A Computationally Efficient Approach for Modeling Complex and Big Survival Data

Kevin He, Yanming Li, Qingyi Wei, and Yi Li

Abstract Modern data collection techniques have resulted in an increasing number of big clustered time-to-event data sets, wherein patients are often observed from a large number of healthcare providers. Semiparametric frailty models are a flexible and powerful tool for modeling clustered time-to-event data. In this manuscript, we first provide a computationally efficient approach based on a minimization–maximization algorithm to fit semiparametric frailty models in large-scale settings. We then extend the proposed method to incorporate complex data structures such as time-varying effects, for which many existing methods fail because of lack of computational power. The finite-sample properties and the utility of the proposed method are examined through an extensive simulation study and an analysis of the national kidney transplant data.

1 Introduction

In recent years, advancing technology has resulted in an increasing number of big time-to-event data sets, wherein patients are often observed from multiple clusters (e.g., healthcare providers). For multi-clusters analysis, fixed effects model with clusters as fixed effects is attractive if the sample sizes across clusters are large. However, as is often seen in multi-cluster studies, there are many clusters with relatively few patients. An alternative to a fixed effects approach is the random effects or frailty model, in which clusters-specific effects are treated as random samples from a specific probability distribution.

A wide variety of random effects models have been studied in survival analysis. Among them, the gamma frailty model [1–3] and the log-normal frailty model [4–6] are the most extensively studied approaches for time-to-event data. One reason for

K. He • Y. Li • Y. Li (✉)
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
e-mail: kevinhe@umich.edu; liyanmin@umich.edu; yili@umich.edu

Q. Wei
Duke University School of Medicine and Duke Cancer Institute, Duke University Medical Center, 27710 Durham, NC, USA
e-mail: qingyi.wei@duke.edu

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_10

the popularity of the gamma frailty model is that it has a closed form Laplace transformation for the survival function. Although the log-normal frailty model has no explicit evaluation of the Laplace transform, it allows more flexibility and has been commonly used to fit clustered frailty models [6].

Despite their popularity, the computational complexity of random effects models have limited their use in big data. First, the numerical calculations may have tremendous costs when the dimensionality of predictors is large [7]. Second, when the number of subjects grows, the difficulty of model construction may also increases dramatically. For instance, big time-to-event data are usually complex, e.g., associations between disease outcomes and risk factors may involve complex functional forms such as time-varying effects [8]. In the context of survival analysis, time-varying effects have been studied for application with relatively small sample sizes [9–14]. To estimate such a model, the data set is typically expanded in a repeated measurement format (counting process style), e.g., the time is divided into small time intervals where one single event occurs in each time interval. The covariate values and outcome in the interval for each subject still under observation are stacked into a large data set. Even with a moderate sample size, such an expansion leads to a extremely large data which will be often infeasible to handle with existing computational capability. As an example, data set with 5000 event (assuming no ties) will lead to an expanded data set with records more than 12 millions, which easily out-powers a computer with 8G memory. To avoid the expansion of large-scale data, an alternative approach based on Kronecker product was suggested by Perperoglou et al. [15], with a Newton's method applied by iteratively updating the gradients and Hessian matrices. However, in large-scale survival analyses with massive sample size and large number of predictors, it is computationally expensive to calculate and invert the Hessian matrix. The commonly used Newton-type method may converge slowly or even fail. Finally, numerical problems may arise with skewed covariates (e.g., binary variables with extreme proportion). Extremely small at-risk sets in certain groups may lead to unstable estimations.

To improve the computation efficiency and fill the gap in the existing literature, we first develop an computationally efficient algorithm for estimating the Cox proportional hazards model in the presence of a large number of covariates. The proposed approach combines the strength of the quasi-Newton and minimization–maximization (MM) algorithm. To address the correlation due to clustering, we then extend the proposed algorithm to semiparametric frailty models. Finally, the proposed algorithm is generalized to estimate time-varying effects in complex and big survival data. The proposed method has a connection with coordinate descent which is widely used in high-dimensional data analysis. It should be noted, however, that our general aim is to estimate each predictor's effect instead of variable selection. This is different than a typical constrained optimization approach. In the latter approach, the dimensionality of the data is often much larger than the sample size and the estimated covariate effects are shrunken via penalization.

2 MM Algorithms for Cox Proportional Hazards Model

2.1 The Model

Let D_i denote the time to death and C_i be the censoring time for patient i, $i = 1, \ldots, n$. The observation time is denoted as $T_i = \min\{D_i, C_i\}$, and the death indicator is given by $\delta_i = I(D_i \leq C_i)$. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be a p-dimensional covariate vector for the ith patient. We assume that, conditional on $\mathbf{X}_i$, D_i is independently censored by C_i. To model the death hazard, consider

$$\lambda_i(t|\mathbf{X}_i) = \lim_{dt \to 0} \frac{1}{dt} Pr(t \leq D_i < t + dt | D_i \geq t, \mathbf{X}_i),$$

which we model by $\lambda_i(t|\mathbf{X}_i) = \lambda_0(t) \exp(\mathbf{X}_i^T \boldsymbol{\beta})$, where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of parameters. The corresponding log-partial likelihood is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{X}_i^T \boldsymbol{\beta} - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{X}_\ell^T \boldsymbol{\beta}\right) \right\} \right], \tag{1}$$

where $R_i = \{\ell : T_\ell \geq T_i\}$ is the at-risk set. Let $\nabla l(\boldsymbol{\beta})$ denote the first derivative of the log-partial likelihood with respect to $\boldsymbol{\beta}$. We have

$$\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{X}_i - \frac{\sum_{\ell \in R_i} \mathbf{X}_\ell \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})}{\sum_{\ell \in R_i} \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})} \right\},$$

