Inference on Clustered Survival Data Using Imputed Frailties

Yi LI, Louise RYAN, Scarlett BELLAMY, and Glen A. SATTEN

This article proposes a new method for fitting frailty models to clustered survival data that is intermediate between the fully parametric and nonparametric maximum likelihood estimation approaches. A parametric form is assumed for the baseline hazard, but only for the purpose of imputing the unobserved frailties. The regression coefficients are then estimated by solving an estimating equation that is the average of the partial likelihood score with respect to the conditional distribution of frailties given the observed data. We prove consistency and asymptotic normality of the resulting estimators and give associated closed-form estimators of their variance. The algorithm is easy to implement and reduces to the ordinary Cox partial likelihood approach when the frailties have a degenerate distribution. Simulations indicate high efficiency and robustness of the resulting estimates. We apply our new approach to a study with clustered survival data on asthma in children in east Boston.

Key Words: Asymptotic normality; Bootstrap; Consistency; Cox models; Frailty models; Imputed frailty partial likelihood score (IFPLS); Monte Carlo estimating equation; S-U algorithm.

1. INTRODUCTION

Dependent or clustered survival data arise frequently in medical research; for example, in familial studies or multi-center clinical trials. In the East Boston Asthma Study, conducted by Rosalind Wright at the Channing Laboratory, Harvard Medical School, efforts are being made to understand the etiology of the rising prevalence and morbidity of childhood asthma, and of the disproportionate burden among urban minority children. A total of 753 subjects from 25 neighborhoods were enrolled at community health clinics throughout east Boston between 1986 and 1992. As children from the same neighborhoods share similar environ-

---

Yi Li is Assistant Professor, and Louise Ryan is Professor, Department of Biostatistics, Harvard University and Dana-Farber Cancer Institute, 44 Binney Street, M232, Boston, MA 02115 (E-mail: yili@jimmy.harvard.edu *and* lryan@hsph.harvard.edu). Scarlett Bellamy is Assistant Professor, Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: sbellamy@cceb.upenn.edu). Glen A. Satten is Mathematical Statistician, Centers for Disease Control, 1600 Clifton Road, Atlanta, GA 30333 (E-mail: gas0@cdc.gov).

Journal of Computational and Graphical Statistics, Volume 12, Number 3, Pages 1–24
DOI: 10.1198/1061860032247

mental and social factors, correct inference needs to take within-cluster dependence into account. Frailty models (Clayton and Cuzick 1985), a natural extension of the Cox model, are increasingly popular for such settings, and typically involve adding random effects into the baseline hazard to model the correlation among observations within the same cluster. A challenge in fitting frailty models, however, is that the standard partial likelihood approach does not apply and one has to estimate the regression coefficients, the variance components, and possibly the baseline hazard simultaneously.

Two major approaches, parametric and nonparametric, are available to fitting frailty models for clustered survival data. Parametric methods involve specifying a parametric form on the baseline hazard and then maximizing the marginal likelihood to obtain the maximum likelihood estimates of unknown parameters (Andersen, Borgan, Gill, and Keiding 1992, p. 668; Shoukri, Attanasio, and Sargeant 1998). On the other hand, with the functional form of the baseline hazard unspecified, Nielsen, Gill, Andersen, and Sorensen (1992), Murphy (1994, 1995), and Parner (1998) discussed the use of nonparametric maximum likelihood estimation (NPMLE) to fit frailty models. For a review, see Oakes and Jeong (1998).

Although parametric MLEs are consistent and most efficient when the baseline hazard is correctly specified, they are generally biased when the baseline hazard is misspecified. In contrast, NPMLEs are robust with respect to such misspecification. There are, however, several obvious drawbacks. First, the computation of NPMLEs is complex and the convergence is usually slow. Second, the efficiency of NPMLEs is low compared with that of parametric MLEs (Shoukri et al. 1998). Third, the variance estimators for NPMLEs are hard to calculate (Murphy 1995; Parner 1998).

Because of the difficulties described above, we propose a method that is intermediate between the fully parametric and nonparametric approaches. A parametric form is assumed for the baseline hazard, but it is used only to impute the unobserved frailties. Once the baseline distribution is specified, the distribution of frailties conditional on the observed data can be calculated. A new estimating equation can be obtained by averaging the score equation for the Cox partial likelihood with respect to the conditional distribution of frailties. We propose to use this average partial likelihood score equation to estimate regression parameters, and give closed-form estimators of the sampling variance of our proposed estimators. Because the average over the conditional distribution of frailties cannot be carried out analytically, and because this distribution depends on unknown parameters to be estimated, we use the S-U algorithm (Satten and Datta 2000) and a novel Monte Carlo estimating equation (MCEE) approach to calculate the estimators. Our method reduces to the ordinary Cox partial likelihood approach when the frailties have a degenerate distribution. The algorithms are easy to implement and the efficiency of the resulting estimates is high. Though a disadvantage of our method is its dependence on the parametric structure of the baseline hazard, simulations indicate that the choice of baseline hazard has only a slight effect on the value of the estimated regression coefficients. A similar approach was proposed by Satten, Datta, and Williamson (1998) for independent interval-censored data analysis, where they imputed unobserved failure times based on a parametric model and, thereafter, calculated the average partial likelihood score based on the imputed data for the estimation of the regression coefficients.

The rest of the article is structured as follows. Section 2 presents the model and derives the average partial likelihood score equations. Section 3 develops asymptotic results. Section 4 details the two algorithms, the S-U algorithm and the Monte Carlo estimating equation, for solving the proposed estimating equations. Section 5 gives consistent variance estimators for the resulting parameter estimates. Sections 6 and 7 conduct simulations to assess the finite sample performance of the proposed methods and apply the methods to the analysis of the East Boston Asthma Study. We conclude with general discussion in Section 8.

2. IMPUTED FRAILTY PARTIAL LIKELIHOOD SCORE EQUATION

A distinctive feature of the imputed frailty partial likelihood score (IFPLS) approach is that unobserved frailties are effectively treated as missing covariates, and their imputed values based on the conditional distributions are used to construct an average partial likelihood score estimating equation. As the partial likelihood score does not involve the baseline hazard, one might expect the resulting estimates to be more robust than the fully parametric maximum likelihood estimates. In recent years a variety of Monte Carlo procedures for solving missing data problems based on data augmentation or imputing missing data have been developed (Tanner 1993). We proceed, in the following, by stating the frailty survival model, followed by the derivation of the imputed frailty partial likelihood score equation.

Let V_{ij}, C_{ij}, and $\mathbf{X}_{ij}$ $(r \times 1)$ be the failure time, the censoring time, and the covariate vector for subject j in cluster i, $j = 1, \ldots, n_i$, $i = 1, \ldots, M$. We assume that the C_{ij} are independently and identically distributed and independent of the V_{ij}, conditional on the observed covariate $\mathbf{X}_{ij}$. The observed data are right censored with only $T_{ij} = \min\{V_{ij}, C_{ij}\}$ and the censoring code $\delta_{ij} = I(V_{ij} \leq C_{ij})$ observed, where $I(\cdot)$ denotes an indicator function. Clusters are assumed to be independent, cluster-specific frailties b_i are iid and failure times within a cluster are independent, conditional on the b_i. Our model specifies that, conditional on the covariates and an unobserved cluster-specific frailty b_i, the survival time V_{ij} is independent and has an intensity function as follows

$$\lambda(t|\mathbf{X}_{ij}, b_i) = \lambda_0(t, \boldsymbol{\alpha}) \exp(\mathbf{X}_{ij}'\boldsymbol{\beta} + b_i), \tag{2.1}$$

where $\boldsymbol{\beta}$ $(r \times 1)$ is a vector of unknown fixed effect parameters, $\lambda_0(t, \boldsymbol{\alpha})$ is a continuous baseline hazard function depending on an unknown parameter $\boldsymbol{\alpha}$, and b_i is a zero mean random variable with density function $f(\cdot; \theta)$ and distribution function $F(\cdot; \theta)$, specified up to an unknown variance parameter θ. A common choice of the distribution of the frailty b_i would be a log-Gamma distribution (Clayton and Cuzick 1985; Murphy 1995) or a normal distribution (Li and Lin 2000).

