The Annals of Statistics
2016, Vol. 44, No. 3, 1298–1331
DOI: 10.1214/15-AOS1406
© Institute of Mathematical Statistics, 2016

SEMIPARAMETRIC EFFICIENT ESTIMATION FOR SHARED-FRAILTY MODELS WITH DOUBLY-CENSORED CLUSTERED DATA

BY YU-RU SU[1] AND JANE-LING WANG[2]

Fred Hutchinson Cancer Research Center and University of California, Davis

In this paper, we investigate frailty models for clustered survival data that are subject to both left- and right-censoring, termed "doubly-censored data". This model extends current survival literature by broadening the application of frailty models from right-censoring to a more complicated situation with additional left-censoring.

Our approach is motivated by a recent Hepatitis B study where the sample consists of families. We adopt a likelihood approach that aims at the nonparametric maximum likelihood estimators (NPMLE). A new algorithm is proposed, which not only works well for clustered data but also improve over existing algorithm for independent and doubly-censored data, a special case when the frailty variable is a constant equal to one. This special case is well known to be a computational challenge due to the left-censoring feature of the data. The new algorithm not only resolves this challenge but also accommodates the additional frailty variable effectively.

Asymptotic properties of the NPMLE are established along with semiparametric efficiency of the NPMLE for the finite-dimensional parameters. The consistency of Bootstrap estimators for the standard errors of the NPMLE is also discussed. We conducted some simulations to illustrate the numerical performance and robustness of the proposed algorithm, which is also applied to the Hepatitis B data.

1. Introduction.

In the past decades, Cox's proportional hazards model [Cox (1972)], along with its generalizations, have been widely explored and the corresponding asymptotic theories have been well established for independently sampled subjects. When subjects are correlated, for example, under a clustered sampling plan in familial-type studies, the approaches for independent samples are no longer suitable. A common approach to accommodate familial or, more generally, clustered data is the shared-frailty model, which assumes independence for subjects from different clusters but a shared-frailty variable for subjects in the same cluster. Such a frailty model is generally useful to explain the dependency

Received September 2013; revised October 2015.

[1] Supported in part by NIH Grants R21ES022332 and R01AG014358.

[2] Supported in part by NIH Grants 1R01AG025218-01 and 1R56AG043995-01.

MSC2010 subject classifications. Primary 62N02; secondary 62E20.

Key words and phrases. Frailty model, semi-parametric efficiency, EM algorithm, Monte Carlo integrations.

of subjects within the same cluster due to shared genes and environmental background.

For the Cox proportional hazards model with a shared frailty, this leads to a proportional hazards model with a multiplicative frailty term w, which is random and unobservable for subjects within the same cluster, and which explains the dependency among subjects. This class of models was first introduced and termed "frailty" by Vaupel, Manton and Stallard (1979), and subsequently studied for right censored data by Nielsen et al. (1992), Murphy (1994, 1995), and Parner (1998) among others. Due to the latent term in the model, the elegant partial likelihood approach [Cox (1972, 1975)] is no longer applicable. Two alternative approaches have been adopted in the literature, one reverts to the full likelihood approach and the other treats the frailty variables as parameters in the estimation process but imposes a penalty in the partial likelihood to regularize the high dimensional parameters induced by frailties [Therneau, Grambsch and Pankratz (2003)].

The full likelihood approach leads to nonparametric maximum likelihood estimators (NPMLE) when the baseline hazard function is modeled nonparametrically and an expectation-maximization (EM) algorithm is proposed in Nielsen et al. (1992) when the frailty distribution follows a Gamma distribution. The corresponding asymptotic theories, including consistency and asymptotic normality of the NPMLE, were well studied by Murphy (1994, 1995) and Parner (1998) for the cases without and with covariates, respectively. All these approaches adopt the Gamma frailty assumption, mainly due to its computational advantages, as the posterior distribution involved in the E-step of the EM algorithm is also a Gamma distribution. Other frailty distributions, such as log-normal or Weibull distribution, could be employed at additional computational cost, since numerical integration methods, such as Monte Carlo (MC) integration, will be needed to estimate the posterior distributions at each step of the EM-algorithm. Besides using a full likelihood approach, Ripatti and Palmgren (2000) investigated the penalized partial likelihood estimator with a log-normal frailty distribution and Therneau, Grambsch and Pankratz (2003) showed that, with a gamma frailty distribution and a special type of penalty, this leads to the same estimates for the regression parameters as those obtained from an EM algorithm.

All the aforementioned approaches are for right-censored data. Our focus in this paper is to study the estimating procedure and accompanying theory for the shared-frailty model when data are subject to both right- and left-censoring, that is, double-censoring. An example is a familial-type study for Hepatitis B patients, whose age at e-antigen seroconversion is the primary focus of the study. However, due to delayed entry into the study, subjects who have e-seroconverted prior to entry into the study were left-censored (only their age at entry is available), while all other subjects are subject to the usual right censorship that is common in longitudinal follow-up studies. This leads to the double-censorship considered in this paper. We make a note here that the terminology "double-censoring" is confusing by itself, as there are two different definitions in the literature. The first definition is that the survival time of interest can only be observed within a certain window determined by the left- and right-censoring times. Outside this window, the survival time is only known to be either less than the left-censoring time or greater than the right-censoring time. This is the situation considered in this paper, and it has also been considered by Turnbull (1974), Chang and Yang (1987), Chang (1990), Mykland and Ren (1996), Cai and Cheng (2004), Zhang and Jamshidian (2004) and Kim, Kim and Jang (2010). Another definition of double-censoring, as adopted in De Gruttola and Lagakos (1989), Kim, De Gruttola and Lagakos (1993) and Kim (2006), refers to a time period where both endpoints of the time period are subject to either left-, right- or interval-censoring. This second type of double censorship is not considered here as our data conforms to the first type of double-censorship.

Estimating the survival function when there is no covariate has been well investigated in the literature. For instance, Chang and Yang (1987) and Chang (1990) address the consistency and asymptotic normality of the self-consistent estimators of the survival function, while Mykland and Ren (1996) and Zhang and Jamshidian (2004) discuss algorithmic issues for self-consistent estimators and maximum likelihood estimators. The NPMLE under the Cox model had not been explored until Kim, Kim and Jang (2010) established its consistency and asymptotic normality. However, the numerical computation of NPMLE remains a challenge and, to the best of our knowledge, the shared-frailty model for the first type of double censoring has eluded the attention of researchers.

We investigate in this paper the nonparametric maximum likelihood approach and study the asymptotic theory for the NPMLEs. Additionally, a workable numerical algorithm to locate the NPMLEs is proposed along with sufficient conditions to ensure convergence of the algorithm. Our approach works with or without a frailty term and resolves the computational difficulties for doubly-censored data without a frailty term. We would like to make a note here that the proposed numerical method without frailty terms is an independent work of that shown in Kim, Kim and Jang (2013) though the idea of treating the left-censorship as missing data in the EM algorithm is similar to their work. This idea was firstly demonstrated in Su (2011), and further studied in this paper. The model is introduced in Section 2 followed by a computational algorithm presented in Section 3. The left censorship present in the data poses computational challenges in contrast to the right- censoring situation due to the lack of a closed-form solution for the score equation during the M-step of the EM-algorithm. We resolve this difficulty by introducing in Section 3 a modified MCEM algorithm which can be seen as a weighted version of a regular MCEM algorithm. Asymptotic properties and estimations of the standard errors of the proposed NPMLEs are discussed in Section 4. Simulation studies are presented in Section 5 to provide numerical support for the new algorithm and an analysis of the motivating example is provided in Section 6.

2. Model and the NPMLE.

Consider a cluster sampling plan for, for example, familial data, when n independent families are sampled and data are collected for each of the n_i subjects in the ith cluster. The goal is to study the association between the response variable, the survival time T of a subject, and its vector covariates Z. Because the survival times of subjects from the same family may be correlated due to shared gene or environmental background, we assume that a shared-frailty variable W, which could be a vector, explains the dependency of all subjects from the same family/cluster. More specifically, let T_{ij} denote the survival time of the jth subject from the ith cluster with frailty variable W_i and z_{ij} be the observed value of its covariates.

The shared-frailty model assumes that, given the value w_i of the frailty variable and the covariate value z_{ij}, the hazard function for this subject takes the form:

$$\lambda(t|w_i, z_{ij}) = w_i \lambda_0(t) \exp(\beta^T z_{ij}), \tag{1}$$

where λ_0 is the baseline hazard rate and β stands for the regression parameter. Besides the violation of the independence assumption, we face another complication for the motivating Hepatitis B data in that the survival time T of a subject is subject to either left- or right-censoring by L and R, respectively. In the following, we denote $\tilde{T} = \max(L, \min(T, R))$, the observed event-time, and $\delta = I(T \leq R)$ and $\eta = I(T \geq L)$, the right- and left-censored indicators, respectively. By the fact that each subject is only subject to one type of censoring, $\delta + \eta = 1$ or 2 always holds. One thing worth pointing out is that the left-censoring time (e.g., the time at recruitment in the Hepatitis B study) is always observed. This is different from the cases subject to right-censoring only. Consequently, the observed data for a subject is either $(\tilde{T}, \delta, \eta, Z, L)$ for uncensored or right-censored individual, or $(\tilde{T}, \delta, \eta, Z)$ for left-censored one where $\tilde{T}$ is exactly equal to L. The following conditions are for the identifiability of this model and the construction of the likelihood function.

C1. The left- and right-censoring times for the jth subject of the ith cluster, denoted respectively as L_{ij} and R_{ij}, are continuously distributed on $[0, \infty)$ with density functions f_L and f_R, respectively; where L_{ij} is the age at the entry of the study, and the right-censoring time R_{ij} is written as $L_{ij} + Y_{ij}$ with $Y_{ij} \geq 0$, since right-censoring can only occur after a subject enters the study.

C2. Let $\mathbf{Z}_i$ be the $n_i \times q$ covariate matrix for the ith cluster, where the jth row, denoted by Z_{ij}, is the covariate vector of the jth subject in the ith cluster. The probability that $\mathbf{Z}_i^T \mathbf{Z}_i$ is full rank is positive. Moreover, if $c^T Z_{ij} = 0$ with positive probability, it implies $c = 0$. These conditions mean that the covariates are independent within and between subjects.

C3. Conditional on W_i and Z_{ij}, (L_{ij}, Y_{ij}) are independent of T_{ij} and their joint distribution does not involve β or the frailty distribution. This implies that both the left- and right-censoring schemes are noninformative.

C4. The frailty variables, $W_1, \ldots, W_n$, are i.i.d. from a density $f_W(\cdot|\gamma)$ with mean 1 and variance γ. The Laplace transform of f_W, denoted by $M_\gamma(t) =$

$E_\gamma[\exp(-Wt)]$, for any $0 < t < \infty$, satisfies the following conditions: $M_\gamma(0^+) = 1$, $M_\gamma(t) > 0$, $0 < -\frac{\partial}{\partial t} M_\gamma(t) < \infty$, and $0 \le \frac{\partial^2}{\partial t^2} M_\gamma(t) < \infty$ for all γ in a compact set in $\mathbb{R}$. Moreover, the frailty W and the covariate $\mathbf{Z}$ are independent.

Conditions C1 and C3 are standard assumptions for survival data in the presence of left- or right-censoring to facilitate the expression of the likelihood function. Conditions C2 and C4 are needed for the identifiability of the frailty model and integrability of those integrals that involve the frailty distribution.

Let f, S, and F stand for the density, survival, and cumulative density functions respectively of the random variable in the subscript, and Λ_0 is the baseline cumulative hazard function of survival time, that is, $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$. The likelihood contributed by a left-censored subject, given the frailty w and covariates z, is $F_T(\tilde{t}|w, z)$. The likelihood contributed by an uncensored or right-censored subject, given the frailty w and covariates z, is $[f_T(\tilde{t}|w, z)]^{\delta\eta}[S_T(\tilde{t}|w, z)]^{(1-\delta)\eta}$ (see Appendix for the detail). Therefore, the likelihood contributed by the observation O_i from the ith cluster can be expressed as

$$(2)\qquad \begin{aligned} L_i(\beta, \Lambda_0, \gamma|O_i) = \int_0^\infty \prod_{j=1}^{n_i} &\big[f_T(\tilde{t}_{ij}|w, z_{ij})\big]^{\delta_{ij}\eta_{ij}} \big[S_T(\tilde{t}_{ij}|w, z_{ij})\big]^{(1-\delta_{ij})\eta_{ij}} \\ &\times \big[F_T(\tilde{t}_{ij}|w, z_{ij})\big]^{(1-\eta_{ij})} f_W(w|\gamma)\,dw. \end{aligned}$$

In (2), we consider Λ_0 instead of λ_0 as the parameter, because Λ_0 can be estimated at the same parametric rate as β and γ. As is common for models with a nonparametric parameter, the maximum likelihood estimator does not exist due to the infinite-dimensional parameter associated with Λ_0. Therefore, we turn to the nonparametric maximum likelihood approach, which leads to a discrete probability measure with positive point mass assigned to all uncensored observations and an additional set of left-censored observations. This is similar to the left-censored case described in Mykland and Ren (1996), which is for a single population without covariates and the frailty term. The following lemma extends their result (Corollary 5) and provides a description of the NPMLE of Λ_0 under double censoring. The proof is similar to theirs and thus omitted.