Let $\nabla^2 l(\boldsymbol{\beta})$ denote the second derivative of the log-partial likelihood with respect to $\boldsymbol{\beta}$. We have

$$-\nabla^2 l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\ell \in R_i} \mathbf{X}_\ell^{\otimes 2} \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})}{\sum_{\ell \in R_i} \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})} - \left\{ \frac{\sum_{\ell \in R_i} \mathbf{X}_\ell \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})}{\sum_{\ell \in R_i} \exp(\mathbf{X}_\ell^T \boldsymbol{\beta})} \right\}^{\otimes 2} \right],$$

where $\otimes$ is the Kronecker product.

2.2 Proposed Method

The proposed method is based on MM algorithm. For some good review on MM methods, the readers are referred to [16–19]. We first consider the Cox proportional hazards model. In a minorization step, we minorize the log-partial likelihood by a surrogate function, which is chosen to separate the parameters. We begin with the

observation that the log-partial likelihood (1) is a concave function of $\boldsymbol{\beta}$. Given the mth step estimate $\hat{\boldsymbol{\beta}}^{(m)}$, an application of Jensen's inequality leads to the following minority surrogate function:

$$l(\boldsymbol{\beta}) \geq \sum_{j=1}^{p}\sum_{i=1}^{n}\alpha_j\delta_i\left[\frac{X_{ij}}{\alpha_j}(\beta_j-\hat{\beta}_j^{(m)})+\mathbf{X}_i^T\hat{\boldsymbol{\beta}}^{(m)}-\log\left\{\sum_{\ell\in R_i}\exp\left(\frac{X_{\ell j}}{\alpha_j}(\beta_j-\hat{\beta}_j^{(m)})\right.\right.\right.$$
$$\left.\left.\left.+\mathbf{X}_\ell^T\hat{\boldsymbol{\beta}}^{(m)}\right)\right\}\right]=g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(m)})=\sum_{j=1}^{p}g(\beta_j|\hat{\boldsymbol{\beta}}^{(m)}), \tag{2}$$

where $g(\beta_j|\hat{\boldsymbol{\beta}}^{(m)})$ is defined implicitly, all $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$ and $\alpha_j > 0$ whenever $X_{ij} \neq 0$. A candidate for α_j is

$$\alpha_j = \frac{\sum_{i=1}^{n}|X_{ij}|}{\sum_{j=1}^{p}\sum_{i=1}^{n}|X_{ij}|}.$$

As we will show in the next paragraph, the choice of α_j is not crucial.

In the maximization step, we maximize (or monotonically increase) the surrogate function to produce the next iteration estimators. For instance, given the mth iteration estimate $\widehat{\boldsymbol{\beta}}^{(m)}$, for $j = 1, \ldots, p$, consider $g(\beta_j|\hat{\boldsymbol{\beta}}^{(m)})$ and update coordinate-wise directions β_j cyclically. Up to a constant, $v > 0$, such a procedure is equivalent to the approach based on coordinate descent; e.g., for $j = 1, \ldots, p$,

$$\hat{\beta}_j^{(m+1)} = \hat{\beta}_j^{(m)} - \alpha_j\{\nabla^2 g(\beta_j|\hat{\boldsymbol{\beta}}^{(m)})\}^{-1}\nabla g(\beta_j|\hat{\boldsymbol{\beta}}^{(m)})), \tag{3}$$

where $\mathbf{X}^T\hat{\boldsymbol{\beta}}^{(m)}$ is treated as an offset. The α_j in (5) and (3) can be considered as part of the step-size control. As long as the ascent property is achieved, the choice of α_j is not crucial.

2.3 Computational Issues

The proposed algorithm maximizes the original log-partial likelihood via the surrogate functions. Simplicity is obtained by separating the variables of optimization problem. That means, we replace the complicated objective functions with a sum of simpler functions, $g(\beta_j|\boldsymbol{\beta}^{(m)})$, each of which depends only on one component of parameter space. The computational speed for optimizing the surrogate functions is linear in p, which is much faster than $O(p^3)$ from inverting the original Hessian matrix. Furthermore, following the argument in Chap. 12 of [18], the ascent property

in the MM algorithm depends only on increasing the surrogate function, not on maximizing it. Therefore, one-step Newton estimators (with step-size control) provide sufficient and rapid updates at each MM step, which further improves the computational efficiency.