Define the standard event counting processes by $N_{ij}(t) = I(T_{ij} \leq t, \delta_{ij} = 1)$, at-risk processes by $Y_{ij}(t) = I(T_{ij} \geq t)$, and the cumulative baseline hazard by $\Lambda_0(t, \boldsymbol{\alpha}) = \int_0^t \lambda_0(s, \boldsymbol{\alpha}) ds$. For notational convenience, we denote $\mathbf{b} = (b_1, \ldots, b_M)$, $\mathbf{T}_i = (T_{i1}, \ldots, T_{in_i})$, $\mathbf{T} = (T_1, \ldots, T_M)$, and similarly for $\mathbf{X}_i, \mathbf{X}, \boldsymbol{\Delta}_i, \boldsymbol{\Delta}, \mathbf{N}_i(t), \mathbf{N}(t), \mathbf{Y}_i(t)$, and $\mathbf{Y}(t)$.

We assume that the covariates $\mathbf{X}_{ij}$ are constant over time. Throughout, unless specified, $F(\cdot)$ represents a distribution function and $f(\cdot)$ a density function, and expectations are taken conditionally on the observed covariates $\mathbf{X}$.

Conditional on the covariates $\mathbf{X}_{ij}$ and the unobserved frailties b_i, following Cox (1972), one may derive the partial likelihood score function for the regression coefficients $\boldsymbol{\beta}$ as follows:

$$\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_0^{\tau}\left\{\mathbf{X}_{ij} - \frac{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}, \mathbf{b})}{\mathcal{S}^{(0)}(t, \boldsymbol{\beta}, \mathbf{b})}\right\} dN_{ij}(t), \tag{2.2}$$

where $\tau < \infty$ is a constant, usually in practice the study duration, and $\mathcal{S}^{(l)}(t, \boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{X}_{ij}^{\otimes l} Y_{ij}(t)\exp(\mathbf{X}_{ij}'\boldsymbol{\beta} + b_i)$. Here, for a vector $\mathbf{u}$, $\mathbf{u}^{\otimes l} = \mathbf{u}\mathbf{u}'$ if $l = 2$, $\mathbf{u}^{\otimes l} = \mathbf{u}$ if $l = 1$ and $\mathbf{u}^{\otimes l} = 1$ if $l = 0$. Using a simple application of the Martingale theory, we show in the Appendix that the marginal expectation of $\mathbf{S}(\cdot)$ is 0, that is,

$$E\{\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})\} = 0, \tag{2.3}$$

which is an important finding as it guarantees the unbiasedness of the estimating equations developed in the following.

Let $F(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\eta})$ denote the conditional distribution of the frailties given observed survival times, covariates, and censoring indicators, where $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \theta)$ denotes the parameters necessary to specify the baseline hazard function and the unconditional distribution of frailties $\mathbf{b}$. The conditional expectation of the full data partial likelihood score for $\boldsymbol{\beta}$ hence is

$$\begin{aligned}\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta}) &= E\{\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\eta}\} \\ &= \int \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) dF(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\eta}),\end{aligned} \tag{2.4}$$

where F has a product form

$$F(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{i=1}^{M} F(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta}).$$

It should be noted that, in our model formulation, a parametric form has been assumed on the baseline hazard for the purpose of fully specifying the conditional distribution F. For instance, in our later simulations, we will be considering a flexible form, that is, a Weibull model:

$$\lambda_0(t, \boldsymbol{\alpha}) = \lambda p(\lambda t)^{p-1}, \tag{2.5}$$

where $\boldsymbol{\alpha} = (\lambda, p)$. It follows that

$$F(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{L(\mathbf{T}_i, \boldsymbol{\Delta}_i|\mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta})F(b_i; \theta)}{L(\mathbf{T}_i, \boldsymbol{\Delta}_i|\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta})},$$

where $L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta})$ is the conditional likelihood for the ith cluster given the frailty b_i under (2.1) and (2.5), that is,

$$L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{j=1}^{n_i} \left\{ \lambda_0(T_{ij}, \boldsymbol{\alpha}) e^{\mathbf{X}'_{ij}\boldsymbol{\beta} + b_i} \right\}^{\delta_{ij}} e^{-\Lambda_0(T_{ij}, \boldsymbol{\alpha}) \exp(\mathbf{X}'_{ij}\boldsymbol{\beta} + b_i)}, \quad (2.6)$$

and $L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta})$ is the marginal likelihood for the ith cluster obtained by integrating b_i out in (2.6) over its unconditional distribution. That is,

$$L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta}) = \int L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta}) dF(b_i; \theta). \quad (2.7)$$

A similar procedure was used by Louis (1982) for finding maximum likelihood estimates from incomplete data using the EM algorithm.

By iterated expectations, $E\{\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta})\} = E\{\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})\} = 0$. The form of (2.4) motivates us to regard frailties as "missing" constants, to impute them by the conditional distribution and to use the imputed values to construct an unbiased estimating equation. A similar observation was made by Satten, Datta, and Williamson (1998), who imputed unobserved failure times in the context of independent interval-censored survival data.

We propose to estimate the regression coefficient $\boldsymbol{\beta}$ from the imputed frailty partial likelihood score equation given the estimate of $\boldsymbol{\eta}$, containing the variance component θ and the unknown parameters $\boldsymbol{\alpha}$ associated with the baseline hazard. As the full data partial likelihood score $S(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})$ is independent of the baseline distribution, we may expect that (2.4) is less sensitive to the misspecification of the parametric form of the baseline hazard than the full likelihood score obtained by differentiating (2.7) with respect to $\boldsymbol{\beta}$.

To estimate $\boldsymbol{\eta}$, we resort to full likelihood maximization. Specifically, given $\boldsymbol{\beta}$, we estimate $\boldsymbol{\eta}$ by solving the full data log-likelihood score,

$$\mathbb{U}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^{M} \mathbf{U}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^{M} \frac{\partial}{\partial \boldsymbol{\eta}} \log \left\{ L(\mathbf{T}_i, \boldsymbol{\Delta}_i | \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta}) \right\}, \quad (2.8)$$

where $L(\cdot)$ is defined in (2.7).

The scheme above is equivalent to simultaneously solving for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in

$$\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta}) = 0, \quad (2.9)$$

$$\mathbb{U}(\boldsymbol{\beta}, \boldsymbol{\eta}) = 0. \quad (2.10)$$

In Section 4, two iterative algorithms are presented to solve these equations.

3. ASYMPTOTIC RESULTS

Under model (2.1) with $\lambda_0(t; \boldsymbol{\alpha})$ correctly specified and by the usual properties of maximum likelihood estimation, the estimating equations (2.9) and (2.10) are unbiased and, hence, would be expected to yield consistent estimates. If $\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta})$ were a sum of independent terms for each cluster, standard approaches would be applied to obtain the asymptotic results.

However, since (2.9) is not derived from an ordinary likelihood function and is not a sum of independent items, an application of the decomposition of the partial likelihood score will be needed to establish the asymptotic results. The asymptotic framework is as cluster number M goes to infinity with cluster sizes, n_i, uniformly bounded.

Denote by $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ the true value of the parameter vector, $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\eta})$, and suppose that $\boldsymbol{\gamma}_0$ is contained in a compact subset of R^{r+q}, say, $\mathcal{B}$, where q is the dimension of $\boldsymbol{\eta}$. For simplicity in the following theoretical development, we assume a constant cluster size, that is, $n_i \equiv n < \infty$. With the assumption that $(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i)$ are iid, we show in the Appendix that one may represent the full data partial likelihood score $\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}_0)$, up to $o_p(1)$, as an iid sum. That is,

$$M^{-1/2}\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}_0) = M^{-1/2} \sum_{i=1}^{M} \boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + o_p(1), \tag{3.1}$$

where

$$\begin{aligned} \boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta}) \quad = \quad & \sum_{j=1}^{n} \left[\delta_{ij} \left\{ \mathbf{X}_{ij} - \frac{s^{(1)}(T_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta})}{s^{(0)}(T_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta})} \right\} \right. \\ & \left. - e^{\mathbf{X}'_{ij}\boldsymbol{\beta} + b_i} \int_0^{T_{ij}} \left\{ \mathbf{X}_{ij} - \frac{s^{(1)}(t, \boldsymbol{\beta}, \boldsymbol{\eta})}{s^{(0)}(t, \boldsymbol{\beta}, \boldsymbol{\eta})} \right\} \frac{dG(t)}{s^{(0)}(t, \boldsymbol{\beta}, \boldsymbol{\eta})} \right]. \end{aligned}$$

Here, $G(t) = E\{N_{ij}(t)\}$, and $s^{(l)}(t, \boldsymbol{\beta}, \boldsymbol{\eta}) = E\{\mathcal{S}^{(l)}(t, \boldsymbol{\beta}, \mathbf{b}); \boldsymbol{\beta}, \boldsymbol{\eta}\}$ for $l = 0, 1, 2$. We further show in the Appendix that, in a small neighborhood of $\boldsymbol{\gamma}_0$, the termwise integration of (3.1) is allowable, enabling us to write

$$M^{-1/2}\mathbb{S}(\boldsymbol{\gamma}) = M^{-1/2} \sum_{i=1}^{M} \boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}) + o_p(1), \tag{3.2}$$

where $\boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}) = \int \boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta}_0) dF(b_i | \mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\eta})$. Hence, one is able to approximate the average partial likelihood score with respect to the conditional distribution of frailties using a sum of iid random variables.

