsymbol{\alpha}^*$, and if $\boldsymbol{\alpha}^{(k)} \neq \boldsymbol{\alpha}^*$ for all k, the sequence of likelihood values $\ell(\boldsymbol{\alpha}^{(k)}|Y_{obs})$ strictly increases to $\ell(\boldsymbol{\alpha}^*|Y_{obs})$.*

For the convergence of our MM algorithms, we need a condition.

Condition A.

(i). $\max_{1\leq i\leq B} D_i \geq 1$.

(ii). Let $O_0 = \{1, 2, ..., q\}$. For any $O \subset O_0$, there exist the pairs (i, j), (i_1, j_1) and (i_2, j_2) such that $I_{ij} = 1$, $t_{i_1 j_1} \geq t_{ij}$, $t_{i_2 j_2} \geq t_{ij}$, and for any $r \in O$ and $s \in O^c = O_0 - O$, $X_{i_1 j_1 r} - X_{ijr} > 0$ and $X_{i_2 j_2 s} - X_{ijs} < 0$.

(iii). Stationary points for $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{obs})$ are separated.

Remark 1. Condition A (i) ensures the boundedness of θ. Condition A (ii) excludes the situation where the parameters $\beta_p, p = 1, \ldots, q$ can be $-\infty$ or ∞ and hence guarantees the boundedness of $\beta_p, p = 1, \ldots, q$. Condition A (iii) corresponds to condition (C5).

Theorem 1. *If Condition A holds, for any initial value $\{\theta^{(0)}, \boldsymbol{\beta}^{(0)}, \Lambda_0^{(0)}\}$ the sequence $\{\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\}$ generated by the MM algorithm that updates the estimates by* (3.3)*,* (3.5) *and* (3.6) *are convergent.*

Proof of Theorem 1. By the construction of MM1,

$$
\begin{aligned}
&\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0|Y_{\text{obs}}) \\
&= \sum_{i=1}^{B} \log \int_0^{+\infty} \tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0)\, \mathrm{d}\omega_i \\
&= \sum_{i=1}^{B} \log \int_0^{+\infty} \left\{ v_i(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) \cdot \frac{\tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0)}{v_i(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})} \right\} \mathrm{d}\omega_i \\
&\geq \sum_{i=1}^{B} \int_0^{+\infty} v_i(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}) \cdot \log \left(\frac{\tau_i(\omega_i|\theta, \boldsymbol{\beta}, \Lambda_0)}{v_i(\omega_i|\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})} \right) \mathrm{d}\omega_i
\end{aligned}
$$

$$= Q_1(\theta, \boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}),$$

where the minorizing function $Q_1(\theta, \boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ satisfies the conditions in (3.1),

$$\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}}) \geq Q_1(\theta, \boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}), \quad \forall\, \theta, \boldsymbol{\beta}, \Lambda_0 \quad \text{and}$$
$$\ell_1(\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)} | Y_{\text{obs}}) = Q_1(\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)} | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}).$$

The initial minorizing function consists of $Q_{11}(\theta | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ and $Q_{12}(\boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$. After profiling out Λ_0, $Q_{12}(\boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ is $Q_{13}(\boldsymbol{\beta} | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ in (3.6), a unimodal function.This shows that $Q_{12}(\boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$ has a unique global maximum and verifies the condition **C6**. Conditions **C1, C2**, and **C4** easily follow from the forms of $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}})$ and $Q_1(\theta, \boldsymbol{\beta}, \Lambda_0 | \theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)})$. Next, we verify the condition **C3** and show the compactness of the level set $\Omega_c = \{\boldsymbol{\beta} = (\theta, \boldsymbol{\beta}, \Lambda_0) : \ell_1(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}}) \geq c\}$. It follows from the continuity of $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}})$ that Ω_c is closed. It remains to prove the boundedness of Ω_c. We have

$$\begin{aligned}
\ell_1(\boldsymbol{\alpha} | Y_{\text{obs}}) &= \ell_2(\boldsymbol{\alpha} | Y_{\text{obs}}) \\
&= \sum_{i=1}^{B} \Bigg[\sum_{j=1}^{M_i} \{ I_{ij} \log(\lambda_0(t_{ij})) \} + \log\left(\Gamma\left(D_i + \frac{1}{\theta}\right)\right) - \log\left(\Gamma\left(\frac{1}{\theta}\right)\right) - \frac{\log(\theta)}{\theta} \\
&\quad + \sum_{j=1}^{M_i} I_{ij} \mathbf{X}_{ij}^T \boldsymbol{\beta} - \left(D_i + \frac{1}{\theta}\right) \log\left(\frac{1}{\theta} + \sum_{j=1}^{M_i} \Lambda_0(t_{ij}) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta})\right) \Bigg].
\end{aligned}$$

Thus for any value of $(\boldsymbol{\beta}, \Lambda_0)$ with $\max_{1 \leq i \leq B} D_i \geq 1$, when θ is unbounded, $\ell_1(\boldsymbol{\alpha} | Y_{\text{obs}})$ tends to minus infinity and hence, by contradiction, θ is bounded in Ω_c. Similarly, considering $\beta_p, p = 1, \ldots, q$, and $d\Lambda_0(t_{ij}), i = 1, \ldots, B; j = 1, \ldots, M_i$, when any of them is unbounded, $\ell_1(\boldsymbol{\alpha} | Y_{\text{obs}})$ tends to minus infinity regardless of the values of the other parameters. It follows that Ω_c is bounded since $\ell_1(\boldsymbol{\alpha} | Y_{\text{obs}}) \geq c$ when $\boldsymbol{\alpha} \in \Omega_c$. Thus **C3** holds. Consequently, with the assumption that stationary points for $\ell_1(\theta, \boldsymbol{\beta}, \Lambda_0 | Y_{\text{obs}})$ are separated, by Lemma 1 the MM1 algorithm is convergent.

Similarly, we establish the convergence properties for MM2 and MM3.