LEMMA 2.1. *Denote the ranked observation time points in ascending order by $\tilde{t}_{(l)}$, $l = 1, \ldots, \sum_{i=1}^n n_i$. The NPMLE of the cumulative baseline hazard function, $\Lambda_0(\cdot)$, is a non-decreasing step function with jumps at all uncensored observations and an additional set of observations from left-censored subjects. The left-censored observations that receive positive mass consist of: the smallest observation at time $\tilde{t}_{(1)}$, if it is left-censored, and for $l \ge 2$ all the left-censored observation at time $\tilde{t}_{(l)}$ such that the observation immediately preceding it at time $\tilde{t}_{(l-1)}$ is right-censored. We denote those time points with positive mass in ascending order by $t^1, \ldots, t^K$ with corresponding jump sizes $\lambda^1, \ldots, \lambda^K$.*

Direct maximization of the NPMLE poses computational challenges due to the latent frailty term and the contribution to the likelihood by left-censored data. The latter, which involves the cumulative distribution function F_T in (2), is a more serious problem, as even when the frailty variable is known to be 1, that is, no clustering effect and all survival data are independent, the profile likelihood has no explicit form. Kim, Kim and Jang (2010) proposed to adopt a Gauss–Seidel algorithm to solve the high-dimensional equations, which works well for small sample sizes but often fails to converge when the sample size grows to several hundreds, which is common for a medical or epidemiological study.

For computational stability and because none of the previous approaches accommodate a frailty variable, we took a different and more appealing approach in this paper to treat the left-censored survival times, along with the frailty variable, as missing variables so that a different and more effective EM algorithm can be employed to overcome the aforementioned computational challenges. Following this idea and conditioning first on the frailty variable and then left-censored data, the likelihood (2) from the ith cluster can be written as

$$\begin{aligned}L_i(\beta, \Lambda_0, \gamma | O_i) = \int \Bigg\{ \prod_{j=1}^{n_i} &\big[f_T(\tilde{t}_{ij} | w, z_{ij})\big]^{\delta_{ij}\eta_{ij}} \big[S_T(\tilde{t}_{ij} | w, z_{ij})\big]^{(1-\delta_{ij})\eta_{ij}} \\ &\times \left[\int f_T(u_{ij} | w, z_{ij}) I(u_{ij} \le \tilde{t}_{ij})\, du_{ij}\right]^{(1-\eta_{ij})} \Bigg\} f_W(w|\gamma)\, dw,\end{aligned} \tag{3}$$

since $F_T(\tilde{t}_{ij}|w, z_{ij}) = \int_0^{\tilde{t}_{ij}} f_T(u|w, z_{ij})\, du = \int f_T(u|w, z_{ij}) I(u \le \tilde{t}_{ij})\, du$.

At first glance, the likelihood in (3) involves many integrations because all left-censored survival times are treated as missing data. However, as shown in Propositions 3.1 later, the actual E-step only involves one-dimensional integration over the frailty distribution due to the appealing structure of the proportional hazards model. This leads to a stable EM-algorithm with only one-dimensional Monte Carlo integration in the E-steps and a low dimensional nonlinear maximization in the M-steps.

3. EM-algorithm.

For ease of presentation, we illustrate the EM-algorithm with gamma frailty but other frailty distributions could be employed with additional computational cost in the E-step. The computational advantage of gamma frailty is that the posterior distribution required in the EM algorithm remains a gamma distribution. This feature allows directly sampling from a known gamma distribution and enhances the computational efficiency of Monte Carlo integration. This will be further illustrated in this section. Treating the frailty term and left-censored times as missing data, the integrand of (3) provides the complete likelihood. Let $U_{ij} = T_{ij} I(T_{ij} \le \tilde{T}_{ij})$ denote the unobserved left-censored time, and is zero otherwise. The resulted complete log-likelihood for the ith cluster is

$$
\begin{aligned}
l_i^C(\beta,\Lambda_0,\gamma) = \sum_{j=1}^{n_i} \bigl[& \delta_{ij}\eta_{ij}\bigl(\log w + \log\lambda_0(\tilde{t}_{ij}) + \beta z_{ij}\bigr) - \eta_{ij} w \Lambda_0(\tilde{t}_{ij})\exp(\beta z_{ij}) \\
& + (1-\eta_{ij})\bigl(\log w + \log\lambda_0(u_{ij}) + \beta z_{ij}\bigr) \\
& - (1-\eta_{ij}) w \Lambda_0(u_{ij})\exp(\beta z_{ij})\bigr] \\
& + \log f_W(w).
\end{aligned} \tag{4}
$$

The complete log-likelihood, denoted as l^C, from all clusters is then the sum of (4) over $i = 1,\dots,n$. For simplicity, we will denote the parameters of interest, (β,Λ_0,γ), or equivalently $(\beta,\lambda^1,\dots,\lambda^K,\gamma)$, by θ, and the corresponding parameter space as $\Theta_{\mathrm{EM}} = \Theta_\beta \times \Theta_{(\lambda^1,\dots,\lambda^K)} \times \Theta_\gamma$, in the following illustration.

3.1. *E-step.* The expected complete log-likelihood contributed by the ith cluster with the posterior parameter value θ' is

$$
\begin{aligned}
E_{\theta'}\bigl[l_i^C(\theta)|O_i\bigr] = \sum_{j=1}^{n_i} \bigl[& \delta_{ij}\eta_{ij}\bigl(E_{\theta'}(\log W_i|O_i) + \log\lambda_0(\tilde{t}_{ij}) + \beta z_{ij}\bigr) \\
& - \eta_{ij} E_{\theta'}(W_i|O_i)\Lambda_0(\tilde{t}_{ij})\exp(\beta z_{ij}) \\
& + (1-\eta_{ij})\bigl(E_{\theta'}(\log W_i|O_i) + E_{\theta'}\bigl(\log\lambda_0(U_{ij})|O_i\bigr) + \beta z_{ij}\bigr) \\
& - (1-\eta_{ij}) E_{\theta'}\bigl(W_i\Lambda_0(U_{ij})|O_i\bigr)\exp(\beta z_{ij})\bigr] \\
& + E_{\theta'}\bigl(\log f_W(W_i)|O_i\bigr),
\end{aligned} \tag{5}
$$

which involves imputation of functions of W_i, U_{ij}, or both of them given the observed data. Fortunately, the imputation of functions of W_i, such as W_i, $\log(W_i)$, and $\log f_W(W_i)$ in (5), have simple forms, if one employs the Bayes rule effectively as described below. For illustration purposes, we consider the imputation of a general function $h(W_i)$.

Consider the three sets of variables, O_i^1, O_i^2 and O_i^3, where

$$
\begin{aligned}
O_i^1 &= \bigl\{(\tilde{t}_{ij},\delta_{ij}),\ \text{for non-left-censored subjects}\bigr\}, \\
O_i^2 &= \{\tilde{t}_{ij}, T_{ij} < \tilde{t}_{ij},\ \text{for left-censored subjects}\},
\end{aligned}
$$

and

$$O_i^3 = \{z_{ij}, j = 1,\dots,n_i\}.$$

By Bayes' rule, the imputation of $h(W_i)$ can be expressed as

$$
\begin{aligned}
E_{\theta'}\bigl(h(W_i)|O_i\bigr) &= \int h(w) f_{W|O}(w|O_i)\,dw \\
&= \frac{\int h(w) f(O_i^2|w,O_i^3) f(w|O_i^1,O_i^3)\,dw}{\int f(O_i^2|w,O_i^3) f(w|O_i^1,O_i^3)\,dw},
\end{aligned} \tag{6}
$$

where $f(O_i^2|w, O_i^3) = \prod_{j=1}^{n_i}[1 - \exp\{-w\Lambda_0(\tilde{t}_{ij})\exp(\beta z_{ij})\}]^{(1-\eta_{ij})}$. Under the gamma-frailty model, the posterior distribution of W_i conditioning on the non-left-censored part is a gamma distribution with parameters

$$1/\gamma + \sum_{j=1}^{n_i} \delta_{ij}\eta_{ij}$$

and

$$\gamma \Big/ \left[1 + \gamma \sum_{j=1}^{n_i} \eta_{ij}\Lambda_0(\tilde{t}_{ij})\exp(\beta z_{ij})\right].$$

The imputation of functions involving U_{ij}, say $\log\lambda_0(U_{ij})$ and $W_i\Lambda_0(U_{ij})$ in (5), is more complicated because of the semiparametric setting on T. For a generic function of U_{ij}, say $h(U_{ij})$, the imputation given the observed data O_i is of the form

$$E_{\theta'}\big(h(U_{ij})|O_i\big) = \int\int h(u) f_{U|(W,O)}(u|w, O_{ij})\,du f_{W|O}(w|O_i)\,dw, \tag{7}$$

where O_{ij} is the observed data from the jth subject within the ith cluster, and $f_{W|O}(w|O_i)$ is defined in (6). Since the cumulative hazard function is a non-decreasing step function with positive jumps at $t^1 < \cdots < t^K$, the ordered observed time points mentioned in Lemma 2.1, and corresponding jump sizes $\lambda^1, \ldots, \lambda^K$, the conditional density $f_{U|(W,O)}(u|w, O_{ij})$ in (7) is

$$\frac{\lambda^k w \exp(\beta z_{ij})\exp\{-w\sum_{k'=1}^{k}\lambda^{k'}\exp(\beta z_{ij})\}}{1 - \exp\{-w\sum_{k':\tilde{t}_{ij}\in R^{k'}}\lambda^{k'}\exp(\beta z_{ij})\}},$$

when $u = t^k$, for any $t^k \le \tilde{t}_{ij}$, and 0 otherwise. Because of this explicit form for $f_{U|(W,O)}(u|w, O_{ij})$, no Monte Carlo integration is needed to evaluate the inner integral in (7), so only one-dimensional Monte Carlo integration is needed for (7) and we arrive at the following proposition.

PROPOSITION 3.1. *The imputations of the two functions involving U_{ij} in the imputed complete log-likelihood can be expressed as follows*:

1. $E_{\theta'}[\log\lambda_0(U_{ij})|O_i] = \sum_{k:t^k\le\tilde{t}_{ij}} \log\lambda^k a_{k,ij}(\theta')$,
2. $E_{\theta'}[W\Lambda_0(U_{ij})|O_i] = \sum_{k:t^k\le\tilde{t}_{ij}} \lambda^k c_{k,ij}(\theta')$,

where

$$a_{k,ij}(\theta') = \int f_{U|(W,O)}(t^k|w, O_{ij}, \theta') f_{W|O}(w|O_i, \theta')\,dw,$$

and

$$c_{k,ij}(\theta') = \sum_{k':k'\ge k}^{t^{k'}\le\tilde{t}_{ij}} b_{k',ij}(\theta'),$$

with

$$b_{k',ij}(\theta') = \int f_{U|(W,O)}\bigl(t^{k'}|w, O_{ij}, \theta'\bigr) w f_{W|O}\bigl(w|O_i, \theta'\bigr)\,dw.$$

Note that this proposition implies that only one-dimensional Monte Carlo integrations are involved in the calculation of $a_{k,ij}(\theta')$ and $c_{k,ij}(\theta')$. Let $M_{(l+1)}$ denote the number of Monte Carlo seeds generated in the $(l+1)$th iteration and $\theta_{(l)}$ be the value of θ obtained in the previous iteration. The detailed imputation procedure in the E-step is provided below.

Step 1. Generate $w_{i1}, \ldots, w_{iM_{(l+1)}}$ from a gamma distribution with parameters $1/\gamma_{(l)} + \sum_{j=1}^{n_i} \delta_{ij}\eta_{ij}$ and $\gamma_{(l)}/[1 + \gamma_{(l)} \sum_{j=1}^{n_i} \eta_{ij} \Lambda_{(l)}(\tilde{t}_{ij}) \exp(\beta_{(l)} z_{ij})]$.

Step 2. Evaluate the following terms at w_{im}, for $m = 1, \ldots, M_{(l+1)}$, and plug-in the current value, $\theta_{(l)}$, for θ:

(a) $f(O_i^2|w_{im}, O_i^3)$,

(b) $h(w_{im}) f(O_i^2|w_{im}, O_i^3)$,

(c) $f_{U|(W,O)}(t^k|w_{im}, O_{ij}) f(O_i^2|w_{im}, O_i^3)$, for $t^k \le \tilde{t}_{ij}$,

(d) $f_{U|(W,O)}(t^k|w_{im}, O_{ij}) w_{im} f(O_i^2|w_{im}, O_i^3)$, for $t^k \le \tilde{t}_{ij}$.

Step 3. Take sample means on the four sets of $M_{(l+1)}$ values in (a) to (d) in step 2 and replace the integrals to be evaluated by the sample means in corresponding forms.