Under model (2.1) with the baseline hazard $\lambda_0(t; \boldsymbol{\alpha})$ correctly specified and under the regularity conditions (C.1)–(C.3) listed in the Appendix, we can establish the consistency of the solution to the proposed estimating equations.

Theorem 1. *Under assumptions (C.1)–(C.3), there exists a sequence of solutions* $\hat{\boldsymbol{\gamma}}$ *to (2.9) and (2.10) such that for any given* $\epsilon > 0$, *there exists a* $K < \infty$ *and an integer* $M_0 > 0$ *such that* $\Pr\{\hat{\boldsymbol{\gamma}} \in \mathcal{N}_{K/\sqrt{M}}(\boldsymbol{\gamma}_0)\} \geq 1 - \epsilon$ *for any* $M \geq M_0$, *where* $\mathcal{N}_\rho(\boldsymbol{\gamma}_0)$ *is the neighborhood around* $\boldsymbol{\gamma}_0$ *with radius* ρ.

Using (3.2), it can be shown readily that

$$M^{-1/2}\{\mathbb{S}(\boldsymbol{\gamma}_0), \mathbb{U}(\boldsymbol{\gamma}_0)\} \xrightarrow{d} N(0, \boldsymbol{\Psi}), \tag{3.3}$$

where $\boldsymbol{\Psi}$, the covariance matrix of $\boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)$ and $\mathbf{U}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)$, is defined below. Asymptotic of the solution to (2.9) and (2.10) follows in part from (3.3) and is summarized in the following theorem.

Theorem 2. *Let $\hat{\gamma}$ be a consistent solution to (2.9) and (2.10). Then under (C.1)–(C.3),*

$$M^{1/2}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, \mathbf{V}),$$

where $\mathbf{V} = \mathbf{A}^{-1}\boldsymbol{\Psi}(\mathbf{A}^{-1})^T$, *with* $\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}'_{12} & \boldsymbol{\Psi}_{22} \end{pmatrix}$. *Here,* $\mathbf{A}$ *is expectation of the Jacobian matrix of the score Equations (2.9) and (2.10) and is defined in the Appendix, and*

$$\begin{aligned} \boldsymbol{\Psi}_{11} &= E\{\boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)\boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)'\}, \\ \boldsymbol{\Psi}_{12} &= E\{\boldsymbol{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)\mathbf{U}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)'\}, \\ \boldsymbol{\Psi}_{22} &= E\{\mathbf{U}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)\mathbf{U}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}_0)'\}. \end{aligned}$$

Both Theorems 1 and 2 are proved in the Appendix, along the line of Datta, Satten, and Williamson (2000).

4. ALGORITHMS FOR SOLVING THE ESTIMATING EQUATIONS

Due to the intractability of the integrals, Equations (2.9) and (2.10) are difficult to solve using standard numerical methods. We describe below two types of Monte Carlo algorithms to solve the score equations: the S-U Algorithm (Satten and Datta 2000) and a new approach we call the Monte Carlo estimating equation (MCEE) method.

4.1 S-U ALGORITHM

The S-U algorithm is a technique for finding the solution of an estimating equation that can be expressed as the expected value of a full data estimating equation, where the expectation is taken with respect to the missing data, given the observed data. This algorithm alternates between two steps: a simulation step wherein the missing values are simulated based on the conditional distributions given the observed data, and an updating step wherein parameters are updated without performing a numerical maximization. An attractive feature of this approach is that it is sequential, that is, the number of Monte Carlo replicates does not have to be specified in advance, and the values of previous Monte Carlo replicates do not have to be stored or regenerated for later use. In the following, we will apply this approach to solve (2.9) and (2.10).

Notice that $\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta})$ can be written

$$\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}; \boldsymbol{\gamma})} \int \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) \frac{L(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{b}|\mathbf{X}; \boldsymbol{\gamma})}{f(\mathbf{b}; \theta)} dF(\mathbf{b}; \theta), \tag{4.1}$$

where $L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}; \boldsymbol{\gamma})$ is the marginal likelihood of the observed dataset, which is the product of (2.7) over all clusters and $L(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{b}|\mathbf{X}; \boldsymbol{\gamma})$ is the joint likelihood of the observed survival times, the censoring indicators and the frailties. Hence, one can apply the importance sampling scheme (Tanner and Wong 1987) to approximate $\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta})$ and $\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta})$ and proceed as follows.

Having obtained approximants $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_j$ to $\hat{\boldsymbol{\gamma}}$, at the jth S-step of the algorithm, we simulate $\mathbf{b}^{(j,l)}, l = 1, \ldots, m$, independently from $f(\mathbf{b}; \hat{\theta}_j)$. Denote $w^{(j,l)}$ by

$$w^{(j,l)} = \frac{L(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{b}^{(j,l)}|\mathbf{X}; \hat{\boldsymbol{\gamma}}_j)}{f(\mathbf{b}^{(j,l)}; \hat{\theta}_j)} = L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}, \mathbf{b}^{(j,l)}; \hat{\boldsymbol{\gamma}}_j),$$

and let

$$\bar{w}_j = \frac{1}{j \cdot m} \sum_{j'=1}^{j} \sum_{l=1}^{m} w^{(j',l)}.$$

As $j \to \infty$, the Law of Large Numbers gives that $\bar{w}_j \xrightarrow{p} L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}; \hat{\boldsymbol{\gamma}})$ provided that $\hat{\boldsymbol{\gamma}}_j \xrightarrow{p} \hat{\boldsymbol{\gamma}}$.

We then write

$$\begin{aligned}
\bar{\mathbf{S}}_j &= \frac{1}{j \cdot m \cdot \bar{w}_j} \sum_{j'=1}^{j} w^{(j',l)} \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(j,l)}; \hat{\boldsymbol{\beta}}_j), \\
\bar{\mathbf{S}}_{\boldsymbol{\beta},j} &= \frac{1}{j \cdot m \cdot \bar{w}_j} \sum_{j'=1}^{j} w^{(j',l)} \Big\{ \mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) \\
&\quad + \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) \mathbf{U}_{b,\boldsymbol{\beta}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(j',l)}; \hat{\boldsymbol{\gamma}}_j) \Big\},
\end{aligned}$$

where

$$\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \int \left\{ \mathbf{X}_{ij} - \frac{\mathcal{S}^{(1)}(t, \beta, b)}{\mathcal{S}^{(0)}(t, \beta, b)} \right\}^2 dN_{ij}(t).$$

Notice that $\mathcal{I}(\cdot)$ is the information for the full dataset with $\mathbf{b}$ given (Fleming and Harrington 1991). With j sufficiently large, $\bar{\mathbf{S}}_j$ and $\bar{\mathbf{S}}_{\boldsymbol{\beta},j}$ will be good estimates for $\mathbb{S}(\hat{\boldsymbol{\gamma}}_j)$ and $\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{S}(\hat{\boldsymbol{\gamma}}_j)$.