To accelerate the convergence of the MM algorithm, we consider a strategy proposed by [18]. Denote the corresponding MM estimation in the $(m+1)$th iteration as $M(\hat{\boldsymbol{\beta}}^{(m)})$ and a composite function $M(M(\cdot))$ by $M \circ M(\cdot)$. Define vector

$$v = M \circ M(\widehat{\boldsymbol{\beta}}^{(m)}) - M(\widehat{\boldsymbol{\beta}}^{(m)})$$

and

$$u = M(\widehat{\boldsymbol{\beta}}^{(m)}) - \widehat{\boldsymbol{\beta}}^{(m)}.$$

Compute the accelerated MM updates as

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = M(\widehat{\boldsymbol{\beta}}^{(m)}) - V(U^T U - U^V)^{-1} U^T \{\widehat{\boldsymbol{\beta}}^{(m)} - M(\widehat{\boldsymbol{\beta}}^{(m)})\}$$

Iterate $\widehat{\boldsymbol{\beta}}$ until converge.

3 MM Algorithms for Penalized Partial Likelihood Estimation of Semiparametric Frailty Model

3.1 The Model

One way to fit the log-normal frailty model is the penalized partial likelihood (PPL) approach developed by McGilchrist and Aisbett [5]. For completeness of exposure, we summarize the algorithm as follows. Let T_{hi} and C_{hi} represent the survival and censoring times, respectively, for the ith patient in the hth cluster. Observation times are denoted by $X_{hi} = T_{hi} \wedge C_{hi}$. The observed death indicators are denoted by $\delta_{hi} = I(T_{hi} \leq C_{hi})$. Let H be the number of clusters, and the total number of subjects be $n = \sum_{h=1}^{H} n_h$, where n_h is the number of subjects in cluster h.

We consider a hazard function

$$\lambda_h(t|\mathbf{X}_i) = \lambda_0(t) \exp(\mathbf{X}_i^T \boldsymbol{\beta} + w_h),$$

where $\mathbf{w} = (w_1, \ldots, w_H)$ is a vector of random effects with independent normal distribution $w_h \sim N(0, \sigma^2)$ for $h = 1, \ldots, H$. Considering the random effects as another set of parameters, the logarithm of the penalized partial likelihood can be written as the sum of the log-partial likelihood and the log of the density of the

random effects

$$l_{\text{ppl}}(\boldsymbol{\beta}, \mathbf{w}, \sigma) = l(\boldsymbol{\beta}, \mathbf{w}) + l_{\text{pen}}(\mathbf{w}, \sigma),$$

where

$$l(\boldsymbol{\beta}, \mathbf{w}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \delta_{hi} \left[\mathbf{X}_{hi}^T \boldsymbol{\beta} + w_h - \log \left\{ \sum_{q\ell \in R_{hi}} \exp\left(\mathbf{X}_{q\ell}^T \boldsymbol{\beta} + w_q \right) \right\} \right], \tag{4}$$

and

$$l_{\text{pen}}(\mathbf{w}, \sigma) = -\frac{1}{2} \sum_{h=1}^{H} \left\{ \frac{w_h^2}{\sigma^2} + \log(2\pi\sigma^2) \right\},$$

where R_{hi} contains all patients still at risk at time T_{hi} regardless the clusters.

The maximization of the penalized partial likelihood includes an inner and an outer loop. The inner loop estimates $\boldsymbol{\beta}$ and $\mathbf{w}$ by a Newton's procedure to maximize $l(\boldsymbol{\beta}, \mathbf{w})$ based on a provisional value of σ (best linear unbiased predictor—BLUP). The outer loop fits the restricted maximum likelihood estimator (REML) for σ^2 based on the BLUPs. Then the procedure is iterated until convergence. Specifically,

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{(m+1)} \\ \widehat{\mathbf{w}}^{(m+1)} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^{(m)} \\ \widehat{\mathbf{w}}^{(m)} \end{bmatrix} - \boldsymbol{\Omega} \begin{bmatrix} \partial l_{\text{ppl}} / \partial \boldsymbol{\beta} \\ \partial l_{\text{ppl}} / \partial \mathbf{w} \end{bmatrix}_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m)},\ \mathbf{w} = \widehat{\mathbf{w}}^{(m)}}$$

where

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}$$

is the inverse of the square $(p + H)$-dimensional Hessian matrix $\mathbf{A}$ with $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \partial^2 l_{\text{ppl}} / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T & \partial^2 l_{\text{ppl}} / \partial \boldsymbol{\beta} \partial \mathbf{w}^T \\ \partial^2 l_{\text{ppl}} / \partial \boldsymbol{\beta} \partial \mathbf{w}^T & \partial^2 l_{\text{ppl}} / \partial \mathbf{w} \partial \mathbf{w}^T \end{bmatrix}$$

More details of this algorithm can be found in Duchateau and Janssen [20].

3.2 Proposed Method

When the number of clusters or the number of covariates is large, it may be computationally expensive to evaluate or invert the square $(p + H)$-dimensional Hessian matrix, prohibiting its application to big data settings. To address this issue,

we extend the MM algorithm to the semiparametric frailty models. Specifically, for a provisional value of σ, we consider the following minority surrogate function:

$$
\begin{aligned}
l_{\text{ppl}}(\boldsymbol{\beta}, \mathbf{w}) \geq & \sum_{j=1}^{p} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \alpha_j \delta_{hi} \left[\frac{X_{hij}}{\alpha_j} (\beta_j - \hat{\beta}_j^{(m)}) + \mathbf{X}_{hi}^T \hat{\boldsymbol{\beta}}^{(m)} + \hat{w}_h^{(m)} \right. \\
& \left. - \log \left\{ \sum_{q\ell \in R_{hi}} \exp \left(\frac{X_{q\ell j}}{\alpha_j} (\beta_j - \hat{\beta}_j^{(m)}) + \mathbf{X}_{q\ell}^T \hat{\boldsymbol{\beta}}^{(m)} + \hat{w}_q^{(m)} \right) \right\} \right] \\
& + \sum_{h=1}^{H} \sum_{i=1}^{n_h} \alpha_h \delta_{hi} \left[\frac{w_h - \hat{w}_h^{(m)}}{\alpha_h} + \mathbf{X}_{hi}^T \hat{\boldsymbol{\beta}}^{(m)} + \hat{w}_h^{(m)} \right. \\
& \left. - \sum_{h=1}^{H} \log \left\{ \sum_{q\ell \in R_{hi}} \exp \left(\frac{Z_{q\ell,h} (w_h - \hat{w}_h^{(m)})}{\alpha_h} + \mathbf{X}_{q\ell}^T \hat{\boldsymbol{\beta}}^{(m)} + \hat{w}_q^{(m)} \right) \right\} \right] \\
& - \frac{1}{2} \sum_{h=1}^{H} \left\{ \frac{w_h^2}{\sigma^2} + \log(2\pi\sigma^2) \right\} \\
= & \, g(\boldsymbol{\beta}, \mathbf{w} | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}) = \sum_{j=1}^{p} g(\boldsymbol{\beta}_j | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}) + \sum_{h=1}^{H} g(w_h | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}),
\end{aligned}
$$

where $Z_{q\ell,h} = 1$ if $q = h$ (i.e., the patient belongs to cluster h) and $Z_{q\ell,h} = 0$ otherwise. In the inner loop, we treat $\hat{w}_h^{(m)}$ as offsets and update coordinate-wise estimate of β_j cyclically: for $j = 1, \ldots, p$

$$
\hat{\boldsymbol{\beta}}_j^{(m+1)} = \hat{\boldsymbol{\beta}}_j^{(m)} - \alpha \left\{ \nabla^2 g(\hat{\boldsymbol{\beta}}_j^{(m)} | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}) \right\}^{-1} \nabla g(\hat{\boldsymbol{\beta}}_j^{(m)} | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}).
$$

Similarly, we treat $\hat{\boldsymbol{\beta}}^{(m)}$ as offsets and update coordinate-wise estimate of w_h cyclically: for $h = 1, \ldots, H$

$$
\hat{w}_h^{(m+1)} = \hat{w}_h^{(m)} - \alpha \left\{ \nabla^2 g(\hat{w}_h^{(m)} | \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}) \right\}^{-1} \nabla g(\hat{w}_h^{(m)} | \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}).
$$

Follows the approach based on [4], an approximated REML estimate for σ^2 is given by

$$
(\hat{\sigma}^2)^{(m+1)} = \frac{\sum_{h=1}^{H} (\hat{w}_h^{(m)})^2}{H - r},
$$

where $r = \alpha \sum_{h=1}^{H} \nabla^2 g(\hat{w}_h^{(m)} | \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{w}}^{(m)}) / (\hat{\sigma}^2)^{(m)}$.

4 MM Algorithm for Semiparametric Frailty Model with Time-Varying Effects

We now extend the MM algorithm to semiparametric frailty models with time-varying effects. Let $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))$ be a p-dimensional vector of potentially time-varying effects. We consider a hazard function

$$\lambda_h(t|\mathbf{X}_i) = \lambda_0(t) \exp(\mathbf{X}_i^T \boldsymbol{\beta}(t) + \mathbf{w}_h).$$

The corresponding log-partial likelihood (4) described in Sect. 3 is replaced by

$$l(\boldsymbol{\beta}, \mathbf{w}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \delta_{hi} \left[\mathbf{X}_{hi}^T \boldsymbol{\beta}(T_{hi}) + w_h - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{X}_{h\ell}^T \boldsymbol{\beta}(T_{hi}) + w_h \right) \right\} \right],$$

To estimate $\boldsymbol{\beta}$, a commonly applied approximation is to span $\boldsymbol{\beta}(\cdot)$ by a set of B-splines on a fixed grid of knots, usually taken to be equally spaced to cover the range of time or equal number of events within each interval. For instance, each $\beta_j(\cdot)$ is an expansion of the form

$$\beta_j(t) = \boldsymbol{\Theta}_j^T \mathbf{B}(t) = \sum_{k=1}^{K} \theta_{jk} B_k(t), \quad j = 1, \ldots, p,$$

where K is the dimension of the basis functions, the $B(t) = (B_1(t), \ldots, B_K(t))^T$ form a basis for a finite-dimensional space, and $\boldsymbol{\Theta}_j = (\theta_{j1}, \ldots, \theta_{jK})$ is a vector of coefficients with θ_{jk} as the corresponding coefficient for the kth component of the jth covariate. Consider parameter vector $\boldsymbol{\theta} = vech(\boldsymbol{\Theta})$, the vectorization of $\boldsymbol{\Theta}$ by row, the log-partial likelihood function is

$$l(\boldsymbol{\theta}, \mathbf{w}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \delta_{hi} \left[\mathbf{X}_{hi}^T \boldsymbol{\Theta} \mathbf{B}(T_{hi}) + w_h - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{X}_{h\ell}^T \boldsymbol{\Theta} \mathbf{B}(T_{hi}) + w_h \right) \right\} \right],$$