The imputed complete log-likelihood can thus be rewritten as

$$\begin{aligned}
&E_{\theta'}\bigl[l^C(\theta)|O\bigr] \\
&\quad = \sum_{i=1}^{n} E_{\theta'}\bigl[l_i^C(\theta)|O_i\bigr] \\
&\quad = \sum_i \biggl\{ \sum_j \delta_{ij}\eta_{ij} E_{\theta'}(\log W_i|O_i) + \sum_j \delta_{ij}\eta_{ij} \log \Lambda\{\tilde{t}_{ij}\} \\
&\qquad + \beta \sum_j \delta_{ij}\eta_{ij} z_{ij} - E_{\theta'}(W_i|O_i) \sum_j \eta_{ij} \sum_{k:t^k \le \tilde{t}_{ij}} \lambda^k \exp(\beta z_{ij}) \\
&\qquad + \sum_j (1-\eta_{ij}) E_{\theta'}(\log W_i|O_i) + \sum_j (1-\eta_{ij}) \sum_{k:t^k \le \tilde{t}_{ij}} \log \lambda^k a_{k,ij}(\theta') \\
&\qquad + \beta \sum_j (1-\eta_{ij}) z_{ij} - \sum_j (1-\eta_{ij}) \sum_{k:t^k \le \tilde{t}_{ij}} \lambda^k c_{k,ij}(\theta') \exp(\beta z_{ij}) \\
&\qquad + E_{\theta'}\bigl(\log f_W(W_i)|O_i\bigr) \biggr\},
\end{aligned}$$

where $\Lambda\{\cdot\}$ represents the jump size of a step function Λ at the specified time point inside the bracket.

3.2. *M-step.* In the M-step, the NPMLE of β,γ and $(\lambda^1, \dots, \lambda^K)$ are located by taking derivatives on (5) and solving a system of equations. The MLE of λ^k is the solution to the following equation:

$$(8)\quad \hat{\lambda}^k = \frac{\delta^k \eta^k + \sum_i \sum_j (1-\eta_{ij}) a_{k,ij}(\theta') I(\tilde{t}_{ij} \in R^k)}{\sum_i \sum_j [\eta_{ij} E_{\theta'}(W|O_i) + (1-\eta_{ij}) c_{k,ij}(\theta')] \exp(\beta z_{ij}) I(\tilde{t}_{ij} \in R^k)},$$

where δ^k and η^k both correspond to the observed time point t^k, and R^k stands for the corresponding risk set defined as $\{k' : t^{k'} \geq t^k\}$. On the other hand, there is no explicit form for the NPMLE of β and v, so a one-step Newton–Raphson method is used to update the estimates in each iteration. The updating formula for β is

$$\beta_{\text{new}} = \beta_{\text{old}} - S_\beta(\beta_{\text{old}})/S'_\beta(\beta_{\text{old}}),$$

where S_β, the score function of β, takes the following form:

$$\begin{aligned} &\sum_i \sum_j [1 - \eta_{ij}(1-\delta_{ij})] z_{ij} \\ (9)\qquad &- \sum_k \Biggl\{ \Biggl[\delta^k \eta^k + \sum_i \sum_j (1-\eta_{ij}) a_{k,ij}(\theta') I(\tilde{t}_{ij} \in R^k) \Biggr] \\ &\times \frac{\sum_i \sum_j [\eta_{ij} E_{\theta'}(W|O_i) + (1-\eta_{ij}) c_{k,ij}(\theta')] z_{ij} \exp(\beta z_{ij}) I(\tilde{t}_{ij} \in R^k)}{\sum_i \sum_j [\eta_{ij} E_{\theta'}(W|O_i) + (1-\eta_{ij}) c_{k,ij}(\theta')] \exp(\beta z_{ij}) I(\tilde{t}_{ij} \in R^k)} \Biggr\}. \end{aligned}$$

One interesting finding from (8) and (9) is that the structure of the resulting forms in the proposed EM algorithm is very similar to those subject to right-censoring. The difference is that, for doubly-censored data every left-censored observation contributes part of its probability mass to each jump points preceding it during each iteration of the EM-algorithm. This redistribution to the left algorithm is similar to the self-consistency property for right-censored data, which redistributes the weight of each right-censored observation to all observations after it. It also reflects the fact that for a left-censored subject, the unobserved event of interest has happened sometime in the past. However, there is a major difference in that some left-censored data also carries positive masses.

Under the assumption of a gamma frailty, the Newton–Raphson algorithm for the parameter γ is based on the following updating rule:

$$\gamma_{\text{new}} = \gamma_{\text{old}} - S_\gamma(\gamma_{\text{old}})/S'_\gamma(\gamma_{\text{old}}),$$

with S_γ, the score function of γ, defined as

$$\frac{n\Gamma'(1/\gamma)}{\Gamma(1/\gamma)} + n \log \gamma - n - \sum_i E_{\theta'}(\log W|O_i) + \sum_i E_{\theta'}(W|O_i).$$

3.3. *Convergence of the algorithm.* Since Monte Carlo errors are induced in the E-step of the MCEM algorithm, the convergence of the proposed algorithm to the true NPMLE is no longer guaranteed. To address this issue, we increase the Monte Carlo sample size $M_{(l+1)}$ with each iteration to enhance the convergence of our algorithm and refer to Chan and Ledolter (1995), Booth and Hobert (1999), Fort and Moulines (2003) and Caffo, Jank and Jones (2005) for the discussions there on how this overcomes the convergence issue. Although the frailty model is formulated under a semi-parametric setting, the problem of locating the NPMLE given an observed sample via the MCEM algorithm is no different from the parametric setting since the jump points $t^1, \dots, t^K$ of the NPMLE of Λ_0 are fixed across iterations. This feature allows us to investigate the convergence issue similar to those in existing parametric literature as long as the stationary points of the observed log likelihood (points where the derivative of the observed log likelihood is zero) are all isolated points and there is no left-censoring involved. See Fort and Moulines (2003) for more details. However, the situation is much more complicated in the presence of left-censoring, because the Monte Carlo approximation in the proposed algorithm is nonstandard. In standard MCEM algorithms the sample mean of the Monte Carlo samples were used but in the our set up a weighted average [as an empirical counterpart to (6)] is employed in step 2 of the EM algorithm to approximate the needed integrals in the likelihood. A new convergence theory is thus needed and we establish this in the proposition below, which provides some sufficient conditions for the convergence of the proposed algorithm in the presence of left-censoring. The proof of Proposition 3.2 is relegated to the supplemental material [Su and Wang (2015)]. In the following context, we denote $L(\theta|O)$ and $l(\theta|O)$ as the observed likelihood and log-likelihood respectively with the cumulative baseline hazard function Λ replaced by a non-decreasing step function. The notation $\mathcal{L} = \{\theta : \frac{dl(\theta|O}{d\theta}) = 0\}$ stands for the set of stationary points of $l(\theta|O)$.

PROPOSITION 3.2. *Under the following conditions* (a)–(e), *the sequence* $\{l(\theta_{(l)}|O)\}$ *of the observed log-likelihood evaluated at* $\{\theta_{(l)}, l = 1, 2, \dots\}$ *converges with probability* 1 *to* $l(\theta^*|O)$, *where* θ^* *is a local maximizer of* $l(\cdot)$, *and* $\{\theta_{(l)}\}$ *converges to* θ^*.

(a) *$f_W(w|\gamma)$ is continuous w.r.t. to w. Moreover, $E_{\theta'}(W_i|O_i)$, $E_{\theta'}(\log W|O)$, $a_{k,ij}(\theta')$, and $c_{k,ij}(\theta')$ are all continuous w.r.t. θ'.*

(b) *$\{\theta \in \Theta_{\mathrm{EM}} : L(\theta) \geq c\}$ is compact for any given constant c and the stationary points of $l(\theta|O)$ are all isolated points in $\mathcal{L}$.*

(c) *The initial value $\theta_{(0)}$ falls in a compact neighborhood $\mathcal{C}^*$ of θ^*, and θ^* is the only point in $\{\theta \in \mathcal{L} : l(\theta|O) = l(\theta^*|O)\}$.*

(d) *For any compact subset $\mathcal{C} \subseteq \Theta_{\mathrm{EM}}$,*

$$\sup_{\theta\in\mathcal{C}} \sup_{t} \sup_{O_{ij}} \mathrm{Var}_\theta\big(h_{(k)}(W, t, O_{ij}) f(O_i^2|W, O_i^3)|O_i^1, O_i^3\big) < \infty, \qquad k = 1, 2, 3, 4,$$

where $h_{(1)}(W, t, O_{ij}) = \log(W)$, $h_{(2)}(W, t, O_{ij}) = W$, $h_{(3)}(W, t, O_{ij}) f_{U|=(W,O)}(t|W, O_{ij})$, *and* $h_{(4)}(W, t, O_{ij}) = f_{U|(W,O)}(t|W, O_{ij})W$.

(e) *The Monte Carlo sample size* $\{M_{(l)}\}$ *satisfies* $\sum_{l=1}^{\infty} M_{(l)}^{-1} < \infty$, *and grows fast enough such that* $l(\theta_{(l)}|O) \geq l(\theta^*|O) - M$ *infinitely often, for some constant* $M > 0$, *and* $\{l(\theta_{(l)}|O)\} \leq l(\theta^*|O)$ *with probability* 1.

Conditions (a)–(c) are also required for standard parametric MCEM [Fort and Moulines (2003)], where condition (a) ensures the continuity of $E_{\theta'}[l_C(\theta|O)]$, the expected complete log-likelihood w.r.t. θ', and condition (c) requires a good initial value which is close to the local maximizer. A sufficient condition for condition (b) is the continuity of the $(K + 3)$th derivative of $l(\theta|O)$, where K varies with the sample size in the semi-parametric setting in contrast to parametric settings where K is fixed. This is the major price for the convergence of the proposed EM algorithm under a semi-parametric model. Condition (d) controls the error induced by the Monte Carlo approximation, and condition (e) specifies the required size of the Monte Carlo samples.

The convergence of $\{l(\theta_{(l)}|O)\}$ suggests a stopping rule based on the difference of the observed likelihood. The algorithm stops when the difference between the observed likelihood at two consecutive iterations is smaller than a pre-specified tolerance of error.

COROLLARY 3.1. *Under the conditions* (a)–(e) *in Proposition* 3.2 *and given a good set of initial values in a neighbor of the NPMLE, the estimators from the proposed MCEM converges a.s. to the NPMLE.*

4. Main theorems. We first list the technical assumptions for the theoretical results of the NPMLE. Hereafter, τ denotes the endpoint time of the study.

A1. The baseline hazard rate function $\lambda_0(t)$ is bounded and positive in $[0, \tau]$. Moreover, the cumulative hazard function is bounded at τ, that is, $\Lambda_0(\tau) < \infty$. Let D_i be the number of right-censored subjects at time τ in the ith cluster. $E(D_i) > 0$.

A2. The expected number of subjects in a family, $E(n_i)$, is bounded above. Also n_i is non-informative to the parameters of interest.

A3. $\Theta_\beta \times \Theta_\gamma$, the parameter space of (β, γ) is compact, and the true value (β_0, γ_0) falls in the interior of the parameter space.

A4. The covariate Z is bounded, that is, there exists $M_Z > 0$, such that $|Z| \leq M_Z$. Moreover, $E_\theta[W \sum_{j=1}^{n_i} \exp(\beta Z_j) I(T_j \geq \tau)]$ exists and is bounded away from 0 over the parameter space.

A5. $E_{\theta_0}[W \exp(\beta_0 Z) I(T \geq t)]$ exists and is bounded away from 0 for all $t \in [0, \tau]$. Moreover, $E_{\theta_0}[W \Lambda_0(\tilde{T}) Z^2 \exp(\beta_0 Z)]$ exists and is greater than 0.

A6. The distribution $f_W(\cdot|\gamma)$ is continuous with respect to γ and has a continuous second derivative with respect to γ. Furthermore, the Fisher information matrix from f_W is of positive definite.

The assumption that $\Lambda_0(\tau) < \infty$ and $E(D_i) > 0$ in A1 are satisfied for a follow up study that needs to end early before all subjects have failed. This is common in medical studies. Assumption A2 is typically satisfied in familial type studies. A3 is a common assumption on the true values of the parameter. Assumptions A4–A5 are technical assumptions for the boundedness of $\hat{\Lambda}_n(\tau)$, the invertibility and the boundedness of the Fisher information operator for the proof of the consistency and asymptotic normality of the proposed estimators. The differentiability of the frailty distribution with respect to γ and the invertibility of its Fisher information are stated in A6.

4.1. *Asymptotic properties of the NPMLE.*

THEOREM 4.1 (Consistency). *Under assumptions* C1–C4 *and* A1–A5, *the NPMLE* $\hat{\theta} = (\hat{\beta}, \hat{\Lambda}, \hat{\gamma})$ *converges strongly to* $\theta_0 = (\beta_0, \Lambda_0, \gamma_0)$ *under the Euclidean norm* $|\cdot|$ *for vector parameters and the supreme norm* $\|\cdot\|$ *for functions on* $[0, \tau]$ *respectively.*

THEOREM 4.2 (Asymptotic normality and efficiency). *Under assumptions* C1–C4 *and* A1–A6, *which imply the consistency* (*for functions on* $[0, \tau]$) *in Theorem* 4.1, *the process* $(\sqrt{n}(\hat{\beta} - \beta), \sqrt{n}(\hat{\Lambda} - \Lambda_0), \sqrt{n}(\hat{\gamma} - \gamma_0))$ *converges in distribution to a normal element G in* $l_\infty(H_p)$, *where* H_p *is a collection of directions as defined at the beginning of Section* A.2, *with mean* 0 *and a covariance structure*

$$\operatorname{cov}(G(h), G(h^*)) = h_1^* \sigma_{\theta_0,1}^{-1}(h) + \int_0^\tau h_2^*(u) \sigma_{\theta_0,2}^{-1}(h)(u)\, d\Lambda_0(u) + h_3^* \sigma_{\theta_0,3}^{-1}(h),$$

$\forall h, h^* \in H_p$, *where* $\sigma_{\theta_0,k}$, $k = 1, 2, 3$, *are the information operators derived in Appendix. Moreover,* $\hat{\beta}$ *and* $\hat{\gamma}$ *are efficient estimators for* β_0 *and* γ_0, *respectively, in the semi-parametric sense.*

We provide the detailed proofs for the two theorems in the Appendix. Basically, the proofs are based on demonstrating the Glivenko–Cantelli property on the terms involved in the NPMLE, and the Donsker property on the score functions.