Theorem 2. *If Condition A holds, for any initial value* $\{\theta^{(0)}, \boldsymbol{\beta}^{(0)}, \Lambda_0^{(0)}\}$, *the sequence* $\{\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\}$ *generated by the MM algorithm that updates the estimates by* (3.3), (3.5) *and* (3.9) *are convergent.*

Theorem 3. *If Condition A holds, for any initial value* $\{\theta^{(0)}, \boldsymbol{\beta}^{(0)}, \Lambda_0^{(0)}\}$, *the sequence* $\{\theta^{(k)}, \boldsymbol{\beta}^{(k)}, \Lambda_0^{(k)}\}$ *generated by the MM algorithm that updates the esti-*

mates by (3.11)*,* (3.16) *and* (3.17) *are convergent.*

The proofs of Theorems 2 and 3 are given in the Supplementary Materials.

5. Numerical Experiments

We conducted two sets of simulation studies to assess the finite-sample performance of the proposed MM algorithms. The simulations were run in a desktop with Intel(R) Core(TM) i7-2600 and CPU 3.40 GHz. The R codes are available from the authors upon request. The stopping criterion was set to be

$$\frac{|\ell(\boldsymbol{\alpha}^{(t+1)}|Y_{\mathrm{obs}}) - \ell(\boldsymbol{\alpha}^{(t)}|Y_{\mathrm{obs}})|}{|\ell(\boldsymbol{\alpha}^{(t)}|Y_{\mathrm{obs}})| + 1} < 10^{-6}.$$

We generated $\sum_{i=1}^{B} M_i$ observations from the proportional hazards gamma frailty model (2.1), with $\lambda_0(t) = \alpha = 5$ and ω_i simulated from a gamma distribution with mean 1 and shape parameter $1/\theta$.

In the first set of simulations, we considered the non-regularized setting and the covariates $\mathbf{X}^T = (X_1, \ldots, X_q)$ were generated from independent uniform distribution between 0 and 0.5. The censoring times were generated to yield a censoring proportion of 30%. To illustrate the advantages of the proposed MM algorithms in high-dimensional settings, we let $q = 30$ or 40. The true coefficient vector $\boldsymbol{\beta}$ was set to $(-5_{10}^T, 2_{10}^T, 4_{10}^T)^T$ or $(-5_{10}^T, 2_{10}^T, 4_{10}^T, -1_{10}^T)^T$, $\theta \in \Omega_\theta = \{3, 10, 16\}$, and $(B, M) \in \Omega_{(B,M)} = \{(40, 20), (30, 30)\}$. We numerically compare the EM algorithm in Klein (1992) and the proposed MM algorithms. As there is much room for improvement for the EM and MM algorithms by using simple off-the-shelf accelerators (Varadhan and Roland (2008); Zhou, Alexander and Lange (2011)), following the suggestion of the referee, we also implemented accelerated EM, MM1, MM2 and MM3 with the squared iterative method (SqS1). Based on 500 replications, the average values of estimated regression parameters (with their empirical standard deviations in parentheses), iteration numbers (K), run times (Time) and the final objective values (L) are summarized in Tables 1-4. For the un-accelerated algorithms, all three MM algorithms converge faster than the EM algorithm and MM1 is the fastest. We also observe that accelerated algorithms indeed substantially save run times, especially MM2 and MM3. For the accelerated algorithms, all three accelerated MM algorithms converge faster than the accelerated EM algorithm and accelerated MM2 or MM3 is the fastest. In terms of estimation accuracy, the un-accelerated MM2 and MM3 algorithms perform the best, exhibiting small biases and empirical standard deviations in all situations.

Table 1. Simulation results for the non-regularized setting with $(B, M, q) = (30, 30, 30)$.