We consider the following minority surrogate function:

$$l_{\text{ppl}}(\boldsymbol{\theta}, \mathbf{w}) \geq \sum_{j=1}^{p} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \alpha_j \delta_{hi} \left[\frac{X_{hij}}{\alpha_j} (\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j^{(m)}) \mathbf{B}(T_{hi}) + \mathbf{X}_{hi}^T \hat{\boldsymbol{\Theta}}^{(m)} \mathbf{B}(T_{hi}) + \hat{w}_h \right.$$
$$\left. - \log \left\{ \sum_{q\ell \in R_{hi}} \exp\left(\frac{X_{q\ell j}}{\alpha_j} (\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j^{(m)}) \mathbf{B}(T_{hi}) + \mathbf{X}_{q\ell}^T \hat{\boldsymbol{\Theta}}^{(m)} \mathbf{B}(T_{hi}) + \hat{w}_q^{(m)} \right) \right\} \right]$$

$$+\sum_{h=1}^{H}\sum_{i=1}^{n_h}\alpha_h\delta_{hi}\left[\frac{w_h-\hat{w}_h^{(m)}}{\alpha_h}+\mathbf{X}_{hi}^T\hat{\boldsymbol{\Theta}}^{(m)}\mathbf{B}(T_{hi})+\hat{w}_h^{(m)}\right.$$

$$\left.-\sum_{h=1}^{H}\log\left\{\sum_{q\ell\in R_{hi}}\exp\left(\frac{Z_{q\ell,h}(w_h-\hat{w}_h^{(m)})}{\alpha_h}+\mathbf{X}_{q\ell}^T\hat{\boldsymbol{\Theta}}^{(m)}\mathbf{B}(T_{hi})+\hat{w}_q^{(m)}\right)\right\}\right]$$

$$-\frac{1}{2}\sum_{h=1}^{H}\left\{\frac{w_h^2}{\sigma^2}+\log(2\pi\sigma^2)\right\}$$

$$=g(\boldsymbol{\theta},\mathbf{w}|\hat{\boldsymbol{\theta}}^{(m)},\hat{\boldsymbol{w}}^{(m)})=\sum_{j=1}^{p}g(\boldsymbol{\theta}_j|\hat{\boldsymbol{\theta}}^{(m)},\hat{\boldsymbol{w}}^{(m)})+\sum_{h=1}^{H}g(w_h|\hat{\boldsymbol{\theta}}^{(m)},\hat{\boldsymbol{w}}^{(m)}).$$

The remaining algorithms are the same as those in Sect. 3.

5 Convergence Properties

The numerical convergence of the MM algorithm can be described by the following proposition:

Proposition 1 *Any sequence of iterates* $\boldsymbol{\beta}^{(m+1)}=M(\boldsymbol{\beta}^{(m)})$ *generated by the iteration map* $M(\boldsymbol{\beta})$ *of the MM algorithm possesses a limit, and that limit is the optimal point.*

Proof of Proposition 1 The inequalities

$$l(\boldsymbol{\beta}^{(m+1)})\geq g(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)})\geq g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)})=l(\boldsymbol{\beta}^{(m)})$$

follow from the choice of $\boldsymbol{\beta}^{(m+1)}$ and the minorization condition (5) described in Sect. 2.2. Given the fact that log-partial likelihood function is smooth, if the parameter space is bounded, then all super-level sets $\{\boldsymbol{\beta}: l(\boldsymbol{\beta})\geq c\}$, for a constant c, are compact, and the maximum value of log-partial likelihood is attained (e.g., Weierstrass's theorem). Note that such a bounded assumption is applicable in most practical applications. Apply Proposition 12.4.4 of [18], then Proposition 1 follows.

6 Simulation Study

Finite-sample properties of the proposed method and their alternative were evaluated under three models: Cox proportional hazards model, semiparametric frailty models with time-independent effects or time-varying effects.

6.1 Setting 1: Cox Proportional Hazards Model

Death times were generated from an exponential model, $\lambda(t|\mathbf{X}_i) = 0.5\exp(\mathbf{X}_i^T\boldsymbol{\beta})$ for $i = 1,\ldots,n$. The sample size was $n = 1000$ and the number of covariates was $p = 100$, generated from independent standard normal distributions. The first five variables had coefficients $1, 1, -1, -1, 1$, while the rest had zero coefficients. Censoring times were generated from uniform distributions, with the percentage of censored subjects being approximately 20–30 %. Each data configuration was replicated 100 times. We compared the proposed MM algorithm described in Sect. 2.1. (termed MM), its accelerated modification described in Sect. 2.3 (termed MM2) and a "cocktail" algorithm proposed by Yang and Zou [21]. Specifically, instead of iteratively update $l_j''(\hat{\boldsymbol{\beta}}^{(m)})$ in formula (3) described in Sect. 2.2, the "cocktail" algorithm used an upper bound for the second derivative which is fixed across iteration

$$\Omega_{jj} = \sum_{i=1}^{n} \frac{\delta_i}{4} \left\{ \max_{\ell \in R_i}(X_{\ell j}) - \min_{\ell \in R_i}(X_{\ell j}) \right\}^2 .$$

Table 1 reports average bias (average over $p = 100$ and 100 simulation replications), average mean square error (MSE), empirical coverage probabilities (termed CP) based on 100 bootstraps, median number of iterations until convergence (termed Step), and average computation time (termed Time). Table 1 clearly indicates that the proposed MM algorithms provide better estimation in terms of both convergence speed and estimation accuracy. Moreover, the accelerated modification further reduced the number of iterations.