Also note that, since (2.7) is a one-dimensional integral, one can conveniently apply Gauss–Hermite Quadrature to approximate this integral so as to obtain closed-form approximations, say, $\hat{\mathbb{U}}(\boldsymbol{\gamma})$ and $\hat{\mathbb{U}}_{\boldsymbol{\eta}}(\boldsymbol{\gamma})$, to $\mathbb{U}(\boldsymbol{\gamma})$, defined in (2.8), and its derivative with respect to $\boldsymbol{\eta}$, $\frac{\partial}{\partial \boldsymbol{\eta}}\mathbb{U}(\boldsymbol{\gamma})$. Finally, define

$$\bar{\boldsymbol{\gamma}}_j = \frac{1}{j} \sum_{j'=1}^{j} \hat{\boldsymbol{\gamma}}_{j'},\ \bar{\mathbf{U}}_j = \frac{1}{j} \sum_{j'=1}^{j} \hat{\mathbb{U}}(\boldsymbol{\gamma}_{j'}),\ \bar{\mathbf{U}}_{\boldsymbol{\eta},j} = \frac{1}{j} \sum_{j'=1}^{j} \hat{\mathbb{U}}_{\boldsymbol{\eta}}(\boldsymbol{\gamma}_{j'}).$$

Then at the jth U-step, the updated value for $\hat{\boldsymbol{\gamma}}$ is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{j+1} &= \bar{\boldsymbol{\beta}}_j - \bar{\mathbf{S}}_{\boldsymbol{\beta},j}^{-1} \bar{\mathbf{S}}_j, \\
\hat{\boldsymbol{\eta}}_{j+1} &= \bar{\boldsymbol{\eta}}_j - \bar{\mathbf{U}}_{\boldsymbol{\eta},j}^{-1} \bar{\mathbf{U}}_j.
\end{aligned}$$

It might be worthwhile to point out that each of the quantities required at this step, such as $\bar{\mathbf{S}}_j$, $\bar{\mathbf{S}}_{\boldsymbol{\beta},j}$, and so on, can be calculated recursively so that the past values of these intermediate variables never need to be stored.

Following Satten and Datta (2000), as $j \to \infty$, $\hat{\boldsymbol{\gamma}}_j$ converges to $\hat{\boldsymbol{\gamma}}$ almost surely. The total sampling variance of $\hat{\boldsymbol{\gamma}}_s$ around $\boldsymbol{\gamma}_0$ is the sum of the variance of $\hat{\boldsymbol{\gamma}}_s$ around $\hat{\boldsymbol{\gamma}}$ due to the S-U algorithm and the sampling variance of $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$ (Satten 1996). In most cases, the S-U algorithm should be iterated until the former is negligible compared to the latter. In theory, the starting value for the S-U algorithm is arbitrary. However, a bad starting value might cause instability at the begining of this algorithm. Hence, in the next section, we introduce an algorithm that can generate a starting value sufficiently close to the true zero of the estimating equations.

4.2 MONTE CARLO ESTIMATING EQUATION (MCEE)

One may notice that (4.1) can also be written

$$\mathbb{S}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}; \gamma)} \int \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) \frac{L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}, \mathbf{b}; \gamma) f(\mathbf{b}; \theta)}{g(\mathbf{b})} dG(\mathbf{b}), \quad (4.2)$$

where $G(\mathbf{b})$ is a known distribution function and $g(\mathbf{b})$ its density function. Hence, we are able to exploit an importance sampling approach to approximate (4.2) by sampling from distribution $G(\cdot)$. The resulting estimating equations are termed Monte Carlo estimating equations. There are many numerical methods, such as bracketing and bisection, Brent's method, and so on, to solve such Monte Carlo estimating equations; for a detailed description of these methods, see Press, Teukolsky, Vetterling, and Flannery (1992). Below we describe a Newton–Raphson-type method.

The key idea behind this algorithm is that one needs only to generate a single set of random numbers at the beginning of the iteration procedure. With the current estimates at each iteration step, one updates the sampling weights for the approximation of the average partial likelihood score over the the conditional distribution of frailties using importance sampling. More precisely, the procedure goes as follows:

1. Choose an initial iteration point $\hat{\boldsymbol{\gamma}}_0$. In practice, one can maximize the marginal likelihood for the observed data, which is the product of (2.7) over all clusters, to obtain $\hat{\boldsymbol{\gamma}}_0$.
2. Independently generate m simulations, $\mathbf{b}^{(0,l)}, l = 1, \ldots, m$, from a candidate density $g(\mathbf{b})$, which should have the same support as $f(\mathbf{b}; \theta)$.
3. With $\hat{\boldsymbol{\gamma}}_j$ obtained at the jth step, calculate the weights $w^{(j,l)} = L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}, \mathbf{b}^{(0,l)}; \hat{\boldsymbol{\gamma}}_j) f(\mathbf{b}^{(0,l)}; \hat{\theta}_j)/g(\mathbf{b}^{(0,l)})$ and $\bar{w}_j = \sum_{l=1}^m w^{(j,l)}$.
4. Estimate $\mathbb{S}(\hat{\boldsymbol{\gamma}}_j)$ and $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{S}(\hat{\boldsymbol{\gamma}}_j)$ using importance sampling by

$$\tilde{\mathbf{S}}_j = \frac{1}{\bar{w}_j} \sum_{l=1}^{m} w^{(j,l)} \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(0,l)}; \boldsymbol{\beta}).$$

and

$$\tilde{\mathbf{S}}_{\boldsymbol{\beta},j} = \frac{1}{\bar{w}_j} \sum_{l=1}^{m} w^{(j,l)} \Big\{ \mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(0,l)}; \hat{\boldsymbol{\beta}}_j) + \mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(0,l)}; \hat{\boldsymbol{\beta}}_j) \mathbf{U}_{b,\boldsymbol{\beta}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(0,l)}; \hat{\boldsymbol{\gamma}}_j) \Big\}.$$

5. As in the S-U algorithm, acquire the Gauss-Hermite Quadrature approximations $\hat{\mathbb{U}}(\boldsymbol{\gamma})$ and $\hat{\mathbb{U}}_{\boldsymbol{\eta}}(\boldsymbol{\gamma})$ and evaluate them at $\hat{\boldsymbol{\gamma}}_j$ to obtain $\hat{\mathbf{U}}_j$ and $\hat{\mathbf{U}}_{\boldsymbol{\eta},j}$.
6. Perform a Newton–Raphson updating step

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{j+1} &= \hat{\boldsymbol{\beta}}_j - \tilde{\mathbf{S}}_{\boldsymbol{\beta},j}^{-1} \tilde{\mathbf{S}}_j, \\ \hat{\boldsymbol{\eta}}_{j+1} &= \hat{\boldsymbol{\eta}}_j - \hat{\mathbf{U}}_{\boldsymbol{\eta},j}^{-1} \hat{\mathbf{U}}_j. \end{aligned}$$

7. Repeat Steps 3–6 until $||\hat{\boldsymbol{\gamma}}_{j+1} - \hat{\boldsymbol{\gamma}}_j|| < \epsilon$, where ϵ is a prespecified positive number.

Under the regularity conditions (Assumptions 1–4) listed by Geweke (1989), $\hat{\boldsymbol{\gamma}}_j$ can be guaranteed to converge to $\hat{\boldsymbol{\gamma}}^*$, which is sufficiently close to $\hat{\boldsymbol{\gamma}}$, the true zero of (2.9) and (2.10). These results are summarized in the following theorem proved in the Appendix.

Theorem 3. *Given a fixed M and m, as $j \to \infty$, $\hat{\boldsymbol{\gamma}}_j$ converges to $\hat{\boldsymbol{\gamma}}^*$ when the initial iteration value $||\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}}^*|| < \epsilon$, where ϵ is a sufficiently small number. Under the regularity conditions listed by Geweke (1989), $\hat{\boldsymbol{\gamma}}^* \to \hat{\boldsymbol{\gamma}}$ almost surely. Moreover,*

$$\sqrt{m}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \tilde{\mathbf{A}}_M^{-1}(\hat{\boldsymbol{\gamma}}) \mathcal{V} (\tilde{\mathbf{A}}_M^{-1})^T(\hat{\boldsymbol{\gamma}})$, $\tilde{\mathbf{A}}_M$ and $\mathcal{V}$ are given in the Appendix.