Par.	EM Original	EM SqS1	MM1 Original	MM1 SqS1	MM2 Original	MM2 SqS1	MM3 Original	MM3 SqS1
				θ=4				
K	359.67	167.23	134.56	64.74	1,618.47	79.73	1,639.17	81.55
T	384.86	325.71	151.37	145.45	248.62	23.66	246.95	28.98
L	−3,814.02	−3,814.64	−3,815.19	−3,814.67	−3,815.78	−3,814.53	−3,815.79	−3,814.75
θ	4.14(0.89)	4.06(0.86)	4.18(0.88)	4.08(0.86)	3.86(0.82)	4.11(0.83)	3.86(0.82)	4.11(0.87)
β_1	−5.13(0.35)	−5.14(0.34)	−5.15(0.34)	−5.15(0.34)	−4.87(0.32)	−5.17(0.34)	−4.87(0.32)	−5.17(0.34)
β_5	−5.12(0.37)	−5.12(0.35)	−5.13(0.35)	−5.13(0.35)	−4.86(0.33)	−5.15(0.36)	−4.86(0.33)	−5.15(0.35)
β_{10}	−5.15(0.34)	−5.14(0.37)	−5.15(0.37)	−5.15(0.37)	−4.88(0.35)	−5.16(0.36)	−4.87(0.35)	−5.17(0.37)
β_{15}	2.05(0.30)	2.05(0.31)	2.06(0.31)	2.06(0.31)	1.95(0.29)	2.06(0.31)	1.95(0.29)	2.07(0.31)
β_{20}	2.03(0.31)	2.04(0.31)	2.05(0.31)	2.04(0.31)	1.94(0.29)	2.05(0.31)	1.93(0.29)	2.05(0.31)
β_{25}	4.11(0.33)	4.12(0.33)	4.13(0.34)	4.12(0.34)	3.90(0.31)	4.13(0.33)	3.90(0.31)	4.14(0.34)
β_{30}	4.12(0.32)	4.09(0.35)	4.10(0.35)	4.10(0.35)	3.88(0.33)	4.10(0.35)	3.88(0.33)	4.11(0.35)
				θ=10				
K	529.27	243.34	75.91	47.75	1,661.73	62.54	1,673.99	68.05
T	598.03	494.14	76.05	67.01	268.82	22.97	242.47	30.55
L	−2,427.07	−2,427.02	−2,427.36	−2,427.11	−2,427.81	−2,427.18	−2,427.82	−2,427.16
θ	10.29(2.53)	10.25(2.52)	10.42(2.56)	10.29(2.52)	9.80(2.42)	10.31(2.50)	9.80(2.42)	10.30(2.50)
β_1	−5.23(0.44)	−5.23(0.43)	−5.23(0.43)	−5.23(0.43)	−4.97(0.41)	−5.22(0.43)	−4.97(0.41)	−5.22(0.43)
β_5	−5.22(0.44)	−5.22(0.44)	−5.22(0.44)	−5.22(0.44)	−4.96(0.41)	−5.21(0.44)	−4.96(0.41)	−5.21(0.44)
β_{10}	−5.24(0.45)	−5.24(0.45)	−5.24(0.45)	−5.24(0.45)	−4.98(0.42)	−5.23(0.45)	−4.98(0.42)	−5.23(0.45)
β_{15}	2.04(0.40)	2.04(0.40)	2.04(0.40)	2.04(0.40)	1.94(0.38)	2.04(0.40)	1.94(0.38)	2.04(0.40)
β_{20}	2.11(0.39)	2.11(0.39)	2.11(0.39)	2.11(0.39)	2.00(0.37)	2.11(0.39)	2.00(0.37)	2.11(0.39)
β_{25}	4.17(0.44)	4.17(0.44)	4.17(0.44)	4.17(0.44)	3.95(0.42)	4.16(0.44)	3.95(0.42)	4.16(0.44)
β_{30}	4.18(0.42)	4.18(0.42)	4.17(0.42)	4.18(0.42)	3.96(0.40)	4.17(0.42)	3.96(0.40)	4.17(0.42)
				θ=16				
K	706.06	316.76	75.44	48.22	1,729.45	59.34	1,739.75	61.29
T	757.43	656	71.96	66.49	286.69	23.53	242.88	27.39
L	−1,722.03	−1,722.00	−1,722.20	−1,722.05	−1,722.58	−1,722.11	−1,722.58	−1,722.12
θ	17.25(4.78)	17.21(4.78)	17.38(4.83)	17.22(4.77)	16.47(4.57)	17.19(4.71)	16.47(4.57)	17.19(4.71)
β_1	−5.26(0.53)	−5.26(0.53)	−5.25(0.53)	−5.26(0.53)	−4.99(0.50)	−5.23(0.52)	−4.99(0.50)	−5.22(0.52)
β_5	−5.25(0.56)	−5.25(0.56)	−5.24(0.56)	−5.24(0.56)	−4.98(0.53)	−5.21(0.55)	−4.98(0.53)	−5.21(0.55)
β_{10}	−5.30(0.56)	−5.30(0.56)	−5.29(0.56)	−5.29(0.56)	−5.03(0.52)	−5.26(0.55)	−5.03(0.52)	−5.26(0.55)
β_{15}	2.12(0.52)	2.12(0.52)	2.12(0.52)	2.12(0.52)	2.01(0.49)	2.11(0.52)	2.01(0.49)	2.11(0.52)
β_{20}	2.11(0.47)	2.11(0.47)	2.10(0.47)	2.11(0.47)	2.00(0.45)	2.10(0.47)	2.00(0.45)	2.10(0.47)
β_{25}	4.23(0.54)	4.22(0.54)	4.22(0.54)	4.22(0.54)	4.01(0.51)	4.20(0.54)	4.01(0.51)	4.20(0.54)
β_{30}	4.21(0.52)	4.20(0.52)	4.20(0.52)	4.20(0.52)	3.99(0.49)	4.18(0.52)	3.99(0.49)	4.18(0.52)

In the second set of simulation studies, we illustrate the utility of the proposed algorithms fo the regularized estimation in sparse high-dimensional regression model with clustered failure time data. We considered the MM2 and MM3 algorithms with the smoothly clipped absolute penalty (SCAD) using local quadratic approximation. For SCAD, as in Fan and Li (2002), we took $a = 3.7$ and the tuning parameter λ was selected using generalized cross-validation. The number of covariates was $q = 30$ and 50. For $q = 30$, we took $(B, M) = (30, 10)$ with the non-zero coefficients $(\beta_2, \beta_{10}, \beta_{26}) = (6, 3, 5)$. For $q = 50$, we took $(B, M) = (40, 10)$ and the nonzero coefficients as $(\beta_2, \beta_{10}, \beta_{26}, \beta_{45}, \beta_{50}) = (6, 3, 5, 2, 3)$. The X_i's were marginally standard normal and the correlation between X_i and X_j was $\rho^{|i-j|}$ with $\rho = 0.25$ or 0.75. The model error $E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}\mathbf{X}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$ was used to evaluate the estimation accuracy. We calculated the relative model error, the ratio of the model error of the regularized estimator and that of the oracle estimator. Based on 500 replications, the median of rela-

Table 2. Simulation results for the non-regularized setting with $(B, M, q) = (40, 20, 30)$.