4.2. *Estimating the standard error of* $\hat{\beta}$. As pointed out in the literature under a semiparametric setting with latent variables, estimation of the standard error of the estimates for finite dimensional parameters involves the inversion of a high-dimensional matrix, where each entity further involves integrals. This often poses computational challenges and is also the case with our setting, where the inverse of the information operator has no explicit form. Thus, even under the right censorship, the straightforward method of utilizing asymptotic variance-covariance matrix, as proposed by Murphy (1995) and Parner (1998), is not applicable to estimate the standard errors. There are two alternative methods in the literature to estimate standard errors under a semiparametric setting: the profile likelihood approach [Murphy, Rossini and van der Vaart (1997)] and the bootstrap method [Tseng, Hsieh and Wang (2005)]. The first approach has also been successfully implemented by Zeng and Cai (2005) in joint modeling right-censored survival data and its longitudinal covariates. Therefore, we explored both approaches in order to compare them. It turns out that the profile likelihood approach in Murphy, Rossini and van der Vaart (1997) and Zeng and Cai (2005) does not work well in our setting, but we are able to modify it and the modified version works well in the simulation study reported in Section 5.

A profile log-likelihood is defined as

$$pl_n(\beta) = \max_{\gamma,\Lambda_0} l(\beta, \Lambda_0, \gamma).$$

The curvature of pl_n around $\hat{\beta}$ provides an estimate for the negative value of the information matrix. However, a direct derivation of the second derivative of pl_n is not feasible since there is no closed form for pl_n due to the integration involved in the likelihood function. Murphy, Rossini and van der Vaart (1997) proposed a second difference method to numerically approximate the information. The second difference method is a numerical approach to approximate the second derivative of a target function pl_n at a point of interest $\hat{\beta}$. We start with the first difference: the basic principle is that if we are interested in estimating the first derivative of pl_n at $\hat{\beta}$, we can use the first difference, $\frac{1}{2h}[pl_n(\hat{\beta}+h) - pl_n(\hat{\beta}-h)]$ for a small h to approximate $pl_n'(\hat{\beta})$. By applying this idea twice the second time on the first differences $pl_n'(\hat{\beta}-h)$ and $pl_n'(\hat{\beta}+h)$, the second difference as defined in Murphy, Rossini and van der Vaart (1997) gives a numerical second differentiation of the target function. However, in the presence of double censoring, their method often results in negative estimates. We were thus motivated to look for an alternative approach to estimate the second derivative of pl_n around $\hat{\beta}$.

The key idea of our approach is, instead of the simple difference method which are very case sensitive, we fit a quadratic curves on pl_n around $\hat{\beta}$ and then take the estimated leading second-order term to estimate the second derivative of pl_n. To implement this method, we evaluate pl_n on d equal-distant points $\beta_1, \dots, \beta_d$ within a window $(\hat{\beta} - h_n, \hat{\beta} + h_n)$, with the half-width h_n taken to be of the order $O(n^{-1/2})$. Although there is no closed form for pl_n, the evaluation can be done by the EM algorithm. A point regarding the calculation of the profile likelihood in our algorithm needs to be addressed as following. Although left-censored data are treated as missing data in the estimation of NPMLE, we use the original form of the likelihood (3) to calculate the profile log-likelihood after obtaining the maximizer $\Lambda(\beta)$ and $\gamma(\beta)$ corresponding to each fixed β. Specifically, we fit a quadratic model $a_0 + a_1\beta + a_2\beta^2$ on the pairs $(\beta_1, pl_n(\beta_1)), \dots, (\beta_d, pl_n(\beta_d))$, the stand error of $\hat{\beta}$ is estimated by $-\hat{a}_2^{-1}$. This method only involves fitting a linear regression model with two predictors, so a moderate number of points $\beta_1, \dots, \beta_d$, say 20, is enough for the implementation.

Although the proposed method needs more computational effort than the method in Murphy, Rossini and van der Vaart (1997), for which pl_n is evaluated at only 3 points, it provides a more stable and accurate estimate for the standard errors. Based on our experience, the performance of both profile likelihood methods depends on the half-width of the window, h_n, and the method by Murphy, Rossini and van der Vaart (1997) is much more sensitive to the choice of h_n. If the window is too narrow, the profile likelihood approach may yield a negative estimate of the standard error due to the highly oscillatory behavior of the profile log-likelihood around $\hat{\beta}$. A wider window may overcome this issue of negative estimate but at a cost of higher biases. For the procedure advocated in Murphy, Rossini and van der Vaart (1997), the bias is always downward and quite serious. Moreover, negative estimates for the standard errors occur much more frequently than our approach based on quadratic approximations. We compare these two profile methods through a simulation study in Section 5, and the simple profile method fails to produce meaningful results.

On the other hand, the bootstrap method has been widely used to estimate the standard error of estimates under many semiparametric models when a simple closed form of the standard error is not available. It provides a numerically valid estimation for the standard error of the estimates by resampling from the observed sample when the number of resampling is fairly large. However, a theoretical justification of the bootstrap method under semiparametric models has not been brought up until Cheng and Huang (2010) and Cheng (2015), which demonstrate the distribution consistency and moment consistency, respectively. Those works provide general theories for us to investigate the consistency of the nonparametric bootstrap standard error under the frailty model subject to double censoring as stated in the following theorem. The proof involves verifying the conditions listed in Theorem 1 in Cheng (2015) and is presented in the Appendix. Below we denote $\hat{\sigma}^*_{\hat{\beta}}$ as the bootstrap sample standard error and $\sigma_{\hat{\beta}}$ as the standard error of $\hat{\beta}$.

THEOREM 4.3 (Consistency of the bootstrap standard error). *Under the assumptions* A1–A6, *the nonparametric bootstrap standard error* $\hat{\sigma}^*_{\hat{\beta}}$ *converges in probability to* $\sigma_{\hat{\beta}}$, *as* $n \to \infty$.

5. Simulation.

5.1. *Evaluate the proposed EM algorithm.* To study the numerical performance of the proposed EM algorithm, four simulation settings were conducted, each based on 100 Monte Carlo samples. For each setting, we consider a binary covariate with equal probability to take the value 0 or 1, and the number of subjects within each family is chosen randomly from $\{2, 3, 4\}$ with equal probabilities, which reflects the structure of the familial data in Section 6, where 49 families participated in the study. The survival times are generated from a Cox model with $\beta = 1$, λ_0 is the hazard function from an exponential distribution with mean 1, and the frailty term is generated from a gamma distribution with mean and variance both equal to 1.

TABLE 1

Simulation results from the proposed EM algorithm. Results on two different simulation settings with two sample sizes, 50 *and* 100, *under each setting. The true baseline hazard is a constant function. The notation* $\hat{\sigma}_{\cdot,\mathrm{MC}}$ *stands for estimated standard deviation from* 100 *Mont Carlo samples. MSE stands for the mean square error of the estimates*

Cases	n	β_0	$\hat{\beta}$	$\hat{\sigma}_{\beta,\mathrm{MC}}$	**MSE**($\hat{\beta}$)	γ_0	$\hat{\gamma}$	$\hat{\sigma}_{\gamma,\mathrm{MC}}$	**MSE**($\hat{\gamma}$)
8% left-censored	50	1	1.0072	0.2173	0.0473	1	0.9532	0.3026	0.0938
	100	1	1.0070	0.1549	0.0240	1	0.9535	0.2293	0.0547
22% left-censored	50	1	1.0200	0.2274	0.0521	1	0.9364	0.3069	0.0982
	100	1	1.0138	0.1559	0.0245	1	0.9580	0.2209	0.0506

The four simulation settings correspond to two cluster sizes, 50 and 100, and the following two types of left censorship: (1) Left-censoring time is generated from an exponential distribution with mean 0.05, and (2) left-censoring time is from an exponential distribution with mean 0.2. In each of the four settings, the right-censoring time is the sum of the left-censoring time and an independent random variable from exponential distribution with mean 8 (cf. Condition C1). For type (1) left censorship above, this resulted in an average of 8% left-censored data and an additional 17% right-censored data, leading to a total of 25% censoring. This reflects a light left-censored case in contrast to the scenario in type (2), where on average 22% of the data are left-censored with an additional 16% right-censored.

The results of the NPMLE for the finite dimensional parameters are listed in Table 1. For the case $n = 50$, the bias for β under light left-censoring is 0.0072 with a standard error of 0.2173. The variance of the gamma-frailty term can be estimated with a bias of 0.0468 and a standard error of 0.3026. Overall, β can be estimated with more precision than γ. Both the biases (and standard errors) for β and γ decreases, to 0.0070 (0.1549) and 0.0465 (0.2293), respectively, as the number of clusters increases to $n = 100$. As expected, the performance of both estimates for β and γ generally deteriorates under the heavier left-censoring scenario (2), but the differences are not large. Considering that a total of 38% of the data are missing under scenario (2), the numerical performance of the procedure seems satisfactory. In addition to the accuracy and precision of the estimator, the proposed EM algorithm also performs well in the aspect of numerical stability. It possesses high convergence rate under all scenarios. In the simulation, we allow the maximum iteration as 100 along with a tolerance of relative error of 0.001. The convergence rates with 50 clusters are 100% and 99% under 8% and 22% of left-censoring, respectively. When the number of clusters increases to 100, the convergence rates achieve 100% under both 8% and 22% of left-censoring.

For estimating the stand error of the estimates, we started by comparing three approaches: the bootstrap method, the profile likelihood method by Murphy, Rossini and van der Vaart (1997) and our version of the profile likelihood method

TABLE 2
Simulation results on estimating the standard error of $\hat{\beta}$. The subscript "BT" stands for estimated standard error based on 50 bootstrap resamples, and the subscript "PL" stands for estimated standard error based on profile likelihood approach with a width of $7/\sqrt{n}$

Cases	n	$\hat{\sigma}_{\beta,\mathrm{MC}}$	$\hat{\sigma}_{\beta,\mathrm{BT}}$	$\hat{\sigma}_{\beta,\mathrm{PL},7}$	$\hat{\sigma}_{\gamma,\mathrm{MC}}$	$\hat{\sigma}_{\gamma,\mathrm{BT}}$	$\hat{\sigma}_{\gamma,\mathrm{PL},7}$
8% left-censored	50	0.2173	0.25442	0.2409	0.3026	0.3019	0.2913
	100	0.1549	0.1703	0.1525	0.2293	0.2053	0.2451
22% left-censored	50	0.2274	0.2561	0.2010	0.3069	0.2871	0.3042
	100	0.1559	0.1769	0.1622	0.2209	0.2087	0.2239

as discussed in Section 4.2. The bootstrap method is similar to the one described in Tseng, Hsieh and Wang (2005). Both versions of the profile likelihood method involve the choice of a window width h, as demonstrated in Section 4.2. Based on our experience in simulations, the performance of the estimated standard errors depends on the choice of the window width and the approach by Murphy, Rossini and van der Vaart (1997) more sensitive to the window width than ours. If the window width is too small, the profile likelihood method may result in unreasonable values of standard error, while a larger width leads to biases. We tried different widths, $h = k/\sqrt{n}$, with $k = 1, 3, 5, 7, 9$ for both profile likelihood approaches but the approach of Murphy, Rossini and van der Vaart (1997) still resulted in many negative estimates up to $h = 7/\sqrt{n}$. Our profile approach resulted in a few negative estimates for small h but none for $h = 7/\sqrt{n}$ and $h = 9/\sqrt{n}$. Naturally, $h = 7/\sqrt{n}$ performed better than $h = 9/\sqrt{n}$. Because of these reasons, we report in Table 2 only our results for $h = 7/\sqrt{n}$ together with the results by the bootstrap method. Both approaches are comparable and produce results close to the Monte Carlo standard deviation, $\hat{\sigma}_{\beta,\mathrm{MC}}$. Since it is difficult to know in reality how to choose the window width, a bootstrap method may be the preferred choice if computational time is not a concern. Otherwise, we recommend our profile likelihood method with a small width h that leads to a positive estimate.