It follows from this theorem that, as in the S-U algorithm, the total sampling variance of the MCEE estimator $\hat{\boldsymbol{\gamma}}^*$ around $\boldsymbol{\gamma}_0$ is the sum of the variance of $\hat{\boldsymbol{\gamma}}^*$ around $\hat{\boldsymbol{\gamma}}$ due to the importance sampling iteration and the sampling variance of $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$. With the number of Monte Carlo simulations, m, sufficiently large, the variation due to the importance sampling can be made arbitrarily small.

In practice, one may be able to combine the S-U and MCEE algorithms: starting with the MCEE algorithm, one obtains an estimate after a small number of iterations and, thereafter, uses it as the starting value for the S-U scheme described earlier. We will numerically examine this hybrid algorithm in the simulation study.

5. VARIANCE ESTIMATORS

Conventionally, the variances of the maximum likelihood estimates are calculated by inverting the Fisher information matrix. However, since (2.9) is not an ordinary likelihood score, a further analysis is needed to derive an estimate of the variance matrix of $\hat{\boldsymbol{\gamma}}$, the zero of (2.9) and (2.10).

As shown in Lemma 2 in the Appendix, $\mathbf{A}$ can be consistently estimated by $\mathbf{A}_M(\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is the solution to (2.9) and (2.10) and $\mathbf{A}_M(\boldsymbol{\gamma})$ is the Jacobian matrix of the score

Equations (2.9) and (2.10). It is given by

$$\mathbf{A}_M(\boldsymbol\beta,\boldsymbol\eta) = \begin{pmatrix} \mathbf{A}_{11}^M(\boldsymbol\beta,\boldsymbol\eta) & \mathbf{A}_{12}^M(\boldsymbol\beta,\boldsymbol\eta) \\ \mathbf{A}_{21}^M(\boldsymbol\beta,\boldsymbol\eta) & \mathbf{A}_{22}^M(\boldsymbol\beta,\boldsymbol\eta) \end{pmatrix} = -\frac{1}{M}\begin{pmatrix} \frac{\partial}{\partial\boldsymbol\beta}\mathbb{S}(\boldsymbol\beta,\boldsymbol\eta) & \frac{\partial}{\partial\boldsymbol\eta}\mathbb{S}(\boldsymbol\beta,\boldsymbol\eta) \\ \frac{\partial}{\partial\boldsymbol\beta}\mathbb{U}(\boldsymbol\beta,\boldsymbol\eta) & \frac{\partial}{\partial\boldsymbol\eta}\mathbb{U}(\boldsymbol\beta,\boldsymbol\eta) \end{pmatrix}, \tag{5.1}$$

where $\mathbf{A}_{21}^M$ and $\mathbf{A}_{22}^M$ are easily obtained by differentiating (2.8) with respect to $\boldsymbol\beta$ and $\boldsymbol\eta$, while $\mathbf{A}_{11}^M$ and $\mathbf{A}_{12}^M$ are given by

$$\mathbf{A}_{11}^M(\boldsymbol\beta,\boldsymbol\eta) = \frac{1}{M}\int \{\mathcal{I}(\mathbf{T},\boldsymbol\Delta,\mathbf{X},\mathbf{b};\boldsymbol\beta) \\ -\mathbf{S}(\mathbf{T},\boldsymbol\Delta,\mathbf{X},\mathbf{b};\boldsymbol\beta)\mathbf{U}_{b,\boldsymbol\beta}(\mathbf{T},\boldsymbol\Delta,\mathbf{X},\mathbf{b};\boldsymbol\beta,\boldsymbol\eta)\}\,dF(\mathbf{b}|\mathbf{T},\boldsymbol\Delta,\mathbf{X};\boldsymbol\beta,\boldsymbol\eta)$$

and

$$\mathbf{A}_{12}^M(\boldsymbol\beta,\boldsymbol\eta) = -\frac{1}{M}\int \mathbf{S}(\mathbf{T},\boldsymbol\Delta,\mathbf{X},\mathbf{b};\boldsymbol\beta)\mathbf{U}_{b,\boldsymbol\eta}(\mathbf{T},\boldsymbol\Delta,\mathbf{X},\mathbf{b};\boldsymbol\beta,\boldsymbol\eta)dF(\mathbf{b}|\mathbf{T},\boldsymbol\Delta,\mathbf{X};\boldsymbol\beta,\boldsymbol\eta).$$

To develop a consistent estimator for $\boldsymbol\Psi$, we begin with $\boldsymbol\phi(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,b_i;\boldsymbol\gamma)$. For each $t<\infty$ and $l=0,1,2$, $s^{(l)}(t,\boldsymbol\beta,\boldsymbol\eta)$ can be consistently estimated by $\hat{s}^{(l)}(t,\boldsymbol\beta,\boldsymbol\eta)$, where

$$\hat{s}^{(l)}(t,\boldsymbol\beta,\boldsymbol\eta) = \frac{1}{M}\sum_{i=1}^{M}\int\left\{\sum_{j=1}^{n} e^{\mathbf{X}_{ij}'\boldsymbol\beta+b_i}Y_{ij}(t)\mathbf{X}_{ij}^{\otimes l}\right\}dF(b_i|\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i;\boldsymbol\beta,\boldsymbol\eta). \tag{5.2}$$

Let $g(s;\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,\boldsymbol\gamma)$ denote the moment generating function of b_i conditional on $\mathbf{T}_i,\boldsymbol\Delta_i$ and $\mathbf{X}_i$, that is, $g(s;\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,\boldsymbol\gamma) = E(e^{s\cdot b_i}|\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i;\boldsymbol\gamma)$. Hence, Equation (5.2) can be simply expressed as

$$\hat{s}^{(l)}(t,\boldsymbol\beta,\boldsymbol\eta) = \frac{1}{M}\sum_{i=1}^{M}\left\{\sum_{j=1}^{n} e^{\mathbf{X}_{ij}'\boldsymbol\beta}Y_{ij}(t)\mathbf{X}_{ij}^{\otimes l}\right\}g(1;\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,\boldsymbol\gamma).$$

It follows that, with the same argument in Lin and Wei (1989), $\boldsymbol\phi$ can be estimated by

$$\hat{\boldsymbol\phi}(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,b_i;\boldsymbol\beta,\boldsymbol\eta) = \sum_{j=1}^{n}\left[\delta_{ij}\left\{\mathbf{X}_{ij}-\frac{\hat{s}^{(1)}(T_{ij},\boldsymbol\beta,\boldsymbol\eta)}{\hat{s}^{(0)}(T_{ij},\boldsymbol\beta,\boldsymbol\eta)}\right\} \right. \\ \left. -\sum_{i',j'}\frac{N_{i'j'}(T_{ij})e^{X_{ij}'\boldsymbol\beta+b_i}}{M\cdot\hat{s}^{(0)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}\left\{\mathbf{X}_{ij}-\frac{\hat{s}^{(1)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}{\hat{s}^{(0)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}\right\}\right].$$

Note that $\hat{\phi}(\cdot)$ resembles the influence function for the full data proportional hazards model (Reid and Crépeau 1985).

Because each $\Psi(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i;\boldsymbol\gamma)$ is the expectation (with respect to b_i conditional on $\mathbf{T}_i,\boldsymbol\Delta_i$ and $\mathbf{X}_i$) of $\boldsymbol\phi(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,b_i;\boldsymbol\gamma)$, it can be consistently estimated by the conditional expectation of $\hat{\phi}(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,b_i;\boldsymbol\gamma)$, which is

$$\hat{\boldsymbol\Psi}(\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i;\boldsymbol\gamma) = \sum_{j=1}^{n}\delta_{ij}\left\{\mathbf{X}_{ij}-\frac{\hat{s}^{(1)}(T_{ij},\boldsymbol\beta,\boldsymbol\eta)}{\hat{s}^{(0)}(T_{ij},\boldsymbol\beta,\boldsymbol\eta)}\right\} \\ -g(1;\mathbf{T}_i,\boldsymbol\Delta_i,\mathbf{X}_i,\boldsymbol\gamma)\sum_{j=1}^{n}e^{\mathbf{X}_{ij}'\boldsymbol\beta}\left[\sum_{i',j'}\frac{N_{i'j'}(T_{ij})}{M\cdot\hat{s}^{(0)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}\left\{\mathbf{X}_{ij}-\frac{\hat{s}^{(1)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}{\hat{s}^{(0)}(T_{i'j'},\boldsymbol\beta,\boldsymbol\eta)}\right\}\right].$$