	EM		MM1		MM2		MM3	
Par.	Original	SqS1	cOriginal	SqS1	Original	SqS1	Original	SqS1
				θ=4				
K	304.62	129.37	121.18	49.39	1,665.75	74.41	1,697.28	74.91
T	245.99	229.2	111.54	94.93	169.5	14.63	226.27	19.73
L	−3,343.54	−3,343.38	−3,343.63	−3,343.39	−3,344.44	−3,343.44	−3,344.46	−3,343.54
θ	4.15(0.78)	4.12(0.77)	4.18(0.79)	4.14(0.78)	3.90(0.74)	4.17(0.80)	3.90(0.74)	4.18(0.79)
β_1	−5.17(0.39)	−5.17(0.38)	−5.19(0.38)	−5.18(0.38)	−4.90(0.36)	−5.21(0.39)	−4.89(0.36)	−5.21(0.39)
β_5	−5.20(0.37)	−5.19(0.37)	−5.21(0.37)	−5.20(0.37)	−4.92(0.35)	−5.23(0.37)	−4.91(0.35)	−5.23(0.37)
β_{10}	−5.16(0.38)	−5.15(0.37)	−5.17(0.38)	−5.16(0.38)	−4.88(0.35)	−5.19(0.38)	−4.88(0.35)	−5.20(0.38)
β_{15}	2.06(0.33)	2.06(0.33)	2.06(0.33)	2.06(0.33)	1.95(0.31)	2.07(0.33)	1.95(0.31)	2.07(0.33)
β_{20}	2.06(0.33)	2.06(0.33)	2.07(0.33)	2.06(0.33)	1.95(0.31)	2.08(0.33)	1.95(0.31)	2.08(0.33)
β_{25}	4.16(0.36)	4.15(0.36)	4.17(0.37)	4.16(0.37)	3.93(0.34)	4.18(0.37)	3.93(0.34)	4.19(0.37)
β_{30}	4.13(0.36)	4.13(0.36)	4.14(0.36)	4.14(0.36)	3.91(0.34)	4.16(0.37)	3.91(0.34)	4.16(0.37)
				θ=10				
K	471.97	195.47	87.14	39.43	1,714.20	64.32	1,732.60	67.46
T	375.95	330.6	74.18	66.11	192.17	15.09	215.29	19.04
L	−2,131.95	−2,131.94	−2,132.11	−2,131.99	−2,132.66	−2,132.04	−2,132.67	−2,132.08
θ	10.53(2.31)	10.52(2.31)	10.58(2.32)	10.54(2.30)	10.03(2.20)	10.54(2.29)	10.03(2.20)	10.55(2.28)
β_1	−5.21(0.50)	−5.21(0.50)	−5.21(0.50)	−5.21(0.50)	−4.94(0.47)	−5.20(0.50)	−4.94(0.47)	−5.21(0.50)
β_5	−5.25(0.50)	−5.25(0.50)	−5.25(0.50)	−5.25(0.50)	−4.97(0.47)	−5.24(0.50)	−4.97(0.47)	−5.24(0.50)
β_{10}	−5.25(0.48)	−5.25(0.48)	−5.24(0.48)	−5.25(0.48)	−4.97(0.46)	−5.24(0.48)	−4.97(0.46)	−5.24(0.48)
β_{15}	2.08(0.43)	2.08(0.43)	2.08(0.44)	2.08(0.43)	1.97(0.41)	2.07(0.44)	1.97(0.41)	2.07(0.43)
β_{20}	2.09(0.42)	2.09(0.42)	2.09(0.42)	2.09(0.42)	1.98(0.40)	2.09(0.42)	1.98(0.40)	2.09(0.42)
β_{25}	4.19(0.44)	4.19(0.44)	4.19(0.44)	4.19(0.44)	3.97(0.41)	4.18(0.44)	3.97(0.41)	4.18(0.44)
β_{30}	4.17(0.47)	4.17(0.47)	4.17(0.47)	4.17(0.47)	3.95(0.44)	4.16(0.47)	3.95(0.44)	4.16(0.47)
				θ=16				
K	665.71	263.78	89.44	43.51	1,796.48	64.27	1,812.29	66.29
T	499.04	403.55	74.43	67.16	218.84	13.76	230.78	18.76
L	−1,518.90	−1,518.89	−1,519.05	−1,518.93	−1,519.44	−1,518.96	−1,519.44	−1,518.96
θ	17.34(4.37)	17.33(4.37)	17.42(4.40)	17.32(4.36)	16.53(4.16)	17.20(4.28)	16.53(4.16)	17.20(4.29)
β_1	−5.37(0.59)	−5.37(0.59)	−5.36(0.58)	−5.36(0.59)	−5.08(0.55)	−5.32(0.58)	−5.08(0.55)	−5.31(0.58)
β_5	−5.29(0.61)	−5.29(0.62)	−5.28(0.61)	−5.29(0.61)	−5.01(0.58)	−5.25(0.61)	−5.01(0.58)	−5.24(0.60)
β_{10}	−5.33(0.60)	−5.33(0.60)	−5.32(0.60)	−5.32(0.60)	−5.05(0.56)	−5.28(0.59)	−5.04(0.56)	−5.28(0.59)
β_{15}	2.09(0.51)	2.09(0.51)	2.09(0.51)	2.09(0.51)	1.98(0.48)	2.07(0.50)	1.98(0.48)	2.07(0.50)
β_{20}	2.10(0.53)	2.10(0.53)	2.10(0.53)	2.10(0.53)	1.99(0.50)	2.09(0.52)	1.99(0.50)	2.08(0.52)
β_{25}	4.29(0.58)	4.29(0.58)	4.29(0.58)	4.29(0.58)	4.06(0.54)	4.26(0.57)	4.06(0.54)	4.25(0.58)
β_{30}	4.29(0.57)	4.28(0.57)	4.28(0.57)	4.28(0.57)	4.05(0.53)	4.25(0.56)	4.05(0.53)	4.24(0.56)

tive model errors (MRME) and the average number of correctly and incorrectly identified zero coefficients are summarized in Table 5. We find that the proposed MM2 and MM3 algorithms mesh well with the SCAD and yield good results in simultaneous parameter estimation and variable selection.

6. The Colorectal Cancer Data Analysis

These data come from a prospective cohort study in the Hospital de Bellvitge, a 960 bed public university hospital in the metropolitan area of Barcelona, Spain. Between January of 1996 and December 1998, 403 patients with initial colorectal cancer and surgery were identified and actively followed until June 2002. The event of interest is readmission, a potential recurrent event since colorectal cancer patients can have several readmissions after discharge. Censoring may occur because of death, migration, or change of hospital. The date of surgery was taken as the beginning of the study period. The first readmission time was considered as the time between the date of the surgical procedure and the first

Table 3. Simulation results for the non-regularized setting with $(B, M, q) = (30, 30, 40)$.