5.2. *Ascent property of the proposed EM algorithm.* One issue commonly encountered in Monte Carlo EM algorithms is the convergence to the true maximizer of the (marginal) likelihood function. As discussed in the literature, maximizing the approximated likelihood by Monte Carlo integration will not locate the MLE exactly due to the presence of Monte Carlo errors. An efficacious EM algorithm should sustain the so-called ascent property which describes the increasing pattern of the targeted marginal likelihood along iterations. Herein, a simulation is conducted to verify the ascent property of the proposed EM algorithm. In order to obtain an analytical form of the marginal likelihood in each iteration, we consider a simple scenario with 100 clusters of size 2. The survival times are generated by a

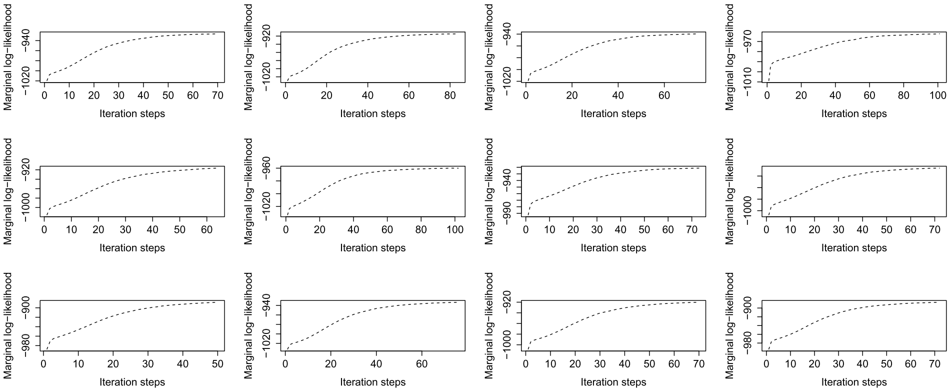

FIG. 1. *Plots of marginal log-likelihood evaluated at the NPMLE calculated in each iteration step based on* 12 *datasets with* 100 *clusters of size* 2.

gamma frailty model with $\beta = 1$, λ_0 as the hazard function of an exponential distribution with mean 1, and $\gamma = 1$. The left-censored rate is about 8% and there is at most 1 left-censored subject within each cluster. The marginal likelihood function is evaluated at the estimated values obtained by maximizing the surrogate likelihood in each iteration. Figure 1 shows the patterns of the marginal log-likelihood along with iteration steps obtained from 12 randomly selected sets of simulations. As observed in the plots, the marginal log-likelihood increase drastically in the first few iterations and continues to climb up till the algorithm converges. The ascent property of the proposed EM algorithm is clearly demonstrated by the trends in the plots.

5.3. *Misspecification on the frailty distribution.* To study the effect of misspecifying the frailty distribution, we conducted some simulations with misspecified frailty distributions. The frailty term is generated from (1) a log-normal distribution with the mean and the standard deviation after logarithm transformation as -0.5 and 1, respectively, and (2) a mixture of two gamma distributions, $\text{Gamma}(2, 0.1)$ and $\text{Gamma}(18, 0.1)$, with equal weights. The two scenarios on frailty distributions represent a unimodal non-gamma distribution with mean 1 and variance about 1.72 and a bimodal distribution with mean 1 and variance about 0.74. We explore the two types of misspecified frailty distributions with the numbers of clusters as 50 and 100, and the settings on other factors similar to the first two simulations in Section 5.1. The left-censoring rate is about 8% with an additional 17% of right censorship in average. Under both scenarios, a gamma frailty model is fitted via the proposed method for estimating the parameters.

The results of the NPMLE obtained from a misspecified model are shown in Table 3. The performance of the estimated regression coefficient $\hat{\beta}$ is comparable to the results under the correct model in Table 1. The biases are slightly greater

TABLE 3

Simulation results with misspecification on the frailty distribution. Results on two scenarios on frailty distributions with 2 *sample sizes*, 50 *and* 100, *under each setting. The true baseline hazard is a constant function. The notation* $\hat{\sigma}_{\cdot,\mathrm{MC}}$ *stands for estimated standard deviation from* 100 *Monte Carlo samples. MSE stands for the mean square error of the estimates*

True frailty distribution	n	β_0	$\hat{\beta}$	$\hat{\sigma}_{\beta,\mathrm{MC}}$	MSE($\hat{\beta}$)	γ_0	$\hat{\gamma}$	$\hat{\sigma}_{\gamma,\mathrm{MC}}$	MSE($\hat{\gamma}$)
Log-normal	50	1	0.9723	0.2371	0.0570	1.72	0.6835	0.2375	0.1566
	100	1	0.9807	0.1697	0.0292	1.72	0.6551	0.1442	0.1398
Mixture of Gammas	50	1	0.9672	0.2460	0.0616	0.74	1.0997	0.3057	0.2228
	100	1	0.9846	0.1593	0.0256	0.74	1.1940	0.2091	0.2498

than that under a correct model, yet still within 4% (0.0277 and 0.0328 based on 50 clusters for unimodal and bimodal models, resp.). Increasing the number of clusters to 100 reduces the bias to less than 2% (0.0193 and 0.0154 for unimodal and bimodal, resp.) and gains efficiency as well. However, the variance component of the frailty cannot be accurately recovered under model misspecification. As demonstrated in Table 3, there can be a non-negligible bias on estimating γ. The relative biases are about 60% and 50% of the true parameter γ_0 for unimodal and bimodal cases, respectively. This is expected under model misspecification as the targets have changed. To summarize, given that the survival regression coefficients are usually the primary interest of a study, the proposed NPMLE is fairly robust against departure of the frailty distribution. In particular, the survival regression coefficient can be estimated with high accuracy and precision even when the frailty component is incorrectly modeled.

6. Numerical example. Our motivating example is a Hepatitis B study for children with chronic Hepatitis B virus (HBV) infection. Hepatitis B is an infectious liver disease causes by HBV. About a quarter of the world populations have been infected. Patients with chronic HBV may infect others over a long period of time and are more likely to develop liver cirrhosis and cancer. It is thus important to control and monitor this disease. HBeAg (Hepatitis B e antigen) is a marker of a patient's degree of infectiousness with positive result indicates the person has high levels of virus and greater infectiousness. E-seroconversion occurs when an infected individual's immune system produces the corresponding antibodies to the e antigen. This is an important therapeutic end point and the primary interest of this study.

Our goal is to understand the seroconversion process of e antigen and its association to two risk factors, one is ALT (Alanine Aminotransferase) measured at the baseline clinical visit and the other is the HBV (Hepatitis B virus) status (yes = 1, and no = 0) of the child's mother. ALT (alanine aminotransferase) is the liver enzyme marker that is followed most closely in those chronically infected with Hepatitis B. An elevated level of ALT indicates the damage on liver cells. Due to the extremely large values of ALT level, a logarithm transformation is often applied and the covariate we used in the analysis is the logarithm of the baseline ALT levels.

The study includes 107 HBeAg positive children from 49 families recruited between 1974 and 1992 and followed up until 2008. Since subjects entered the study at different ages and some of them had completed e-seroconversion before the first clinical visit, the survival time of those patients are thus left- censored. For those who have not developed e-seroconversion at the time of entry into the study, they are subject to the usual right censorship. Thus, this data set is subject to the double censorship considered in this paper. The left-censoring rate is about 2.8% and the right-censoring rate is about 19.4%. Detailed description of the data can be found in Wu et al. (2006), which included a subset of the sample and focused on a different problem. Although the left-censoring rate is low in this data, ignoring the left-censored subjects may result in a bias sample as left-truncated data. To avoid the issue of bias samples, we retain those subjects in the dataset.

Due to the familial structure, a frailty model is employed to accommodate the dependence among subjects from the same family. We consider a shared-frailty model with these two covariates and apply the proposed approach in Sections 2 and 3 to obtain statistical inferences. The results are provided in Table 4. The mother's HBV status, is insignificant but negatively associated with the incidence rate of HB e antigen seroconversion. In the final model, the regression coefficient of logarithm of baseline ALT level is 0.6091 with a p-value 0.0030 indicating a positive and significant effect on the incidence rate of e-seroconversion, which may seem surprising at first but is consistent with clinical observation that patients with higher level of ALT when entering the study tend to have a higher incidence rate to e-seroconvert. This could explained as higher ALT levels are more likely to trigger the development of antibodies to HBV e antigen. The estimated variance of the frailty term is 1.4065 with an estimated standard error of 0.6865. That is, children from the same family tend to have correlated seroconversion time. The estimated cumulative baseline hazard function is shown in Figure 2 along with a 95% pointwise confidence band obtained from bootstrap.

TABLE 4
The fitted results on HB study under full and reduced models

Model	**Parameters**	**Estimates**	**Esti. SD**	***p*-value**
Full	Mom HBV carrier	−0.0311	0.4348	0.9430
	Baseline ALT	0.6191	0.2101	0.0032
	v	0.5303	0.2370	–
Reduced	Baseline ALT	0.6091	0.2053	0.0030
	v	0.4723	0.2350	–

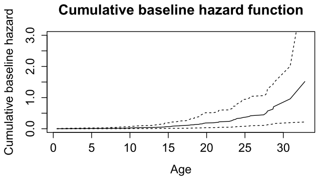

FIG. 2. *The solid line stands for the estimated cumulative baseline hazard function obtained from the proposed method. A pointwise* 95% *confidence band from bootstrap is presented by the dash lines.*

We close this section with a remark on the usage of the baseline ALT. Due to the sampling plan, the ALT measurements taken from the left-censored subjects at their first clinical visits are post-seroconversion; hence, it can be an issue that the significant results are likely contributed by a reverse causation. The validity of using the baseline ALT obtained from the left-censored subjects can be justified by the following two reasons: (1) the left censoring proportion is very low (only 3 out of 107 subjects) for this data, so unlikely to induce a serious bias, and (2) ALT levels tend to stabilize to a normal level after seroconversion, so the ALT level at entry of the study for a left-censored data is likely to be lower than the ALT levels prior to seroconversion. The implication is that the actual p-value should be smaller than the ones reported in Table 4 leading to an even more significant finding. Thus, the significance finding observed in this paper is not a result of the reverse causation. To provide further assurance, a separate analysis is conducted on the same data but omitting the three left-censored subjects. This resulted in a left-truncation (left-censored data are truncated) and right-censoring (LTRC) scenario. The new algorithm we developed for LTRC clustered survival data resulted in an estimate of 0.7238 (S.E. = 0.2570) for the regression coefficient of ALT (p-value = 0.0049). Thus, the new analysis underscores the significant association between the baseline ALT level and seroconversion time.

7. Conclusions. In this paper, we propose a likelihood approach to estimate the unknown components in a shared-frailty model for clustered survival data that are subject to double-censoring. We show that the nonparametric maximum likelihood method leads to $\sqrt{n}$-consistent and semiparametrically efficient estimator for all finite-dimensional parameter and $\sqrt{n}$-consistent for the cumulative baseline hazard function.

Two estimates for the standard deviation of the NPMLE for finite-dimensional parameters are investigated, one based on the bootstrap method and the other based on a new quadratic approximation for the profile likelihood. Both approaches are supported by numerical evidence and lead to reliable and stable estimates. They complement each other in that the bootstrap method is conceptually simpler but computationally costly. The quadratic approximation method is computationally efficient and a remedy for the simple profile likelihood approach proposed in

Murphy, Rossini and van der Vaart (1997), which often leads to negative estimates of the standard errors when data are doubly censored.

In addition to theoretical contributions, a new and effective algorithm is proposed to estimate the nonparametric maximum likelihood estimates through a modified EM algorithm by treating the unobserved frailty terms and all left-censored survival times as missing data. The distinctive features of the proposed algorithm are: (i) it provides a computationally simple and stable algorithm that involves only one-dimensional Monte Carlo integrations, with respect to the latent frailty, in the E-step of the EM-algorithm, (ii) it involves simple and tractable maximization in the M-step of the EM-algorithm and (iii) for a special and simpler case where the frailty variable is constant, it involves no Monte Carlo integration and overcomes the computational instability of an existing method [Kim, Kim and Jang (2010)] that tackles the full nonparametric likelihood by the Gauss–Seidel method, which involves solving high-dimensional equation systems. Thus, we not only provide a viable solution to a new problem but also resolve a lingering computational issue for independent left- or doubly-censored data.

APPENDIX

A.1. Construction of the likelihood contributed by uncensored and right-censored subjects. We focus on uncensored subjects. An analogous argument can be extended to right-censored subjects. For an uncensored subject, the observed data are $(\tilde{T} = \tilde{t}, \delta = 1, \eta = 1, Z = z, L = l)$, where $\tilde{t} > l$. Under the assumption of independence between W and Z, the conditional density f_{obs} of the observed data given $W = w$ is

$$\begin{aligned} f_{\mathrm{obs}}(\tilde{t}, 1, 1, z, l|w) &= f_{(\tilde{T},\delta,\eta,L)}(\tilde{t}, 1, 1, l|z, w) f_Z(z) \\ &= f_{(T,\delta,L)}(\tilde{t}, 1, l|z, w) f_Z(z). \end{aligned} \tag{10}$$

The last equation holds since whenever $\tilde{t} > l$ it implies that $\eta = 1$. By substituting $\delta = 1$ with $R \geq T$,

$$\begin{aligned} f_{(T,\delta,L)}(\tilde{t}, 1, l|z, w) f_Z(z) &= P(T = \tilde{t}, L = l, R \geq T|z, w) f_Z(z) \\ &= P(T = \tilde{t}, L = l, L + Y \geq \tilde{t}|z, w) f_Z(z). \end{aligned} \tag{11}$$

By the conditional independence in C3 between T and (L, Y) given (Z, W) the right-hand side of (11) becomes

$$f_T(\tilde{t}|z, w) P(L = l, L + Y \geq \tilde{t}|z, w) f_Z(z). \tag{12}$$

The noninformative assumption on L and Y in C3 implies that the second term above does not involve any parameter of interest, hence the observed left-censoring time does not contribute information to the likelihood.