Because $\hat{\Psi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})$ is an estimate of the contribution of the ith cluster to the score $\mathbb{S}(\hat{\gamma})$, the matrix $\boldsymbol{\Psi}$ can be estimated by $\hat{\boldsymbol{\Psi}} = \begin{pmatrix} \hat{\boldsymbol{\Psi}}_{11} & \hat{\boldsymbol{\Psi}}_{12} \\ \hat{\boldsymbol{\Psi}}_{21} & \hat{\boldsymbol{\Psi}}_{22} \end{pmatrix}$, where

$$\begin{aligned}
\hat{\boldsymbol{\Psi}}_{11} &= \frac{1}{M}\sum_{i=1}^{M} \hat{\boldsymbol{\Psi}}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})\hat{\boldsymbol{\Psi}}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})', \\
\hat{\boldsymbol{\Psi}}_{12} &= \hat{\boldsymbol{\Psi}}_{21}' = \frac{1}{M}\sum_{i=1}^{M} \hat{\boldsymbol{\Psi}}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})\hat{U}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})', \\
\hat{\boldsymbol{\Psi}}_{22} &= \frac{1}{M}\sum_{i=1}^{M} \hat{\mathbf{U}}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})\hat{\mathbf{U}}_i(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \hat{\gamma})'.
\end{aligned}$$

Hence, the asymptotic variance of $M^{1/2}(\hat{\gamma} - \gamma_0)$, $\mathbf{V}$, can be estimated by

$$\hat{\mathbf{V}} = \mathbf{A}_M^{-1}(\hat{\gamma})\hat{\boldsymbol{\Psi}}(\mathbf{A}_M^{-1})^T(\hat{\gamma}). \tag{5.3}$$

A simple alternative to calculate the variance involves a bootstrap procedure on the basis of clusters (Efron 1979). Specifically, we resample M clusters, with replacement, from $(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i)|_{i=1}^{M}$ to obtain a new dataset $(\mathbf{T}_{(i)}, \boldsymbol{\Delta}_{(i)}, \mathbf{X}_{(i)})|_{i=1}^{M}$. Given this new dataset, we solve (2.9) and (2.10) for the estimates of $\boldsymbol{\beta}$ and θ. This procedure can be repeated for B times to obtain a sequence of estimates, $(\tilde{\boldsymbol{\beta}}^{(l)}, \tilde{\theta}^{(l)})$. The bootstrap variance estimates can hence be calculated using the sample variances

$$\text{var}_{\text{boot}}(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1}\sum_{l=1}^{B}(\tilde{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}}_{\text{boot}})(\tilde{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}}_{\text{boot}})',$$

and

$$\text{var}_{\text{boot}}(\hat{\theta}) = \frac{1}{B-1}\sum_{l=1}^{B}(\tilde{\theta}^{(l)} - \bar{\theta}_{\text{boot}})^2,$$

where $\bar{\boldsymbol{\beta}}_{\text{boot}} = \frac{1}{B}\sum_{l=1}^{B}\tilde{\boldsymbol{\beta}}^{(l)}$ and $\bar{\theta}_{\text{boot}} = \frac{1}{B}\sum_{l=1}^{B}\tilde{\theta}^{(l)}$. In practice, it is adequate to choose a moderate number of resamplings, B, say, in the range 25 to 100 (Lange 1999, p. 301). We chose $B = 40$ in our simulation studies. For a review of other nonparametric techniques for obtaining the variance estimates, such as the jackknife procedure, the smoothed bootstrap method and the half-sampling approach; see Efron (1981) and Efron and Tibshirani (1993).

6. SIMULATION

Simulations were performed to assess the finite sample performance of the proposed imputed frailty partial likelihood score (IFPLS) estimators, compared with the parametric likelihood MLEs and the nonparametric MLEs. The robustness and efficiency of the proposed estimators were of particular interest. The performance of the variance estimator (5.3) and the bootstrap variance estimator was also evaluated.

Table 1. Comparisons of IFPLS Estimates, Parametric MLEs and NPMLEs. The baseline hazard is correctly specified in the imputation and parametric MLE calculation. The true regression coefficients are : $\beta_x = 1$, $\theta = 0.25$. SE_e is the empirical standard error, SE_a is the standard error calculated by (5.3) and SE_b is the bootstrap standard error.

		IFPLS				*Parametric MLE*		*NPMLE*	
Censoring		*Estimate*	*SE_e*	*SE_a*	*SE_b*	*Estimate*	*SE_e*	*Estimate*	*SE_e*
20%	β	1.017	0.348	0.330	0.349	1.012	0.318	1.022	0.424
	θ	0.263	0.198	0.204	0.217	0.256	0.185	0.246	0.230
40%	β	0.982	0.450	0.459	0.450	1.029	0.431	1.014	0.507
	θ	0.258	0.265	0.243	0.249	0.259	0.221	0.267	0.286
60%	β	1.019	0.559	0.563	0.551	1.013	0.523	1.046	0.643
	θ	0.253	0.272	0.282	0.289	0.278	0.252	0.268	0.312
80%	β	1.021	0.707	0.718	0.722	0.965	0.676	1.056	0.781
	θ	0.287	0.499	0.483	0.503	0.289	0.467	0.223	0.532

To calculate the IFPLS estimates, we used a combination of the S-U and MCEE algorithms: we obtained an estimate by the MCEE algorithm after a small number of iterations, and used it as the initial value for the S-U scheme. As stated earlier, the sole purpose of doing so was to guarantee the stability when applying the S-U algorithm.

In each simulated dataset, survival times V_{ij} were generated within each cluster by the conditional hazard $\lambda_{ij}(t) = \lambda_0(t)\exp(\beta X_{ij} + b_i), j = 1, \ldots, n_i, i = 1, \ldots, M$, where the X_{ij} were generated from random uniform $U[0, 1]$. Censoring times C_{ij} were simulated from uniform $U[0, c]$. The frailties were generated according to $N(0, \theta)$.

We considered the following combinations of experiments: the cluster number M was set to be 40, while the cluster sizes n_i were distributed according to the discrete uniform distribution with masses on the integers 1 to 5; $\beta = 1$ and $\theta = 0.25$; c was chosen to yield four different censoring proportions (20%, 40%, 60%, and 80%). We chose different models for the baseline hazard $\lambda_0(t)$ depending on the purpose of simulations, that is, comparison of efficiency or study of robustness. For each parameter configuration, a total of 400 replicate datasets were made.

We first studied the finite sample performance of the IFPLS estimators and compare them with the parametric MLEs and the NPMLEs in terms of efficiency. Specifically, survival times were generated with the Weibull baseline hazard (2.5), where the shape parameter p was set to 2 and the scale λ to 0.25. In the calculation of the IFPLS estimators and the parametric MLEs, the baseline hazard was correctly specified as the Weibull model (2.5). To obtain an initial value for the S-U algorithm, we first ran two iterations of the MCEE algorithm, where the number of Monte Carlo simulations was set to 3,000. At each simulation step of the S-U algorithm, the number of Monte Carlo simulations was set to 1,000. In our experience, usually convergence was achieved within about 10 steps. We also computed the variance estimators using (5.3) and the bootstrap procedure for the IFPLS estimates, The results are displayed in Table 1. The IFPLS method gave highly efficient estimates compared with the full likelihood MLEs. For example, with a censoring proportion 20%, the asymptotic relative efficiency (ARE) for β and θ calculated by the IFPLS method are 92.8% and 92.3%. In addition, the estimated standard errors using (5.3) and the bootstrap agree well with the empirical standard errors.

Table 2. Comparisons of IFPLS Estimates, Parametric MLEs and NPMLEs. The baseline hazard was misspecified in the IFPLS and parametric MLE calculation. The true regression coefficients are : $\beta_x = 1$, $\theta = 0.25$. SE_e is the empirical standard error, SE_a is the standard error calculated by (5.3) and SE_b is the bootstrap standard error.