	EM		MM1		MM2		MM3	
Par.	Original	SqS1	Original	SqS1	Original	SqS1	Original	SqS1
				θ=4				
K	278.74	118.59	73.92	35.07	2,042.08	47.77	2,074.68	46.28
T	523.06	450.68	106.64	98.5	459.57	17.54	426.42	14.85
L	−2,904.03	−2,904.03	−2,904.11	−2,904.05	−2,905.25	−2,904.94	−2,905.27	−2,904.12
θ	4.15(0.95)	4.15(0.95)	4.15(0.95)	4.15(0.95)	3.92(0.90)	4.12(0.90)	3.92(0.90)	4.12(0.94)
β_1	−5.22(0.41)	−5.22(0.41)	−5.21(0.41)	−5.21(0.41)	−4.91(0.38)	−5.15(0.37)	−4.91(0.38)	−5.17(0.40)
β_5	−5.18(0.40)	−5.18(0.40)	−5.17(0.40)	−5.18(0.40)	−4.88(0.37)	−5.11(0.39)	−4.88(0.37)	−5.14(0.39)
β_{10}	−5.22(0.40)	−5.22(0.40)	−5.21(0.40)	−5.22(0.40)	−4.92(0.37)	−5.15(0.39)	−4.92(0.37)	−5.17(0.39)
β_{15}	2.09(0.36)	2.09(0.36)	2.08(0.36)	2.08(0.36)	1.96(0.34)	2.05(0.36)	1.96(0.34)	2.07(0.35)
β_{20}	2.09(0.36)	2.09(0.36)	2.08(0.36)	2.08(0.36)	1.96(0.34)	2.02(0.35)	1.96(0.34)	2.07(0.36)
β_{25}	4.20(0.39)	4.20(0.39)	4.20(0.39)	4.20(0.39)	3.95(0.36)	4.17(0.37)	3.95(0.36)	4.16(0.38)
β_{30}	4.17(0.39)	4.17(0.39)	4.16(0.39)	4.16(0.39)	3.92(0.37)	4.10(0.37)	3.92(0.37)	4.13(0.39)
β_{35}	−1.04(0.36)	−1.04(0.36)	−1.04(0.36)	−1.04(0.36)	−0.98(0.34)	−1.02(0.39)	−0.98(0.34)	−1.03(0.36)
β_{40}	−1.03(0.37)	−1.03(0.37)	−1.03(0.37)	−1.03(0.37)	−0.97(0.35)	−1.00(0.37)	−0.97(0.34)	−1.02(0.36)
				θ=10				
K	519.23	216.64	113.71	51.1	2,186.26	67.91	2,206.81	68.21
T	896.43	704.39	208.78	191.99	436.46	38.7	509.66	36.64
L	−1,767.95	−1,767.90	−1,767.98	−1,767.93	−1,768.71	−1,767.99	−1,768.72	−1,767.98
θ	10.85(2.95)	10.82(2.91)	10.82(2.91)	10.81(2.91)	10.25(2.75)	10.71(2.87)	10.25(2.75)	10.72(2.87)
β_1	−5.36(0.56)	−5.36(0.56)	−5.35(0.55)	−5.36(0.55)	−5.05(0.51)	−5.30(0.55)	−5.04(0.51)	−5.30(0.55)
β_5	−5.35(0.54)	−5.35(0.55)	−5.34(0.54)	−5.34(0.54)	−5.04(0.51)	−5.29(0.54)	−5.04(0.51)	−5.29(0.54)
β_{10}	−5.34(0.57)	−5.34(0.57)	−5.33(0.57)	−5.34(0.57)	−5.03(0.53)	−5.28(0.56)	−5.03(0.53)	−5.28(0.57)
β_{15}	2.14(0.50)	2.14(0.50)	2.13(0.50)	2.13(0.50)	2.01(0.47)	2.11(0.50)	2.01(0.47)	2.11(0.50)
β_{20}	2.12(0.52)	2.12(0.52)	2.12(0.52)	2.12(0.52)	2.00(0.49)	2.10(0.52)	1.99(0.49)	2.10(0.52)
β_{25}	4.28(0.55)	4.28(0.55)	4.27(0.55)	4.28(0.55)	4.03(0.51)	4.23(0.54)	4.02(0.51)	4.23(0.54)
β_{30}	4.28(0.53)	4.28(0.53)	4.27(0.53)	4.27(0.53)	4.02(0.49)	4.23(0.53)	4.02(0.49)	4.23(0.53)
β_{35}	−1.12(0.50)	−1.12(0.50)	−1.12(0.50)	−1.12(0.50)	−1.06(0.47)	−1.11(0.49)	−1.06(0.47)	−1.11(0.49)
β_{40}	−1.07(0.52)	−1.07(0.52)	−1.07(0.52)	−1.07(0.52)	−1.01(0.49)	−1.06(0.52)	−1.01(0.49)	−1.06(0.52)
				θ=16				
K	699.35	290.43	126.27	57.66	2,374.03	86.71	2,392.47	86.16
T	1120.24	953.63	240.81	206.98	474.54	52.27	558.5	53.06
L	−1,244.49	−1,244.44	−1,244.51	−1,244.46	−1,245.07	−1,244.51	−1,245.07	−1,244.5
θ	17.76(6.29)	17.71(6.28)	17.74(6.32)	17.69(6.26)	16.75(5.89)	17.48(6.15)	16.75(5.89)	17.51(6.16)
β_1	−5.50(0.75)	−5.51(0.75)	−5.49(0.74)	−5.5 (0.74)	−5.17(0.69)	−5.42(0.73)	−5.17(0.69)	−5.43(0.73)
β_5	−5.57(0.76)	−5.57(0.76)	−5.56(0.75)	−5.56(0.76)	−5.23(0.70)	−5.49(0.74)	−5.23(0.70)	−5.49(0.74)
β_{10}	−5.53(0.70)	−5.53(0.70)	−5.52(0.70)	−5.53(0.70)	−5.19(0.65)	−5.45(0.69)	−5.19(0.64)	−5.46(0.69)
β_{15}	2.23(0.70)	2.23(0.70)	2.23(0.70)	2.23(0.70)	2.09(0.65)	2.20(0.69)	2.09(0.65)	2.20(0.69)
β_{20}	2.17(0.66)	2.18(0.66)	2.17(0.66)	2.17(0.66)	2.04(0.62)	2.14(0.65)	2.04(0.62)	2.15(0.65)
β_{25}	4.41(0.69)	4.42(0.69)	4.41(0.69)	4.41(0.69)	4.14(0.63)	4.35(0.68)	4.14(0.63)	4.36(0.68)
β_{30}	4.42(0.72)	4.42(0.72)	4.41(0.72)	4.42(0.72)	4.15(0.66)	4.36(0.71)	4.14(0.66)	4.36(0.71)
β_{35}	−1.08(0.67)	−1.08(0.67)	−1.08(0.67)	−1.08(0.67)	−1.01(0.62)	−1.06(0.66)	−1.01(0.62)	−1.06(0.66)
β_{40}	−1.11(0.61)	−1.11(0.61)	−1.11(0.61)	−1.11(0.61)	−1.04(0.57)	−1.09(0.60)	−1.04(0.57)	−1.10(0.60)