6.2 Setting 2: Log-Normal Frailty Model

Death times were generated from the log-normal frailty model with constant baseline hazards 0.5 and the random effects were generated from normal distribution with mean 0 and standard deviation 0.4. We considered 100 clusters with sample size within each cluster following a Poisson distribution with rate 50. The covariates were generated from the same distribution as those in Setting 1. We compared the proposed MM algorithm described in Sect. 3, its accelerated version (MM2) and the PPL based on the Newton's procedure (R package *coxme*). Table 2 reports average

Table 1 Setting 1: Cox proportional hazards model

Method	Bias	MSE	CP	Step	Time (s)
Coxtail	0.0410	0.0034	0.974	244.43	38.04
MM	0.0412	0.0028	0.967	18.03	2.69
MM2	0.0413	0.0028	0.967	9.01	2.07

Table 2 Setting 2: log-normal frailty model

Method	Bias of $\hat{\boldsymbol{\beta}}$	MSE of $\hat{\boldsymbol{\beta}}$	CP of $\hat{\boldsymbol{\beta}}$	Bias of $\hat{\sigma}$	MSE of $\hat{\sigma}$	CP of $\hat{\sigma}$	Step	Time (s)
PPL	0.0142	0.0003	0.952	0.0007	0.0008	0.735	NA	49.04
MM	0.0142	0.0003	0.952	0.0009	0.0008	0.735	39.38	42.41
MM2	0.0142	0.0003	0.952	0.0007	0.0008	0.735	21.94	39.89

bias, average mean square error (MSE) and empirical coverage probabilities (termed CP) for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$, median number of iterations until convergence (termed Step), and the average computation time (termed Time).

Note that the asymptotic variance for the estimates of the regression parameters and random effects variance estimate in PPL approach were provided by McGilchrist and Aisbett [5] and McGilchrist [4]. This issue, however, requires further investigation in our settings as the proposed method is an iterative profile likelihood-type of algorithm. A useful tool might be bootstrap. Specifically, the empirical coverage probabilities studied in this subsection were based on a nonparametric bootstrap algorithm proposed by Therneau and Grambsch [22]: (1) choose H clusters by sampling with replacement from the H clusters in the study; (2) let the bootstrap sample be the subjects from the selected clusters; and (3) fit the proposed procedure to this bootstrap sample. This procedure was repeated 100 times. The estimates $\hat{\boldsymbol{\beta}}^{\star}$ and $\hat{\sigma}^{\star}$ were stored for each bootstrap sample. The standard errors of the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ were calculated based on the variability of $\hat{\boldsymbol{\beta}}^{\star}$ and $\hat{\sigma}^{\star}$.

The proposed MM algorithm has comparable performances with the PPL in this setting. For all methods studied, the CPs of $\hat{\boldsymbol{\beta}}$ are closed to the nominal value, 0.95. However, the estimated standard error of the random effects variance estimate underestimates the standard error, and the corresponding CPs are substantially lower than the nominal value of 0.95. This corresponds to the conclusion drawn by Morris [23] for linear mixed models, e.g., the variances of the BLUPs are biased downwards. Due to this bias, bootstrapping BLUP's results in underestimated variation in the data. Further investigation of the properties will be necessary.

6.3 Setting 3: Log-Normal Frailty Model with Time-Varying Effects

The number of clusters and the covariate distribution were the same as those in setting 2. We let β_1 be a time-varying effect such that $\beta_1(t) = 3\sin(3\pi t/4)$. Other covariate effects were the same as previous settings. Ten basis functions were used for implementing B-spline based methods. Each data configuration was replicated 100 times. The average bias of $\hat{\sigma}$ is 0.002 and the median number of iterations until convergence is 33.9. Figure 1 depicts that the proposed MM estimators are sufficiently accurate.

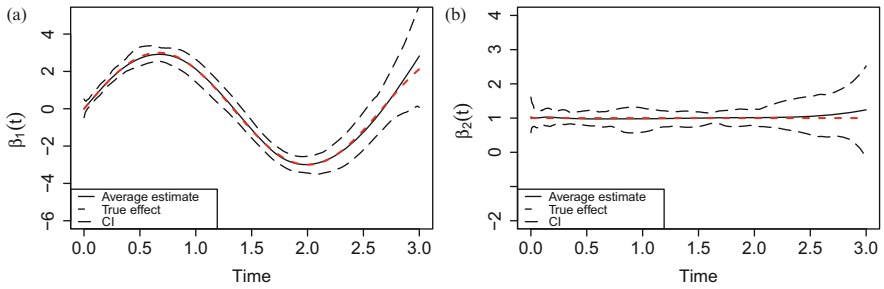

Fig. 1 Estimated coefficients in simulations. (**a**) Time-varying effect. (**b**) Time-independent effect