		IFPLS				*Parametric MLE*		*NPMLE*	
Censoring		*Estimate*	*SE_e*	*SE_a*	*SE_b*	*Estimate*	*SE_e*	*Estimate*	*SE_e*
20%	β	0.951	0.313	0.319	0.323	0.887	0.301	1.045	0.393
	θ	0.238	0.192	0.201	0.207	0.203	0.163	0.256	0.257
40%	β	0.931	0.415	0.421	0.408	0.873	0.464	1.038	0.567
	θ	0.239	0.244	0.224	0.238	0.187	0.153	0.246	0.301
60%	β	0.943	0.526	0.513	0.538	0.861	0.553	1.034	0.641
	θ	0.217	0.252	0.242	0.267	0.168	0.179	0.253	0.331
80%	β	0.931	0.672	0.703	0.686	0.854	0.702	0.951	0.842
	θ	0.192	0.479	0.474	0.462	0.160	0.399	0.271	0.485

We next examined the robustness of the proposed IFPLS estimators, compared with the parametric MLEs and the NPMLEs when the baseline hazard was misspecified. Specifically, the true baseline hazard was $\lambda_0(t) = pe^{-pt}$ with $p = 2$, while in the calculation of the IFPLS estimators and the parametric MLEs, the baseline hazard was incorrectly specified as the Weibull model (2.5). The results are shown in Table 2. It appears that the IFPLS estimates are only slightly biased. For instance, with a censoring proportion 20%, compared with the true values, the relative biases (RBs) for β and θ calculated by the IFPLS method are 5.5% and 7.2%, compared with the corresponding RBs of 11.2% and 16.6% in the full likelihood MLEs. The NPMLEs gave consistent estimates but with large standard errors. With the current configuration of parameters, the estimated standard errors using (5.3) and the bootstrap agree well with the empirical standard errors.

7. AN APPLICATION: EAST BOSTON ASTHMA STUDY

We applied the proposed method to the analysis of the East Boston Asthma Study, which was conducted by Rosalind Wright at the Channing Laboratory, Harvard Medical School, to understand etiologies of rising prevalence and morbidity of childhood asthma, and of the disproportionate burden among urban minority children. For our analysis, we focus on an assessment of the role of a familial history of asthma may attribute to disparities in disease burden. This has been largely an unexplored aspect to explain children's asthma prevalence. The investigator, in particular, was interested in the relationship between the maternal asthma status (with a variable name "MEVAST", coded as 1 = EVER HAD ASTHMA and 0 = NEVER HAD ASTHMA) and children's asthma status, controlled for the effects of race (0 = WHITE and 1 = NON-WHITE) and gender (0 = MALE and 1 = FEMALE).

Subjects were enrolled at community health clinics throughout east Boston and questionnaire data was collected during regularly scheduled well-baby visits. In addition to basic demographic data, residential addresses were recorded and geocoded for each study subject. Geocoding the dataset allowed one to link various community-level (cluster) covariates to

Table 3. Analysis Results of the Boston Asthma Study Data. Estimates were calculated by the Naive (ignoring within-cluster correlation), NPMLE, Fully Parametric MLE and IFPLS algorithms. Numbers inside the parentheses are standard errors (SEs). The SEs of the IFPLS estimates were obtained using the closed-form variance estimators, the SEs of the parametric estimates were obtained by inverting the minus second-order derivative of the log-likelihood and the SEs of the NPMLEs were acquired by inverting the negative second derivative of the log profile likelihood.

Parameter	*Naive*	*IFPLS*	*Parametric*	*NPMLE*
θ	—	0.032 (0.07)	0.034(0.05)	0.029 (0.10)
MEVASTH	0.706 (0.28)	0.652(0.30)	0.664(0.29)	0.647(0.35)
RACE	0.173 (0.21)	0.205 (0.23)	0.212 (0.21)	0.199 (0.26)
GENDER	−0.135 (0.20)	−0.160 (0.22)	−0.165 (0.21)	−0.159 (0.26)

individuals in the east Boston dataset from U.S. Census data at both the census tract and census block-group level (clusters). In the analysis, the census block-groups were treated as independent clusters.

We fitted a random intercept frailty model to the data

$$\lambda_{ij}(t|b_i) = \lambda_0(t)\exp\{\beta_M \times \text{MEVAST}_{ij} + \beta_R \times \text{RACE}_{ij} + \beta_G \times \text{GENDER}_{ij} + b_i\}, \quad (7.1)$$

where the frailty b_i follows $N(0, \theta)$. Here, the subscripts i and j indicate the cluster (census block) level and the individual level, respectively.

We applied the proposed IFPLS method, with the baseline hazard taking the Weibull model (2.5), and calculated the estimates of unknown parameters using a combination of the S-U and MCEE algorithms. For comparison purposes, we also fitted a naive model, which ignores the within cluster correlation, that is, assume that $\theta = 0$ in (7.1). In addition, the NPMLEs and fully parametric MLEs were also calculated. The results are presented in Table 3. As demonstrated by the IFPLS method ($p = 0.030$) and the fully parametric MLE ($p = 0.026$), a higher risk of asthma was significantly associated with a history of maternal asthma, after controlling for the effects of race and gender. However, the NPMLE yielded a nonsignificant p value of 0.064. Ignoring within-cluster correlation inflated the estimate of the regression coefficient. Note that the standard error of the estimate of θ in Table 3 cannot be directly used to test for $\text{H}_0 : \theta = 0$, since the null hypothesis is on the boundary of the parameter space and the Wald statistic is not asymptotically distributed as a chi-square (Lin 1997). An appropriate score test for detecting the heterogeneity across clusters was developed by Gray (1995), who considered institutional variations in a multi-center cancer clinical trial.

8. DISCUSSION

This article extends the Cox partial likelihood approach to fit frailty models for clustered survival data. The frailties are treated as unobserved covariates and are imputed based on the conditional distributions. For this purpose, a parametric form is assumed on the baseline hazard. We estimate the regression coefficients by solving the average partial likelihood score (PLS) equations. Simulations indicate high efficiency of the resulting estimates. Despite the dependence of our proposed method on the parametric structure of the baseline hazard, the estimation of regression coefficients is only slightly affected when the baseline hazard is incorrectly specified.

We have also proposed a closed-form "sandwich" variance estimator, along with a bootstrap variance estimator, for the IFPLS estimates. Both are easy to calculate and simulations indicate that they agree well with the empirical variance estimates.

A Weibull structure was assumed on the baseline hazard in our numerical experiments. In practice, this assumption can be made more flexible to adjust to the given data. For instance, the baseline hazard can be modeled by smooth splines with knots at fixed finite time points. Explicitly, we may consider a linear spline model:

$$\log \lambda_0(t, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 t + \sum_{j=1}^{L} \alpha_{1j}(t - \xi_j)_+,$$

where z_+ denotes the positive part of z, $0 < \xi_1 < \cdots < \xi_L < \infty$ are given knots for a fixed L and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_{11}, \ldots, \alpha_{1L})$ are unknown coefficients. One may choose the knots by examining the smoothed naive hazard curve (Ramlau-Hansen 1983).

We expect that this proposed IFPLS methodology shall have broad applications. A first straightforward application would be on the multivariate random effects survival models with b_i in (2.1) replaced by $\mathbf{B}_{ij}'\mathbf{b}_i$, where $\mathbf{B}_{ij}(a \times 1)$ is the known covariate vector associated with the frailty $\mathbf{b}_i$ and the $\mathbf{b}_i$ are iid with a distribution function $F(\cdot;\boldsymbol{\theta})$, depending on an unknown length-a variance component $\boldsymbol{\theta}$. This approach can also be easily explored to model clustered interval-censored survival data, survival data with measurement error in covariates, spatial survival data, where the unobserved random quantities, such as true survival times, true covariates and region-specific random effects, can be imputed simultaneously from the conditional distributions to construct unbiased average partial likelihood estimating equations. We will report on these in subsequent articles.