readmission to the hospital related to colorectal cancer. Accordingly, the readmission times are considered as the difference between the last discharge date and the current hospitalization date. We only considered readmissions related to colorectal cancer and the dataset consists of 861 readmission times recorded on 403 patients. As readmission times from the same patient are expected to be highly correlated, we considered the proportional hazards gamma frailty model for analyzing such data. The frailty can be interpreted as the aggregate effect of unmeasured individual-specific covariates, such as genes, diet, living environment and lifestyles, etc. In our regression analysis, the included covariates were sex, type of treatment (chemotherapy versus radiotherapy), tumour stage (Duke's classification: A–B, C, or D), and Charlson's index (0, 1–2, $\geq$3) which measures the risk of readmission for comorbidity.

Table 4. Simulation results for the non-regularized setting with $(B, M, q) = (40, 20, 40)$.

	EM		MM1		MM2		MM3	
Par.	Original	SqS1	Original	SqS1	Original	SqS1	Original	SqS1
				θ=4				
K	238.66	96.54	70.16	30.89	2,105.84	47.83	2,155.95	49.74
T	407.31	347.86	125.6	111.4	368.7	23.61	360.7	25.15
L	−2,557.04	−2,557.04	−2,557.09	−2,557.05	−2,558.17	−2,557.12	−2,558.2	−2,557.12
θ	4.17(0.86)	4.17(0.85)	4.16(0.85)	4.17(0.85)	3.93(0.81)	4.14(0.84)	3.92(0.81)	4.15(0.84)
β_1	−5.22(0.45)	−5.22(0.45)	−5.21(0.45)	−5.22(0.45)	−4.91(0.42)	−5.18(0.45)	−4.9 (0.42)	−5.19(0.44)
β_5	−5.25(0.45)	−5.25(0.45)	−5.24(0.45)	−5.24(0.45)	−4.93(0.42)	−5.21(0.45)	−4.93(0.42)	−5.21(0.45)
β_{10}	−5.26(0.44)	−5.26(0.44)	−5.25(0.44)	−5.26(0.44)	−4.95(0.41)	−5.22(0.43)	−4.94(0.41)	−5.23(0.43)
β_{15}	2.07(0.41)	2.07(0.41)	2.06(0.41)	2.07(0.41)	1.94(0.38)	2.05(0.41)	1.94(0.38)	2.06(0.41)
β_{20}	2.09(0.42)	2.09(0.42)	2.08(0.42)	2.09(0.42)	1.96(0.39)	2.07(0.42)	1.96(0.39)	2.07(0.42)
β_{25}	4.16(0.44)	4.16(0.44)	4.15(0.44)	4.16(0.44)	3.91(0.41)	4.13(0.44)	3.90(0.41)	4.14(0.44)
β_{30}	4.18(0.44)	4.18(0.44)	4.17(0.44)	4.18(0.44)	3.93(0.41)	4.15(0.44)	3.92(0.41)	4.16(0.44)
β_{35}	−1.03(0.42)	−1.03(0.43)	−1.03(0.42)	−1.03(0.42)	−0.97(0.40)	−1.03(0.42)	−0.97(0.40)	−1.03(0.42)
β_{40}	−1.06(0.41)	−1.06(0.41)	−1.05(0.41)	−1.05(0.41)	−0.99(0.39)	−1.05(0.41)	−0.99(0.39)	−1.05(0.41)
				θ=10				
K	455.65	178.28	99.48	45.25	2,276.92	69.04	2,308.73	69.36
T	657.63	480.33	135.89	121.38	361.77	30.68	353.35	31.07
L	−1,542.52	−1,542.52	−1,542.59	−1,542.54	−1,543.28	−1,542.61	−1,543.29	−1,542.61
θ	11.05(2.91)	11.05(2.91)	11.03(2.93)	11.02(2.90)	10.43(2.74)	10.89(2.85)	10.43(2.74)	10.91(2.85)
β_1	−5.46(0.62)	−5.46(0.62)	−5.44(0.62)	−5.45(0.62)	−5.12(0.58)	−5.37(0.61)	−5.12(0.58)	−5.38(0.61)
β_5	−5.41(0.61)	−5.41(0.61)	−5.39(0.61)	−5.40(0.61)	−5.07(0.57)	−5.32(0.59)	−5.07(0.57)	−5.33(0.60)
β_{10}	−5.40(0.66)	−5.40(0.65)	−5.38(0.65)	−5.39(0.65)	−5.06(0.60)	−5.31(0.63)	−5.06(0.60)	−5.32(0.64)
β_{15}	2.14(0.54)	2.14(0.54)	2.14(0.54)	2.14(0.54)	2.01(0.51)	2.11(0.53)	2.00(0.51)	2.11(0.53)
β_{20}	2.19(0.56)	2.19(0.56)	2.18(0.56)	2.19(0.56)	2.05(0.52)	2.15(0.55)	2.05(0.52)	2.16(0.55)
β_{25}	4.32(0.58)	4.32(0.58)	4.30(0.58)	4.31(0.58)	4.05(0.54)	4.25(0.57)	4.04(0.54)	4.25(0.57)
β_{30}	4.32(0.59)	4.32(0.59)	4.30(0.59)	4.31(0.59)	4.04(0.55)	4.25(0.58)	4.04(0.55)	4.25(0.58)
β_{35}	−1.10(0.56)	−1.10(0.56)	−1.10(0.56)	−1.10(0.56)	−1.03(0.53)	−1.08(0.55)	−1.03(0.53)	−1.08(0.55)
β_{40}	−1.08(0.55)	−1.08(0.55)	−1.08(0.55)	−1.08(0.55)	−1.01(0.51)	−1.06(0.54)	−1.01(0.51)	−1.06(0.54)
				θ=16				
K	659.78	251.92	123.95	54.58	2,479.56	90.95	2,507.71	90.23
T	935.35	692.06	162.08	146.8	409.84	40.56	413.09	46.89
L	−1,099.3	−1,099.29	−1,099.34	−1,099.31	−1,099.89	−1,099.38	−1,099.9	−1,099.37
θ	17.70(5.17)	17.69(5.17)	17.66(5.14)	17.65(5.14)	16.66(4.81)	17.34(4.99)	16.65(4.81)	17.37(5.00)
β_1	−5.61(0.78)	−5.61(0.78)	−5.59(0.77)	−5.60(0.78)	−5.25(0.72)	−5.49(0.75)	−5.25(0.72)	−5.50(0.76)
β_5	−5.61(0.77)	−5.62(0.77)	−5.60(0.77)	−5.60(0.77)	−5.25(0.71)	−5.50(0.74)	−5.25(0.71)	−5.50(0.75)
β_{10}	−5.68(0.79)	−5.68(0.79)	−5.66(0.78)	−5.67(0.79)	−5.31(0.72)	−5.56(0.76)	−5.31(0.72)	−5.57(0.76)
β_{15}	2.25(0.71)	2.25(0.71)	2.24(0.71)	2.25(0.71)	2.10(0.65)	2.20(0.69)	2.10(0.65)	2.21(0.69)
β_{20}	2.21(0.68)	2.21(0.68)	2.20(0.68)	2.21(0.68)	2.06(0.63)	2.17(0.66)	2.06(0.63)	2.17(0.67)
β_{25}	4.52(0.76)	4.52(0.76)	4.51(0.76)	4.51(0.76)	4.22(0.70)	4.42(0.73)	4.22(0.70)	4.43(0.74)
β_{30}	4.46(0.78)	4.46(0.78)	4.44(0.78)	4.45(0.78)	4.16(0.72)	4.36(0.76)	4.16(0.72)	4.37(0.76)
β_{35}	−1.13(0.72)	−1.13(0.72)	−1.13(0.72)	−1.13(0.72)	−1.06(0.67)	−1.11(0.71)	−1.06(0.67)	−1.11(0.71)
β_{40}	−1.11(0.71)	−1.11(0.71)	−1.11(0.71)	−1.11(0.71)	−1.04(0.66)	−1.09(0.70)	−1.04(0.66)	−1.09(0.70)