A.2. Proof of Theorem 4.1.

OUTLINE OF THE PROOF. We shall use a subscript n, the number of families, for the NPMLE since the asymptotic properties are constructed according to n. We would like to point out here that the NPMLE exists with probability 1 under our setting. This can be verified by an apagogic argument analogous to pages 2140–2141 in Zeng and Cai (2005). Consistency of the NPMLE can be demonstrated by first showing that $\hat{\Lambda}_n(\tau)$ is bounded almost surely as $n \to \infty$. This implies that $\hat{\Lambda}_n$ can be regarded as a bounded measure. Then by Helly's selection theorem and the compactness of the parameter space $\Theta_\beta \times \Theta_\gamma$, every subsequence of $\hat{\theta}_n = (\hat{\Lambda}_n, \beta_n, \gamma_n)$ has a subsequence $\{q(n)\}$ of $\{n\}$ such that $\hat{\theta}_{q(n)} = (\beta_{q(n)}, \Lambda_{q(n)}, \gamma_{q(n)})$ converges to a certain inner point $\theta^* = (\beta^*, \Lambda^*, \gamma^*)$, where Λ^* is continuous as shown in Zeng and Cai (2005), and $\hat{\Lambda}_{q(n)}$ converges uniformly to it in the whole parameter space. The proof will be completed if we can show that $\theta^* = \theta_0$. However, we do not know what θ^* is, since there is no close form solution for $\hat{\theta}_n$. Therefore, we rely on an intermediate function $\bar{\Lambda}_n(\cdot)$, which converges to Λ_0 uniformly on $[0, \tau]$. The claim that $\theta^* = \theta_0$ can next be established similar to the arguments in the literature [Dupuy, Grama and Mesbah (2006), Murphy (1994)]. Below, we provide details of the proof.

To prove the boundedness of $\hat{\Lambda}_n(\tau)$, we take derivatives on the observed log likelihood function with respect to all λ^k's, and it can be shown that

$$
\begin{aligned}
&\hat{\Lambda}_n(\tau)\\
&\quad= \sum_{k=1}^{K} \frac{[\delta^k \eta^k + \sum_{i=1}^{n} \sum_{j=1}^{n_i} (1-\eta_{ij}) a_{k,ij}(\hat{\theta}) I(\tilde{t}_{ij} \in R^k)] I(\tilde{t}^k \le \tau)}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} [\eta_{ij} E_{\theta'}(W|O_i) + (1-\eta_{ij}) c_{k,ij}(\hat{\theta})] \exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tilde{t}^k)}\\
&\quad\le \sum_{k=1}^{K} \frac{[\delta^k \eta^k + \sum_{i=1}^{n} \sum_{j=1}^{n_i} (1-\eta_{ij}) a_{k,ij}(\hat{\theta}) I(\tilde{t}_{ij} \in R^k)] I(\tilde{t}^k \le \tau)}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \eta_{ij} E_{\theta'}(W|O_i) \exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tilde{t}^k)}\\
&\quad\le \sum_{k=1}^{K} \frac{[\delta^k \eta^k + \sum_{i=1}^{n} \sum_{j=1}^{n_i} (1-\eta_{ij}) a_{k,ij}(\hat{\theta}) I(\tilde{t}_{ij} \in R^k)] I(\tilde{t}^k \le \tau)}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \eta_{ij} E_{\theta'}(W|O_i) \exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tau)}\\
&\quad= \sum_{k=1}^{K} \delta^k \eta^k I(\tilde{t}^k \le \tau) + \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (1-\eta_{ij}) a_{k,ij}(\hat{\theta}) I(\tilde{t} \ge \tilde{t}^k) I(\tilde{t}^k \le \tau)\\
&\qquad \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n_i} \eta_{ij} E_{\theta'}(W|O_i) \exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tau).
\end{aligned} \tag{13}
$$

The second term in the numerator of (13) is bounded above by $\sum_{i=1}^{n} \sum_{j=1}^{n_i} (1 - \eta_{ij})$, since the sum of $a_{k,ij}(\hat{\theta})$ over all k such that $\tilde{t}^k \le \tilde{t}_{ij}$ is bounded above by 1.

Then

$$\begin{aligned}\hat{\Lambda}_n(\tau) &\le \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij}\eta_{ij} I(\tilde{t}_{ij} \le \tau) + \sum_{i=1}^n \sum_{j=1}^{n_i} (1-\eta_{ij})}{\sum_{i=1}^n \sum_{j=1}^{n_i} \eta_{ij} E_{\theta'}(W|O_i)\exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tau)} \\ &\le \frac{1/n \sum_{i=1}^n 2n_i}{1/n \sum_{i=1}^n \sum_{j=1}^{n_i} \eta_{ij} E_{\theta'}(W|O_i)\exp(\hat{\beta} z_{ij}) I(\tilde{t}_{ij} \ge \tau)}.\end{aligned} \tag{14}$$

The upper bound in (14) converge a.s. to a finite number as n tends to infinite by the Law of Large Numbers and assumptions A1, A2 and A4. This implies the boundedness of $\hat{\Lambda}_n(\tau)$. Consequently, the NPMLE $\hat{\Lambda}_n$ is a finite measure on $[0, \tau]$. According to Helly's selection lemma, every subsequence of $\hat{\Lambda}_n$ has a further subsequence $\hat{\Lambda}_{q(n)}$ such that $\|\hat{\Lambda}_{q(n)} - \Lambda^*\|$ converges to 0 with probability 1 on $[0, \tau]$.

Before defining an intermediate term used in this proof, we rewrite $\hat{\Lambda}_n$ as

$$\hat{\Lambda}_n(t^*) = \frac{1}{n}\sum_{r=1}^n \sum_{s=1}^{n_r} \frac{\delta_{rs}\eta_{rs} I(\tilde{t}_{rs} \le t^*) + nP_n[Q_1(t, O, \hat{\theta})]|_{t=\tilde{t}_{rs}} I(\tilde{t}_{rs} \le t^*)}{P_n[Q_2(t, O, \hat{\theta})]|_{t=\tilde{t}_{rs}}},$$

where

$$Q_1(t, O_i, \theta) = \sum_{j=1}^{n_i} (1-\eta_{ij}) a_{t,ij}(\theta) I(\tilde{t}_{ij} \ge t),$$

$$Q_2(t, O_i, \theta) = \sum_{j=1}^{n_i} \bigl[\eta_{ij} E_\theta(W|O_i) + (1-\eta_{ij}) c_{t,ij}(\theta)\bigr] \exp(\beta z_{ij}) I(\tilde{t}_{ij} \ge t),$$

and

$$P_n[f(O)] = \frac{1}{n}\sum_{i=1}^n f(O_i)$$

stands for the empirical process. Then the intermediate term $\bar{\Lambda}$ is defined as

$$\begin{aligned}\bar{\Lambda}_n(t^*) &= \frac{1}{n}\sum_{r=1}^n \sum_{s=1}^{n_r} \frac{\delta_{rs}\eta_{rs} I(\tilde{t}_{rs} \le t^*) + nP_n[Q_1(t, O, \theta_0)]|_{t=\tilde{t}_{rs}} I(\tilde{t}_{rs} \le t^*)}{P_n[Q_2(t, O, \theta_0)]|_{t=\tilde{t}_{rs}}} \\ &= \frac{1}{n}\sum_{r=1}^n \sum_{s=1}^{n_r} \frac{\delta_{rs}\eta_{rs} I(\tilde{t}_{rs} \le t^*)}{P_n[Q_2(t, O, \theta_0)]|_{t=\tilde{t}_{rs}}} \\ &\quad + \frac{1}{n}\sum_{r=1}^n \sum_{s=1}^{n_r} \frac{nP_n[Q_1(t, O, \theta_0)]|_{t=\tilde{t}_{rs}} I(\tilde{t}_{rs} \le t^*)}{P_n[Q_2(t, O, \theta_0)]|_{t=\tilde{t}_{rs}}}.\end{aligned}$$

Since the class $\{Q_2(\cdot, O, \theta_0)$ on $[0, \tau]\}$ can be shown to be Glivenko–Cantelli by establishing the uniform boundedness and bounded variation of Q_2, assumption A5 then implies the convergence of the first term to $E_{\theta_0}[\sum_{j=1}^{n_i} \delta_{ij}\eta_{ij} \times$

$\frac{I(T_{ij} \le t^*)}{P[Q_2(t,O,\theta_0)]|t=T}]$ as n goes to infinite. Likewise, the pointwise convergence of the second term to $E_{\theta_0}[\sum_{j=1}^{n_i}(1-\eta_{ij})E_{\theta_0}^{T}(\frac{I(T \le t^*)}{P[Q_2(t,O,\theta_0)]|_{t=T}}|T \le T_{ij})]$, for each t^* in $[0,\tau]$ can be established. It is easy to see that sum of the above two limits is $\Lambda_0(t)$. Glivenko–Cantelli lemma and the continuity of Λ_0 imply that $\|\bar{\Lambda}_n - \Lambda_0\|$ converges to 0 with probability 1. By the definition of $\hat{\Lambda}_n(t^*)$ and $\bar{\Lambda}_n(t^*)$, we have that $\hat{\Lambda}_n(t^*)$ is absolutely continuous with respect to $\bar{\Lambda}_n(t^*)$, and

$$\hat{\Lambda}_n(t^*) = \int_0^{t^*} \frac{P_n\{Q_3(v,O,\theta_0)\}}{P_n\{Q_3(v,O,\hat{\theta})\}}\, d\bar{\Lambda}_n(v),$$

where $Q_3(v,O,\theta) = \sum_{s=1}^{n_r} \frac{\delta_{rs}\eta_{rs}I(v \le T)+nP_n[Q_1(v,O,\theta)]I(v \le T)}{P_n[Q_2(v,O,\theta)]}$. By applying the Glivenko–Cantelli property along with the dominance convergence theorem the convergence of $\hat{\theta}_{q(n)}$ to θ^* implies that

$$\Lambda^*(t^*) = \int_0^{t^*} \frac{E_{\theta_0}\{Q_3(v,O,\theta_0)\}}{E_{\theta^*}\{Q_3(v,O,\theta^*)\}}\, d\Lambda_0(v).$$

Then the absolute continuity of Λ^* with respect to Λ_0 holds. Moreover, $\frac{d\hat{\Lambda}_n(t)}{d\bar{\Lambda}_n(t)}$ converges uniformly to $\frac{d\Lambda^*(t)}{d\Lambda_0(t)}$.

Now we consider the following difference in log-likelihood:

$$\frac{1}{n}l_n(\hat{\theta}_n) - \frac{1}{n}l_n(\beta_0, \bar{\Lambda}_n, \gamma_0) \ge 0.$$

The left-hand side converges a.s. to $E_{\theta_0}[l(\theta^*) - l(\theta_0)]$ by Lebesgue's theorem. Since the limit is the Kullback–Leibler divergence which is non-positive, the only possibility is that the limit is exactly zero. By the identifiability under conditions C1 to C4, we conclude that $\theta^* = \theta_0$. The proof is now complete. □

A.3. Proof of Theorem 4.2.

PROOF OF ASYMPTOTIC NORMALITY. The proof will follow the framework of Theorem 3.3.1 in van der Vaart and Wellner (1996) and involves several key steps.

Let $H_p = \{h = (h_1, h_2, h_3) : |h_1| + \|h_2\|_{\mathrm{v}} + |h_3| \le p\}$, where h_1 and $h_3 \in \mathbb{R}^1$, h_2 is a function of bounded variation on $[0, \tau]$, and $\|h_2\|_{\mathrm{v}}$ denotes the sum of the absolute value of h_2 at 0 and the total variation of h_2 on $[0, \tau]$. Here, we consider $\theta = (\beta, \Lambda, \gamma)$ as a functional on H_p defined as

$$\theta(h) = (\beta, \Lambda, \gamma)(h_1, h_2, h_3) = h_1\beta + \int_0^{\tau} h_2(u)\, d\Lambda(u) + h_3\gamma.$$

Hence, the parameter space Θ is a subspace of $l_\infty(H_p)$. To verify the Fréchet differentiability of the score function, for a fixed $h = (h_1, h_2, h_3) \in H_p$, we shall consider an one-dimensional submodel $\theta_t = (\beta + th_1, \Lambda_t(h_2), \gamma + th_3)$, where $t \in$

$\mathbb{R}^1$ and $\Lambda_t(h_2)(\cdot) = \int_0^{\cdot}[1 + th_2(u)]\,d\Lambda(u)$. Here, for sufficiently small $|t|$, $\Lambda_t(h_2)$ satisfies the requirements of a cumulative hazard function since h_2 is a function of bounded variation on $[0, \tau]$.