7 Analysis

The motivating data were obtained from the Organ Procurement and Transplantation Network (OPTN). The United Network for Organ Sharing (UNOS) administers the OPTN under contract with the US Department of Health and Human Services (HHS). The complete data set can be requested from the Organ Procurement and Transplantation Network (https://optn.transplant.hrsa.gov/). Included in the analysis were adult patients ($\geq$ 18 years of age at transplant) who underwent deceased-donor kidney transplantation between January 1990 and December 2008. Adjustment covariates in this study included age, race, gender, donation after cardiac death (DCD), expanded criteria donor (ECD), BMI, dialysis time, indicator of previous kidney transplant, cold ischemic time, and comorbidity conditions (e.g., glomerulonephritis, polycystic kidney disease, diabetes, hypertension). Graft failure was considered to occur when the transplanted kidney ceased to function. Failure time (recorded in years) was defined as the time from transplantation to graft failure or death, whichever occurred first. The final sample size was $n = 146,248$ from 282 transplant centers.

The proposed MM algorithm described in Sect. 4 was employed to investigate the potential time-varying effects. Figure 2 shows a fitted subset of the potential time-varying coefficients with the approximate 95 % point-wise confidence intervals. These results suggested that the effect of diabetes and black race varies over time, resulting in a strengthening of associations with death over time. However, the results for glomerulonephritis, polycystic kidney disease, and hypertension should be interpreted with caution. As shown in Fig. 2, their effects were minimal in the early stage of the follow-up period, but were amplified in the late stage. This may be due to the small at-risk sets at the late stage, resulting in very wide confidence intervals.

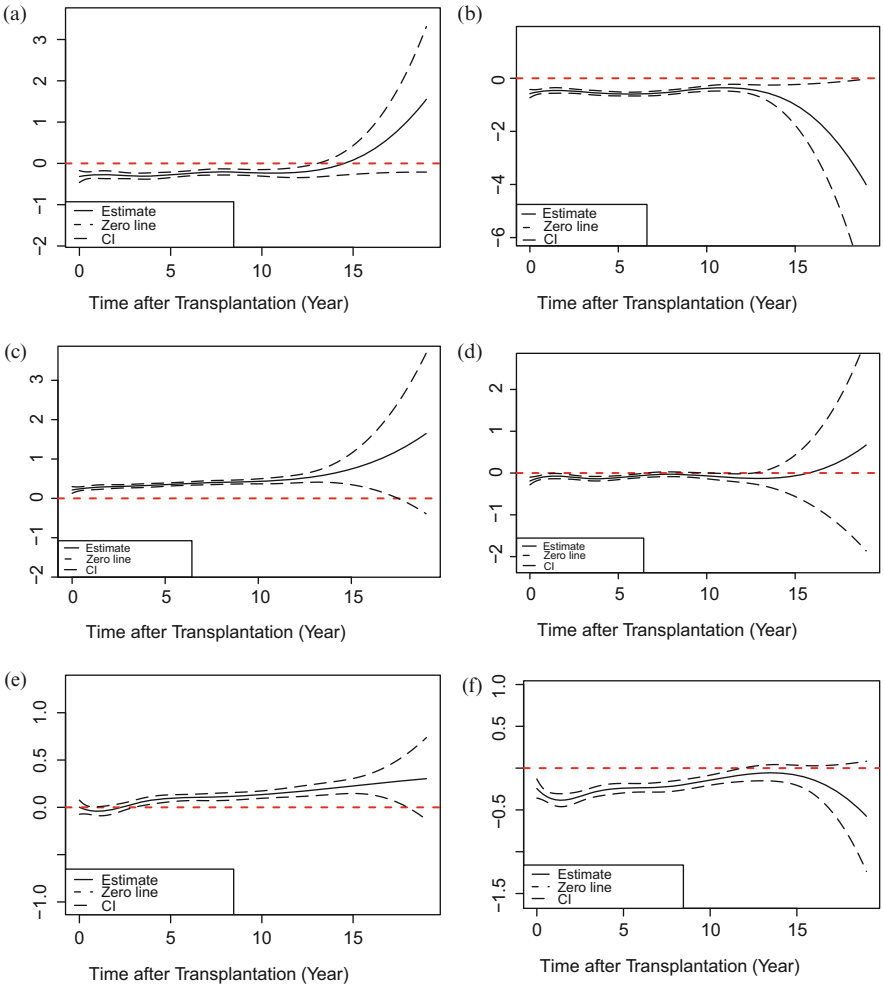

Fig. 2 Real data application: the data were obtained from the Organ Procurement and Transplantation Network (OPTN). (**a**) Glomerulonephritis. (**b**) Polycystic kidney disease (**c**) Diabetes (**d**) Hypertension (**e**) Race: Black (**f**) Race: Hispanic

8 Discussion

Statistical analysis of big clustered time-to-event data presents daunting statistical challenges as well as exciting opportunities. The computation and inversion of the Hessian matrix of the log-partial likelihood is very expensive and may exceed computation memory. To handle problems with large numbers of parameters, we propose a novel algorithm, which combines the strength of quasi-Newton, MM algorithm, and coordinate descent. The proposed algorithm improves upon

the traditional semiparametric frailty models in several aspects. For instance, the proposed algorithms avoid calculation of high-dimensional second derivatives of the log-partial likelihood, and hence, are competitive in term of computation speed and memory usage. Simplicity is obtained by separating the variables of the optimization problem. The proposed methods also provide a useful tool for modeling complex data structures such as time-varying effects.