Table 5. Regularized estimation in sparse high-dimensional gamma frailty models.

	MRME	Zeros		MRME	Zeros	
		Correct	Incorrect		Correct	Incorrect
Method		$\rho = 0.25$			$\rho = 0.75$	
			q=30			
SCAD (MM2)	1.1634	26.92	0	1.1999	26.902	0
SCAD (MM3)	1.3720	26.932	0	1.4435	26.95	0
			q=50			
SCAD (MM2)	1.5687	45	0	1.6122	44.992	0
SCAD (MM3)	2.2237	44.994	0	2.8581	44.99	0

As in the simulation studies, the proposed MM algorithms and the EM algorithm gave similar estimates of θ and $\boldsymbol{\beta}$ and hence we only report the estimates based on the MM1 algorithm. For interval estimation, we repeately generated

Table 6. The estimated regression coefficients for the colorectal cancer data.

Covariate	MLE	SE	95% Bootstrap CI†	95% Bootstrap CI‡
Chemotherapy	−0.2136	0.1522	[−0.5079, 0.0887]	[−0.5017, 0.0978]
Female	−0.5260	0.1370	[−0.7953, −0.2583]	[−0.7959, −0.2682]
Tumour stage A–B	−1.0351	0.1834	[−1.4071, −0.6880]	[−1.4137, −0.7041]
Tumour stage C	−0.7396	0.1847	[−1.1169, −0.3931]	[−1.1155, −0.3892]
Index 0	−0.3938	0.1414	[−0.6675, −0.1132]	[−0.6653, −0.1233]
Index 1–2	0.0520	0.3134	[−0.5643, 0.6642]	[−0.5426, 0.6873]

SE, the empirical standard error based on the boostrap samples; CI†, Normal- based Bootstrap CI; CI‡, Bootstrap percentile CI.

bootstrap samples and obtained bootstrap estimates $(\hat{\theta}^*_g, \hat{\beta}^*_g), g = 1, \ldots, G$ with $G = 1{,}000$. We constructed the normal-based bootstrap confidence interval and the bootstrap percentile interval as follows. The normal-based $100(1-\alpha)\%$ bootstrap interval for θ was $(\bar{\theta}^* \pm z_{\alpha/2}\hat{\text{se}}^*(\hat{\theta}))$, where $\bar{\theta}^* = (1/G)\sum_{g=1}^G \hat{\theta}^*_g$ and $\hat{\text{se}}^*(\hat{\theta}) = \sqrt{(1/(G-1))\sum_{g=1}^G(\hat{\theta}^*_g - \bar{\theta}^*)^2}$. The bootstrap $100(1-\alpha)\%$ percentile interval was given by $(\hat{\theta}^*_L, \hat{\theta}^*_U)$, with $\hat{\theta}^*_L$ and $\hat{\theta}^*_U$ the $(\alpha/2)G$th and $(1-\alpha/2)G$th order statistics of $\{\hat{\theta}^*_g\}_{g=1}^G$. The confidence intervals for β and $\Lambda_0(.)$ were calculated similarly. The estimate of θ was $\hat{\theta} = 0.6136$ with a standard error of 0.1422. The 95% normal-based bootstrap CI and 95% bootstrap percentile CI of θ were $[0.3034, 0.8609]$ and $[0.3289, 0.8765]$, respectively. Both of them exclude 0, which suggests a strong dependence between admission times from the same patient. The results for the regression parameters are reported in Table 6. It can be seen that sex and tumour stage are confirmed to have significant effects on readmission time. The risks of patients with Charlson's index 0 are shown to be significantly different from those with Charlson's index $1-2$ or ≥ 3 while the risks for the latter two groups are not found to be different from each other. In Figure 1, we plot the estimated baseline cumulative hazard rate (solid line) along with its pointwise 95% normal-based boostrap confidence band (dotted line) and 95% bootstrap percentile confidence band (dash line).