Let $\tilde{\theta}$ denote a certain value of θ, the score function for t at θ along the direction h is

$$S_{\tilde{\theta}}(\theta)(h) = h_1 S_{\tilde{\theta},1}(\theta) + S_{\tilde{\theta},2}(\theta)(h_2) + h_3 S_{\tilde{\theta},3}(\theta), \tag{15}$$

where

$$\begin{aligned} S_{\tilde{\theta},1}(\theta) = E_{\theta_0}\Biggl\{\sum_{j=1}^{n^*} &\delta_j \eta_j z_j - \eta_j z_j \exp(\beta z_j)\Lambda(\tilde{T}_j)E_{\tilde{\theta}}(W|O) \\ &+ (1-\eta_j)z_j - (1-\eta_j)z_j \exp(\beta z_j) E_{\tilde{\theta}}\bigl(W\Lambda(U_j)|O\bigr)\Biggr\}, \end{aligned}$$

$$\begin{aligned} S_{\tilde{\theta},2}(\theta)(h_2) = E_{\theta_0}\Biggl\{\sum_{j=1}^{n^*} &\delta_j \eta_j h_2(\tilde{T}_j) + (1-\eta_j)E_{\tilde{\theta}}\bigl(h_2(U_j)|O\bigr) \\ &- \eta_j \exp(\beta z_j) E_{\tilde{\theta}}(W|O)\int_0^{\tilde{T}_j} h_2(u)\,d\Lambda(u) \\ &- (1-\eta_j)\exp(\beta z_j) E_{\tilde{\theta}}\biggl(W\int_0^{U_j} h_2(u)\,d\Lambda(u)\Big|O\biggr)\Biggr\}, \end{aligned}$$

and

$$S_{\tilde{\theta},3}(\theta) = E_{\theta_0}\biggl\{E_{\tilde{\theta}}\biggl[\frac{\partial f_W(W|V)/\partial t}{f_W(W|V)}\Big|O\biggr]\biggr\}.$$

The corresponding Fréchet derivative of the score at the true value θ_0 can be shown to be

$$\nabla_\theta S_{\theta_0}(\theta_0)(h) = -\beta\sigma_{\theta_0,1}(h) - \int_0^\tau \sigma_{\theta_0,2}(h)(u)\,d\Lambda(u) - \gamma\sigma_{\theta_0,3}(h), \tag{16}$$

where

$$\begin{aligned} \sigma_{\theta_0,1}(h) = E_{\theta_0}\Biggl\{\sum_{j=1}^{n^*}\biggl[&h_1\bigl(\eta_j W\Lambda_0(\tilde{T}_j) + (1-\eta_j)W\Lambda_0(U_j)\bigr)z_j^2 \exp(\beta_0 z_j) \\ &+ \biggl(\eta_j W\int_0^{\tilde{T}_j} h_2(u)\,d\Lambda_0(u) \\ &+ (1-\eta_j)W\int_0^{U_j} h_2(u)\,d\Lambda_0(u)\biggr)z_j \exp(\beta_0 z_j)\biggr]\Biggr\}, \end{aligned}$$

$$\sigma_{\theta_0,2}(h)(u) = E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\big[h_1\big(\eta_j W I(u \le \tilde{T}_j) + (1-\eta_j) W I(u \le U_j)\big) z_j \exp(\beta_0 z_j)$$
$$+ W h_2(u)\big(\eta_j I(u \le \tilde{X}_j) + (1-\eta_j) I(u \le U_j)\big)\exp(\beta_0 z_j)\big]\Bigg\},$$

and

$$\sigma_{\theta_0,3}(h) = E_{\theta_0}\bigg\{h_3 \frac{\partial^2}{\partial t^2}\log f_W(W|\gamma_0 + t\gamma)\bigg|_{t=0}\bigg\}.$$

We shall term $\sigma_{\theta_0} = (\sigma_{\theta_0,1}, \sigma_{\theta_0,2}, \sigma_{\theta_0,3})$ the Fisher information operator. Since the Fréchet derivative $\nabla_\theta S_{\theta_0}(\theta_0)$ is a linear form of σ, it suffices to show the continuous invertibility of σ by proving: (i) its one-to-one property, and (ii) it can be expressed as a sum of a continuously invertible operator and a compact operator. The one-to-one property (i) can be illustrated by apagogical argument which is a consequence of the identifiability of the model.

To demonstrate (ii), we consider the following decomposition of the Fisher information operator:

$$\sigma_{\theta_0}(h) = \sigma_{\theta_0,L}(h) + \sigma_{\theta_0,C}(h),$$

where

$$\sigma_{\theta_0,L}(h) = \Bigg(h_1 E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\big[\big(\eta_j W \Lambda_0(\tilde{T}_j) + (1-\eta_j) W \Lambda_0(U_j)\big) z_j^2 \exp(\beta_0 z_j)\big]\Bigg\},$$
$$h_2(u) E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\big[W\big(\eta_j I(u \le \tilde{X}_j) + (1-\eta_j) I(u \le U_j)\big)\exp(\beta_0 z_j)\big]\Bigg\},$$
$$h_3 E_{\theta_0}\bigg\{\frac{\partial^2}{\partial t^2}\log f_W(W|\gamma_0 + t\gamma)\bigg|_{t=0}\bigg\}\Bigg)$$

and

$$\sigma_{\theta_0,C}(h) = \Bigg(E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\Bigg[\bigg(\eta_j W \int_0^{\tilde{T}_j} h_2(u)\,d\Lambda_0(u)$$
$$+ (1-\eta_j) W \int_0^{U_j} h_2(u)\,d\Lambda_0(u)\bigg) z_j \exp(\beta_0 z_j)\Bigg]\Bigg\},$$
$$h_1 E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\big[\big(\eta_j W I(u \le \tilde{T}_j)$$
$$+ (1-\eta_j) W I(u \le U_j)\big) z_j \exp(\beta_0 z_j)\big]\Bigg\}, 0\Bigg).$$

The continuous invertibility of $\sigma_{\theta_0,L}$ is straightforward under assumptions A1 to A6. To show the compactness of $\sigma_{\theta_0,C}$, we consider a sequence $h_n = (h_{1,n}, h_{2,n}, h_{3,n}) \in H_p$, and prove the existence of a convergent subsequence of $\sigma_{\theta_0,C}(h_n)$. By applying Helly's selection theorem along with the Bolzanno–Weierstrass theorem, we obtain a subsequence $h_{q(n)}$ of h_n which converges to a limit $h^* = (h_1^*, h_2^*, h_3^*)$. Since the norm of the distance between $\sigma_{\theta_0,C}(h_{q(n)})$ and $\sigma_{\theta_0,C}(h^*)$ can be expressed as

$$\begin{aligned}
&\Bigg| E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}\bigg[\bigg(\eta_j W\int_0^{\tilde{T}_j}(h_{2,q(n)} - h_2^*)(u)\,d\Lambda_0(u) \\
&\qquad + (1-\eta_j)W\int_0^{U_j}(h_{2,q(n)} - h_2^*)(u)\,d\Lambda_0(u)\bigg)z_j\exp(\beta_0 z_j)\bigg]\Bigg\}\Bigg| \\
&\qquad + \Bigg\|(h_{1,q(n)} - h_1^*)E_{\theta_0}\Bigg\{\sum_{j=1}^{n^*}[(\eta_j W I(u \le \tilde{T}_j) \\
&\qquad + (1-\eta_j)W I(u \le U_j))z_j\exp(\beta_0 z_j)]\Bigg\}\Bigg\|_{\mathbf{v}},
\end{aligned} \tag{17}$$

assumption A1 to A6 imply that (17) is bounded above by

$$c\bigg[\int_0^\tau |(h_{2,q(n)} - h_2^*)(u)|\,d\Lambda_0(u) + |h_{1,q(n)} - h_1^*|\bigg],$$

for some constant c. The dominated convergence theorem gives the convergence of the upper bound to zero, and then implies the convergence of $\sigma_{\theta_0,C}(h_{q(n)})$ to $\sigma_{\theta_0,C}(h^*)$. The operator $\sigma_{\theta_0,C}$ has been shown to be compact and then the continuous invertibility of σ_{θ_0} holds.

In the following step, we demonstrate the convergence of the difference between the empirical score process $S_{n,\hat{\theta}_n}$ and the mean score process S_{θ_0} evaluated at the true θ_0. The definition of $S_{n,\hat{\theta}_n}(\theta)$ and $S_{\theta_0}(\theta)$ are defined as follows. For the empirical score process, we define

$$S_{n,\hat{\theta}_n}(\theta)(h) = h_1 S_{n,\hat{\theta}_n,1}(\theta) + S_{n,\hat{\theta}_n,2}(\theta)(h_2) + h_3 S_{n,\hat{\theta}_n,3}(\theta),$$

where

$$\begin{aligned}
S_{n,\hat{\theta}_n,1}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\sum_{j=1}^{n^*}\delta_j\eta_j z_j - \eta_j z_j\exp(\beta z_j)\Lambda(\tilde{T}_j)E_{\hat{\theta}_n}(W|O) \\
&+ \delta_j(1-\eta_j)z_j - \delta_j(1-\eta_j)z_j\exp(\beta z_j)E_{\hat{\theta}_n}\big(W\Lambda(U_j)|O\big)\Bigg\},
\end{aligned}$$

$$S_{n,\hat{\theta}_n,2}(\theta)(h_2) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\sum_{j=1}^{n^*}\delta_j\eta_j h_2(\tilde{T}_j) + \delta_j(1-\eta_j)E_{\hat{\theta}_n}\big(h_2(U_j)|O\big)$$

$$- \eta_j \exp(\beta z_j) E_{\hat{\theta}_n}(W|O) \int_0^{\tilde{T}_j} h_2(u)\, d\Lambda(u)$$

$$- \delta_j(1-\eta_j) \exp(\beta z_j) E_{\hat{\theta}_n}\biggl(W \int_0^{U_j} h_2(u)\, d\Lambda(u) \Big| O\biggr)\biggr\},$$

$$S_{n,\hat{\theta}_n,3}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \biggl\{ E_{\hat{\theta}_n}\biggl[\frac{\partial f_W(W|\gamma)/\partial t}{f_W(W|\gamma)} \Big| O\biggr]\biggr\}.$$

For the mean score process, we define

$$S_{\theta_0}(\theta)(h) = h_1 S_{\theta_0,1}(\theta) + S_{\theta_0,2}(\theta)(h_2) + h_3 S_{\theta_0,3}(\theta),$$

where

$$S_{\theta_0,1}(\theta) = E_{\theta_0}\biggl\{\sum_{j=1}^{n^*} \delta_j \eta_j z_j - \eta_j z_j \exp(\beta z_j) \Lambda(\tilde{T}_j) W$$

$$+ \delta_j(1-\eta_j) z_j - \delta_j(1-\eta_j) z_j \exp(\beta z_j) W \Lambda(U_j)\biggr\},$$

$$S_{\theta_0,2}(\theta)(h_2) = E_{\theta_0}\biggl\{\sum_{j=1}^{n^*} \delta_j \eta_j h_2(\tilde{T}_j) + \delta_j(1-\eta_j) h_2(U_j)$$

$$- \eta_j \exp(\beta z_j) W \int_0^{\tilde{T}_j} h_2(u)\, d\Lambda(u)$$

$$- \delta_j(1-\eta_j) \exp(\beta z_j) W \int_0^{U_j} h_2(u)\, d\Lambda(u)\biggr\},$$

$$S_{\theta_0,3}(\theta) = E_{\theta_0}\biggl\{\frac{\partial f_W(W|\gamma)/\partial t}{f_W(W|\gamma)}\biggr\}.$$

To illustrate the convergence of the process $\sqrt{n}(S_{n,\hat{\theta}_n}(\theta_0) - S_{\theta_0}(\theta_0))(h)$, the main point is to demonstrate the Donsker property of the classes of functions shown in $S_{n,\hat{\theta}_n}(\theta_0)$. The Donsker property on the class $\{h_1 S_{n,\hat{\theta}_n,1}(\theta_0) + h_3 S_{n,\hat{\theta}_n,3}(\theta_0) : |h_1|, |h_3| \le p\}$ holds due to the boundedness assumption in A4 and A5, and the fact that it is a parametric class, parameterized by h on a bounded subset, of measurable score function. This is illustrated by van der Vaart [Example 19.7 in van der Vaart (1998)]. Moreover, according to the fact that a class of functions that are both uniformly bounded on $[0, \tau]$ and of bounded variation is Donsker, the Donsker property holds for the class $\{S_{n,\hat{\theta}_n,2}(\theta_0)(h_2) : h_2 \in \mathrm{BV}_p\}$, where BV_p is the space of functions of bounded variation whose total variations are smaller than p on $[0, \tau]$. This leads to the convergence of $\sqrt{n}(S_{n,\hat{\theta}_n}(\theta_0) - S_{\theta_0}(\theta_0))(h)$ to a tight element on $l_\infty(H_p)$.