The overall C index [24] has been routinely used in the medical literature as a natural extension of the ROC curve to survival analysis. A key component in the assessment of model performance is its ability to distinguish subjects who will develop an event from those who will not. In large-scale multi-cluster time-to-event data, a within cluster strategy (e.g., only subjects within each cluster are compared) can greatly reduce the number of calculations. This advantage is especially important for large-scale data exemplified in our study. Risk prediction in time-varying effects model, however, is challenging as it is more complex than evaluating the performance of Cox proportional hazard models.

As suggested by the reviewer, the penalized partial likelihood (PPL) approach is closely connected with the hierarchical likelihood (H-likelihood) method [25, 26]. By treating the frailties as parameters, these approaches avoid integration of unobserved frailties over the frailty distribution. Instead, frailties are jointly estimated with other parameters of interest. This property is particularly appealing when the frailty distribution is not a conjugate prior. However, when the censoring rate is high, parameter estimates may be biased and further bias correction can be helpful [26].

Acknowledgements This work was supported in part by Health Resources and Services Administration contract 234-2005-37011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. Yi Li's research is partly supported by the Chinese Natural Science Foundation (11528102).

References

1. Clayton, D.G.: A model for association in bivariate life table and its application in epidemiological studies of familiar tendency in chronic disease incidence. Biometrika **65**, 141–151 (1978)
2. Clayton, D.G., Cuzick, J.: Multivariate generalization of the proportional hazards model (with discussion). J. R. Stat. Soc. Ser. A **148**, 82–117 (1985)
3. Klein, J.P.: Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics **48**, 795–806 (1992)
4. McGilchrist, C.A.: REML estimation for survival models with frailty. Biometrics **49**, 221–225 (1993)
5. McGilchrist, C.A., Aisbett, C.W.: Regression with frailty in survival analysis. Biometrics **47**, 461–466 (1991)
6. Yamaguchi, T., Ohashi, Y.: Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. Stat. Med. **18**, 1961–1971 (1999)

7. He, K., Kalbfleisch, J.D., Li, Y., Li, Y.J.: Evaluating readmission rates in dialysis facilities with or without adjustment for hospital effects. Lifetime Data Anal. **19**(4), 490–512 (2013)
8. Dekker, F.W., de Mutsert, R., van Dijk, P.C., Zoccali, C., Jager, K.J.: Survival analysis: time-dependent effects and time-varying risk factors. Kidney Int. **74**(8), 994–997 (2008)
9. Zucker, D.M., Karr, A.F.: Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. Ann. Stat. **18**(1), 329–353 (1990)
10. Gray, R.J.: Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. Am. J. Kidney Dis. **87**(420), 942–951 (1992)
11. Gray, R.J.: Spline-based tests in survival analysis. Biometrics **50**(3), 640–652 (1994)
12. Hastie, T., Tibshirani, R.: Varying-coefficient models. J. R. Stat. Soc. Ser. B **55**, 757–796 (1993)
13. Verweij, P.J.M., van Houwelingen, H.C.: Time-dependent effects of fixed covariates in cox regression. Biometrics **51**, 1550–1556 (1995)
14. Berger, U., Schäer, J., Ulm, K.: Dynamic Cox modelling based on fractional polynomials: time-variations in gastric cancer prognosis. Stat. Med. **22**(7), 1163–1180 (2003)
15. Perperoglou, A., le Cessie, S., van Houwelingen, H.C.: A fast routine for fitting Cox models with time varying effects of the covariates. Comput. Methods Prog. Biomed. **25**, 154–161 (2006)
16. Hunter, D.R., Lange, K.: A tutorial on MM algorithms. Am. Stat. **58**, 30–37 (2004)
17. Lange, K., Hunter, D.R., Yang, I.: Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Stat. **9**, 1–20 (2000)
18. Lange, K.: Optimization, 2nd edn. Springer Texts in Statistics. Springer, New York (2012)
19. Wu, T.T., Lange, K.: The MM alternative to EM. Stat. Sci. **29**, 492–505 (2010)
20. Duchateau, L., Janssen, P.: Springer Texts in Statistics. Springer, New York (2008)
21. Yang, Y., Zou, H.: A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. Stat. Interface **6**, 167–173 (2013)
22. Therneau, T.M., Grambsch, P.M.: Modeling Survival Data, Extending the Cox Model. Springer, New York (2000)
23. Morris, J.S.: He BLUPs are not "best" when it comes to bootstrapping. Stat. Probab. Lett. **56**, 425–430 (2002)
24. Pencina, M.J., D'Agostino, R.B.: Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat. Med. **23**(13), 2109–2023 (2004)
25. Lee, Y., Nelder, J.A.: Hierarchical generalized linear models. J. R. Stat. Soc. Ser. B **58**, 619–678 (1996)
26. Jeon, J., Hsu, L., Gorfine, M.: Bias correction in the hierarchical likelihood approach to the analysis of multivariate survival data. Biostatistics **13**(3), 384–97 (2012)