7. Concluding Remarks

Frailty models have been widely used for the analysis of multivariate failure time data, allowing not only the regression analysis for the times to event(s) but also the modeling of the dependence structure between times to event(s). The estimation and inference procedures are often based on the nonparametric maximum likelihood estimation since the model parameter contains the unknown

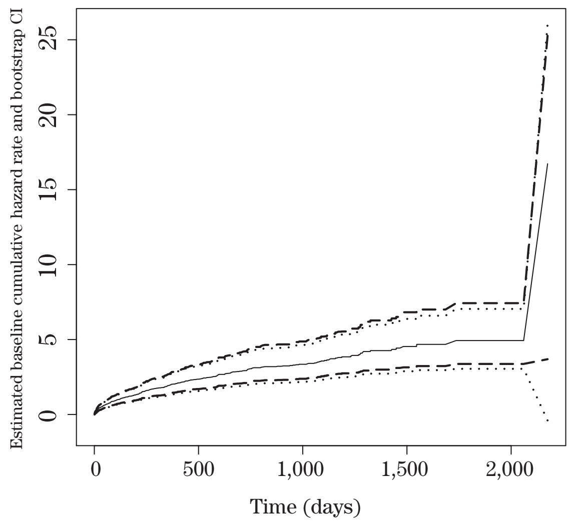

Figure 1. The estimated baseline cumulative hazard rate (solid line), pointwise normal-based bootstrap confidence band (dotted line) and pointwise bootstrap percentile confidence band (dash line).

baseline hazard function, it not can be separated from the other parameters as in univariate failure time modeling. Due to the presence of the high-dimensional nonparametric component, the numerical implementation of the nonparametric likelihood method relies on the EM algorithm and often involves Newton's method and large matrix inversion.

As a viable alternative to the EM algorithm, the MM principle can separate the high-dimensional minorizing function into a sum of univairate function by its construction. This avoids matrix inversion and provides a broader scope for creating more efficient algorithms in statistical optimization problems. For general gamma frailty survival models, we advocate the MM principle and develop a class of profile MM algorithms, shown to exhibit certain theoretical and numerical advantages. Although the MM algorithms are developed for the gamma frailty model, a parallel approach can essentially be developed for the frailty models with a general frailty distribution. We will investigate this in our future work.

Supplementary Materials

The online supplementary material includes the proofs of Theorems 2 and 3.

Acknowledgements

The authors are grateful to the Editor, an associate editor, and referees for many helpful comments. This work has been supported by the University of Hong Kong Seed Fund for Basic Research (201611159026), the University of Hong Kong Seed Fund for Translational and Applied Research (201711160015), the University of Hong Kong, Zhejiang Institute of Research and Innovation Seed Fund, the Hong Kong General Research Fund (17308018) and by the National Natural Science Foundation of China (11771199, 71772153 and 11601524).

References

Andersen, P. K., Klein, J. P., Knudsen, K. M. and Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* **53**, 1475–1484.

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

Chi, E. C. and Lange, K. (2014). A look at the generalized heron problem through the lens of majorization–minimization. *The American Mathematical Monthly* **121**(2), 95–108.

Chi, E. C., Zhou, H. and Lange, K. (2014). Distance majorization and its applications. *Mathematical Programming, Series A* **146**, 409–436.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiologic studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A* (Statistics in Society) **148**, 82–117.

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

Dempster, A. P., Laird, N. M. and Rubin, D. R. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* (Statistical Methodology) **39**, 1–38.

Ding, J. L., Tian, G. L. and Yuen, K. C. (2015). A new MM algorithm for constrained estimation in the proportional hazards model. *Computational Statistics and Data Analysis* **84**, 135–151.

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

Hunter, D. R. (2004). MM algorithms for generalized bradley-terry models. *The Annals of Statistics* **32**, 384–406.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9**, 60–77.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33**, 1617–1642.

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

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 165–174.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* Wiley: Hoboken, New Jersey.

Lee, S. and Huang, J. Z. (2013). A coordinate descent MM algorithm for fast computation of sparse logistic PCA. *Computational Statistics and Data Analysis* **62**, 26–38.

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

Lange, K., Chi, E. C. and Zhou, H. (2014). A brief survey of modern optimization for statisticians (with discussions). *International Statistical Review* **82**, 46-89.

Lange, K. and Zhou, H. (2014). MM algorithms for geometric and signomial programming. *Mathematical Programming, Series A* **143**, 339–356.

Liu, L., Wolfe, R. A. and Huang, X. (2004). Shared frailty model for recurrent events and a terminal event. *Biometrics* **60**, 747–756.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics* **23**, 182–198.

Mkhadri, A., N'Guessan, A. and Hafidi, B. (2010). An MM algorithm for constrained estimation in a road safety measure modeling. *Communication in Statistics—Simulation and Computation* **39**, 1057–2010.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487–493.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**, 183–214.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica* **15**, 831–840.

Varadhan, R. and Roland, C.(2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithms. *Scandinavian Journal of Statistics* **35**, 335–353.

Yen, T. J. (2011). A majorization–minimization approach to variable selection using spike and slab priors. *The Annals of Statistics* **39**, 1748–1775.

Zeng, D. and Lin, D. Y. (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* **65**, 746–752.

Zeng, D., Lin, D. Y. and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Staistica Sinica* **18**, 355–377.

Zeng, D. and Lin, D. Y. (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association* **102**, 167–180.

Zeng, D., Chen, Q. and Ibrahim, J. (2009). Gamma frailty transformation models for multivariate survival times. *Biometrika* **96**, 277–291.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *Journal of the Royal Statistical Society, Series*

B (Statistical Methodology) **69**, 507–64.

Zhou, H., Alexander, D. and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing* **21**, 261–273.

School of Mathematics, Yunnan Normal University, Kunming, Yunnan, P.R. China.

E-mail: u3003396@connect.hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P.R. China.

E-mail: xujf@hku.hk

Department of Mathematics, Southern University of Science and Technology, No. 1088, Xueyuan Road, Shenzhen City, Guangdong Province, P. R. China.

E-mail: tiangl@sustc.edu.cn

(Received November 2016; accepted September 2017)