Next, we verify condition (a) in Theorem 3.3.1 in van der Vaart and Wellner (1996). From now on, we denote the score functions based on one cluster by $s_{\tilde{\theta},O}(\theta)(h) = h_1 s_{\tilde{\theta},O,1}(\theta) + s_{\tilde{\theta},O,2}(\theta)(h_2) + h_3 s_{\tilde{\theta},O,3}(\theta)$, where

$$\begin{aligned}
s_{\tilde{\theta},O,1}(\theta) &= \sum_{j=1}^{n^*} \delta_j \eta_j z_j - \eta_j z_j \exp(\beta z_j) \Lambda(\tilde{T}_j) E_{\tilde{\theta}}(W|O) \\
&\quad + \delta_j(1-\eta_j) z_j - \delta_j(1-\eta_j) z_j \exp(\beta z_j) E_{\tilde{\theta}}\big(W\Lambda(U_j)|O\big), \\
s_{\tilde{\theta},O,2}(\theta)(h_2) &= \sum_{j=1}^{n^*} \delta_j \eta_j h_2(\tilde{T}_j) + \delta_j(1-\eta_j) E_{\tilde{\theta}}\big(h_2(U_j)|O\big) \\
&\quad - \eta_j \exp(\beta z_j) E_{\tilde{\theta}}(W|O) \int_0^{\tilde{T}_j} h_2(u)\, d\Lambda(u) \\
&\quad - \delta_j(1-\eta_j) \exp(\beta z_j) E_{\tilde{\theta}}\biggl(W \int_0^{U_j} h_2(u)\, d\Lambda(u) \Big| O\biggr), \\
s_{\tilde{\theta},O,3}(\theta) &= E_{\tilde{\theta}}\biggl[\frac{\partial f_W(W|\gamma)/\partial t}{f_W(W|\gamma)} \Big| O\biggr].
\end{aligned}$$

According to Lemma 3.3.5 in van der Vaart and Wellner (1996), it suffices to show the following two steps: (i) the class of random functions $\{s_{\theta,O}(\theta)(h) - s_{\theta_0,O}(\theta_0)(h)\} : \|\theta - \theta_0\| < \delta, h \in H\}$, for certain $\delta > 0$, is Donsker, and (ii) $\sup_{h\in H} E_{\theta_0}\{s_{\theta,O}(\theta)(h) - s_{\theta_0,O}(\theta_0)(h)\}^2 \to 0$ as $\theta \to \theta_0$. The Donsker property for the class in (i) can be verified in a similar way as shown previously for condition (b) by looking at $s_{\theta,O,k}(\theta) - s_{\theta_0,O,k}(\theta_0)$, $k = 1, 2, 3$. The second step follows from the dominated convergence theorem. Therefore, condition (a) holds.

We have now verified conditions (a), (b) and (c) in Theorem 3.3.1 in van der Vaart and Wellner (1996). Along with the consistency of $\hat{\theta}_n$ shown in Theorem 4.1, the weak convergence of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is concluded. □

PROOF OF SEMIPARAMETRIC EFFICIENCY. The Fréchet differentiability and the $\sqrt{n}$-consistency of $\hat{\theta}$ shown previously imply

$$\sqrt{n}\, \nabla_{\hat{\theta}_n - \theta_0} S_{\theta_0}(\theta_0)(h) = \sqrt{n}\big(S_{n,\theta_0}(\theta_0)(h) - S_{\theta_0}(\theta_0)(h)\big) + o_p(1), \tag{18}$$

where $o_p(1)$ is a random term converges in probability to zero element in $l_\infty(H_p)$. Since the continuous invertibility of the Fisher information operator σ has been verified, its inverse operator, denoted as σ^{-1}, exists and for each given h we have $\tilde{h} = (\tilde{h}_1, \tilde{h}_2, \tilde{h}_3) = \sigma^{-1}(h)$. By replacing h by $\tilde{h}$ on the right-hand-side of (18) and according to (16), we obtain the following equation:

$$\begin{aligned}
&\sqrt{n}[S_{n,\theta_0}(\theta_0)(\tilde{h}) - S_{\theta_0}(\theta_0)(\tilde{h})] + o_p(1) \\
&\quad = \sqrt{n}\, \nabla_{\hat{\theta}_n - \theta_0} S_{\theta_0}(\theta_0)(\tilde{h}) = \sqrt{n}\, \nabla_{\hat{\theta}_n - \theta_0} S_{\theta_0}(\theta_0)(\sigma^{-1}(h)) \qquad (19) \\
&\quad = \sqrt{n}\biggl[-(\hat{\beta}_n - \beta_0)h_1 - \int_0^\tau h_2(u)\, d(\hat{\Lambda}_n - \Lambda_0)(u) - (\hat{\gamma}_n - \gamma_0)h_3\biggr].
\end{aligned}$$

Hence, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly to a tight Gaussian element in $l_\infty(H_p)$. By taking $h_2 = 0$ in (19), we observe that the influence function for $(\hat{\beta}_n h_1, \hat{\gamma}_n h_3)$ is a linear span of the score functions. By applying Proposition 3.3.1 in van der Vaart and Wellner (1996), the semiparametric efficiency of $(\hat{\beta}, \hat{\gamma})$ is concluded. □

A.4. Proof of Theorem 4.3. We complete the proof of the consistency of the bootstrap standard error by verifying the three conditions M1–M3 listed in Theorem 1 in Cheng (2015). In the following, we denote the log-likelihood contributed by a cluster by $l(\theta)$. Condition M1, which states the quadratic behavior of the log-likelihood, can be illustrated by considering the second-order Taylor expansion on the expected log-likelihood $E_{\theta_0} l(\theta)$. By the identifiability of the model, θ_0 maximizes $E_{\theta_0} l(\theta)$; hence the expected difference between $l(\theta)$ and $l(\theta_0)$ can be expressed by a linear form of the information operator defined in (16), with θ replaced by $\theta - \theta_0$, plus the remainder term. By assumptions A1–A6, $E_{\theta_0}[l(\theta) - l(\theta_0)]$ is bounded above by a certain constant times $|\beta - \beta_0|^2 + |\gamma - \gamma_0|^2 + \|\Lambda - \Lambda_0\|_\infty^2$. Thus, condition M1 holds for the current model. Condition M3 in Theorem 1 in Cheng (2015) requires the $\sqrt{n}$-consistency of the NPMLE $\hat{\Lambda}$ and the bootstrap NPMLE $\hat{\Lambda}^*$. The $\sqrt{n}$-consistency of $\hat{\Lambda}$ is illustrated in Theorem 4.1. Analogously, the $\sqrt{n}$-consistency of $\hat{\Lambda}^*$ can be verified since the log-likelihood from a nonparametric bootstrap sample can be expressed as $\sum_{i=1}^n M_{ni} l_i(\theta)$, where M_{ni} is the frequency of the ith cluster being resampled, and $(M_{n1}, \dots, M_{nn}) \sim \text{Multinomial}(n, (n^{-1}, \dots, n^{-1}))$.

Condition M2 in Theorem 1 in Cheng (2015) describes the moment condition of the empirical process over a class of functions defined as $\mathcal{N}_\delta = \{l(\theta) - l(\theta_0) : \theta \in \Theta, |\beta - \beta_0| \leq \delta, |\gamma - \gamma_0| \leq \delta, \|\Lambda - \Lambda_0\|_\infty \leq \delta\}$ for some $\delta > 0$. Here, we introduce some notation for later use. Let N_δ be the envelop function of the class $\mathcal{N}_\delta$. Define empirical processes $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P) f$ and $\mathbb{G}_n^* f = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) f$, where $Pf = \int f\, dP$, $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$, and $\mathbb{P}_n^* f = \frac{1}{n}\sum_{i=1}^n f(X_i^*)$. The notation with an $*$ denote the corresponding terms based on bootstrap samples. Moreover, we define $\|\mathbb{G}_n\|_{\mathcal{N}_\delta} = \sup_{f \in \mathcal{N}_\delta} |\mathbb{G}_n f|$ and $\|\mathbb{G}_n^*\|_{\mathcal{N}_\delta} = \sup_{f \in \mathcal{N}_\delta} |\mathbb{G}_n^* f|$, and use the notation "$a(b) \lesssim b$" to mean that $a(b)$ is smaller than b, for all b, up to an universal constant.

Since the function $l(\theta) - l(\theta_0)$, for fixed θ_0, is globally Lipschitz continuous with the Lipschitz coefficient function as a finite constant function under the assumption A1–A6, we have the $L_{p'}$-norm of the envelop function

$$\|N_\delta\|_{L_{p'}(P)} \lesssim \delta. \tag{20}$$

It also implies that

$$\|N_\delta\|_{L_{p'}(P)} < \infty. \tag{21}$$

By the compactness of the finite-dimensional parameter space $\Theta_\beta \times \Theta_\gamma$ and the fact that the class of bounded monotone functions is VC-hull class, the class $\mathcal{N}_\delta$

has finite uniform entropy integral. This fact along with (20) imply

$$\left(E_{\theta_0}\|\mathbb{G}\|_{\mathcal{N}_\delta}^{p'}\right)^{1/p'} \lesssim \delta. \tag{22}$$

Moreover, under nonparametric sampling scheme, (21) and (22) lead to the following inequality according to Appendix A.5 in Cheng (2015):

$$\left(E_{\theta_0}\|\mathbb{G}^*\|_{\mathcal{N}_\delta}^{p'}\right)^{1/p'} \lesssim \delta. \tag{23}$$

The two inequalities in (22) and (23) complete the verification of condition M2, and hence the theorem.

Acknowledgements. We thank the Editor, the Associate Editor and two anonymous reviewers; their insightful suggestions greatly improve this paper. Furthermore, we are grateful to Dr. Masanao Yajima at Fred Hutchinson Cancer Research Center for his helpful comments.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data" (DOI: 10.1214/15-AOS1406SUPP; .pdf). Owing to the space constraints, we present the proof of Proposition 3.2 in the supplemental material [Su and Wang (2015)].

REFERENCES

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc. B* **61** 265–285.

CAFFO, B. S., JANK, W. and JONES, G. L. (2005). Ascent-based Monte Carlo expectation-maximization. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 235–251. MR2137323

CAI, T. and CHENG, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* **91** 277–290. MR2081301

CHAN, K. S. and LEDOLTER, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.* **90** 242–252. MR1325132

CHANG, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* **18** 391–404. MR1041399

CHANG, M. N. and YANG, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15** 1536–1547. MR0913572

CHENG, G. (2015). Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation. *Scand. J. Stat.* **42** 665–684.

CHENG, G. and HUANG, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.* **38** 2884–2915. MR2722459

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509

DE GRUTTOLA, V. and LAGAKOS, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45** 1–11. MR0999438

DUPUY, J.-F., GRAMA, I. and MESBAH, M. (2006). Asymptotic theory for the Cox model with missing time-dependent covariate. *Ann. Statist.* **34** 903–924. MR2283397

FORT, G. and MOULINES, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259. MR2001649

KIM, Y.-J. (2006). Regression analysis of doubly censored failure time data with frailty. *Biometrics* **62** 458–464. MR2227493

KIM, M. Y., DE GRUTTOLA, V. and LAGAKOS, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49** 13–22.

KIM, Y., KIM, B. and JANG, W. (2010). Asymptotic properties of the maximum likelihood estimator for the proportional hazards model with doubly censored data. *J. Multivariate Anal.* **101** 1339–1351. MR2609496

KIM, Y., KIM, J. and JANG, W. (2013). An EM algorithm for the proportional hazards model with doubly censored data. *Comput. Statist. Data Anal.* **57** 41–51. MR2981071

MURPHY, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22** 712–731. MR1292537

MURPHY, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23** 182–198. MR1331663

MURPHY, S. A., ROSSINI, A. J. and VAN DER VAART, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127

MYKLAND, P. A. and REN, J.-J. (1996). Algorithms for computing self-consistent and maximum likelihood estimators with doubly censored data. *Ann. Statist.* **24** 1740–1764. MR1416658

NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K. and SØRENSEN, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Stat.* **19** 25–43. MR1172965

PARNER, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.* **26** 183–214. MR1611788

RIPATTI, S. and PALMGREN, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56** 1016–1022. MR1806744

SU, Y.-R. (2011). Survival analysis for incomplete data. Ph.D. thesis, Univ. California, Davis.

SU, Y. and WANG, J. (2015). Supplement to "Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data." DOI:10.1214/15-AOS1406SUPP.

THERNEAU, T. M., GRAMBSCH, P. M. and PANKRATZ, V. S. (2003). Penalized survival models and frailty. *J. Comput. Graph. Statist.* **12** 156–175. MR1965213

TSENG, Y.-K., HSIEH, F. and WANG, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92** 587–603. MR2202648

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69** 169–173. MR0381120

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

VAUPEL, J. W., MANTON, K. G. and STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16** 439–454.

WU, J. F., CHEN, C. C., HSIEH, R. P., SHIH, H. H., CHEN, Y. H., LI, C. R., CHIANG, C. Y., SHAU, W. Y., NI, Y. H., CHEN, H. L., HSU, H. Y. and CHANG, M. H. (2006). HLA typing associated with hepatitis B E antigen seroconversion in children with chronic hepatitis B virus infection: A long-term prospective sibling cohort study in Taiwan. *J. Pediatr.* **148** 647–651.

ZENG, D. and CAI, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann. Statist.* **33** 2132–2163. MR2211082

ZHANG, Y. and JAMSHIDIAN, M. (2004). On algorithms for the nonparametric maximum likelihood estimator of the failure function with censored data. *J. Comput. Graph. Statist.* **13** 123–140. MR2044874

BIOSTATISTICS AND BIOMATHEMATICS
PUBLIC HEALTH SCIENCE DIVISION
FRED HUTCHINSON CANCER RESEARCH CENTER
SEATTLE, WASHINGTON 98109
USA
E-MAIL: ysu@fredhutch.org

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA 95616
USA
E-MAIL: jlwang.ucdavis@gmail.com