APPENDIXES: TECHNICAL DETAILS

A. NOTATION AND REGULARITY CONDITIONS

For the ith cluster, we introduce the score function with respect to the conditional density of the frailties as follows

$$\begin{aligned}
\mathbf{U}_{b_i,\boldsymbol{\gamma}}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}) &= \frac{\partial}{\partial \boldsymbol{\gamma}} \log\{f(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \gamma)\} \\
&= \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log\{f(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \gamma)\}, \right. \\
&\qquad \left. \frac{\partial}{\partial \boldsymbol{\eta}} \log\{f(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \gamma)\}\right] \\
&= [\mathbf{U}_{b_i,\boldsymbol{\beta}}'(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}), \mathbf{U}_{b_i,\boldsymbol{\eta}}'(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma})].
\end{aligned} \tag{A.1}$$

Denote by $\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})$, $\mathbf{U}_{b,\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})$, and $\mathbf{U}_{b,\boldsymbol{\eta}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})$ the sum of the corresponding terms in (A.1) over all clusters. We further denote a ρ-neighborhood of $\boldsymbol{\gamma}$ by $\mathcal{N}_\rho(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma}' \in \mathcal{B} : ||\boldsymbol{\gamma}' - \boldsymbol{\gamma}|| < \rho\}$, where $||\cdot||$ denotes an Euclidean norm. With the notation introduced above and established in Sections 2 and 3, we stipulate the following regularity conditions:

(C.1) The sequence $(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i)$ is iid and $\mathbf{X}_i$ are bounded.

(C.2) The sequences $\{\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\gamma}}\}, \{\frac{\partial \mathbf{U}}{\partial \boldsymbol{\gamma}}\}, \{\boldsymbol{\phi}\mathbf{U}'_{b,\boldsymbol{\gamma}}\}, \{\frac{\partial \mathbf{U}_{b,\boldsymbol{\gamma}}}{\partial \boldsymbol{\gamma}}\}$ and $\{\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\beta}}\}$ each satisfy the uniform weak law of large numbers (UWLLN) conditions at $\boldsymbol{\gamma}_0$ as explained below.

(C.3) The expectation matrix, $\mathbf{A}$, of the Jacobian matrix of the score Equations (2.9) and (2.10), is invertible, where $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$, and

$$\begin{aligned}
\mathbf{A}_{11} &= \mathbf{Q}(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) - E\{\boldsymbol{\phi}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i,b_i;\boldsymbol{\gamma}_0)\mathbf{U}_{b_i,\boldsymbol{\beta}}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i,b_i;\boldsymbol{\gamma}_0)'\},\\
\mathbf{A}_{12} &= E\{\boldsymbol{\phi}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i,b_i;\boldsymbol{\gamma}_0)\mathbf{U}_{b_i,\boldsymbol{\eta}}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i,b_i;\boldsymbol{\gamma}_0)'\},\\
\mathbf{A}_{21} &= E\{\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{U}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i;\boldsymbol{\gamma}_0)'\},\\
\mathbf{A}_{22} &= E\{\frac{\partial}{\partial \boldsymbol{\eta}}\mathbf{U}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i;\boldsymbol{\gamma}_0)'\},\\
\mathbf{Q}(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) &= \int \left\{ \frac{s^{(2)}(t,\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)}{s^{(0)}(t,\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)} - \frac{s^{(1)}(t,\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)^{\otimes 2}}{s^{(0)}(t,\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)^2} \right\} dG(t),
\end{aligned}$$

and all the expectations involved are taken under the true parameter $\boldsymbol{\gamma}_0$.

Remark: An iid random sequence indexed by γ, say, $q_i(\gamma), i = 1, 2, \ldots$, is said to satisfy the UWLLN conditions at a fixed point $\boldsymbol{\gamma}_0$ if, for each i, (1) $E\{q_i(\boldsymbol{\gamma})\}$ is continuous at $\boldsymbol{\gamma}_0$, and, (2) there exists a $\delta > 0$ such that, $E\{\sup_{||\boldsymbol{\gamma}-\boldsymbol{\gamma}_0||<\delta} ||q_i(\boldsymbol{\gamma})||\} < \infty$, and (3)

$$\lim_{\rho \to 0+} \sup E\{Q_i(\boldsymbol{\gamma})\} = 0$$

for any $\gamma \in \mathcal{N}_\delta(\boldsymbol{\gamma}_0)$, where $\mathbf{Q}_i(\boldsymbol{\gamma}) = \sup\{||q_i(\boldsymbol{\gamma}_1) - q_i(\boldsymbol{\gamma})|| : \boldsymbol{\gamma}_1 \in \mathcal{N}_\rho(\boldsymbol{\gamma})$. Datta (1988) showed that the UWLLN conditions ensure the uniform convergence of $M^{-1}\sum_{i=1}^M q_i(\boldsymbol{\gamma})$ to $E\{q_1(\boldsymbol{\gamma})\}$ uniformly in $\mathcal{N}_{\delta/2}(\boldsymbol{\gamma}_0)$. Practically, the seemingly stringent Condition (3) can be substituted with a more familiar sufficient (Lipschitz) condition

$$||q_i(\boldsymbol{\gamma}_1) - q_i(\boldsymbol{\gamma}_2)|| < ||\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2|| G_i$$

for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathcal{N}_\delta(\boldsymbol{\gamma}_0)$, where $E(G_i) < \infty$. For more detailed discussion on the UWLLN conditions, see Satten, Datta, and Williamson (1998).

B. PROOFS

Proof of Equation (2.3): Following Fleming and Harrington (1991, p. 44), we rewrite (2.2)

$$\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_0^\tau \left\{\mathbf{X}_{ij} - \frac{\mathcal{S}^{(1)}(t,\boldsymbol{\beta},\mathbf{b})}{\mathcal{S}^{(0)}(t,\boldsymbol{\beta},\mathbf{b})}\right\} dM_{ij}(t),$$

where $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(s)\lambda_0(s)\exp(\mathbf{X}'_{ij}\boldsymbol{\beta} + b_i)ds$ is a martingale adapted to the filtration $\mathcal{F}_t^{\mathbf{X},\mathbf{b}} = \sigma\{\mathbf{N}(s),\mathbf{Y}(s),\mathbf{X},\mathbf{b},0 \le s \le t\}$, where $\sigma\{\cdot\}$ denotes a σ-algebra. Because the integrand in (2.2) is an $\mathcal{F}_t^{\mathbf{X},\mathbf{b}}$-predictable process, $\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})$ is a locally square integrable martingale adapted to $\mathcal{F}_t^{\mathbf{X},\mathbf{b}}$ (Fleming and Harrington 1991). Hence,

$$E\{\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})|\mathcal{F}_0^{\mathbf{X},\mathbf{b}}\} = E\{\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})|\mathbf{X},\mathbf{b}\} = 0,$$

and, moreover, by the conditional expectation theorem (Fleming and Harrington 1991, p. 22),

$$E\{\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})|\mathcal{F}_0^{\mathbf{X}}\} = E\{\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})|\mathbf{X}\} = 0$$

because $\mathcal{F}_0^{\mathbf{X}} \subset \mathcal{F}_0^{\mathbf{X},\mathbf{b}}$. Here, $\mathcal{F}_0^{\mathbf{X},\mathbf{b}} = \sigma\{\mathbf{X},\mathbf{b}\}$ and $\mathcal{F}_0^{\mathbf{X}} = \sigma\{\mathbf{X}\}$. □

Lemma 1. *As $M \to \infty$, for each $K > 0$, $M^{-1/2}\mathbb{S}(\boldsymbol{\gamma}) = M^{-1/2}\tilde{\mathbb{S}}(\boldsymbol{\gamma}) + o_p(1)$ uniformly in $\mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)$, where $\tilde{\mathbb{S}}(\boldsymbol{\gamma}) = \sum_{i=1}^M \boldsymbol{\Psi}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i;\boldsymbol{\gamma})$.*

Proof: Consider $\mathbf{R}_M(\boldsymbol{\beta}) = M^{-1/2}\{\mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta}) - \tilde{\mathbf{S}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta})\}$, where

$$\tilde{\mathbf{S}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\beta}) = \sum_{i=1}^{M}\boldsymbol{\phi}(\mathbf{T}_i,\boldsymbol{\Delta}_i,\mathbf{X}_i,b_i;\boldsymbol{\beta},\boldsymbol{\eta}_0).$$

Then

$$\begin{aligned} M^{-1/2}\{\mathbb{S}(\boldsymbol{\gamma}) - \tilde{\mathbb{S}}(\boldsymbol{\gamma})\} &= \int \mathbf{R}_M(\boldsymbol{\beta})dF(\mathbf{b}|\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\boldsymbol{\gamma}) \\ &= \int \mathbf{R}_M(\boldsymbol{\beta})\exp\{\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma}^*)' \\ &\quad \times(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\}dF(\mathbf{b}|\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\boldsymbol{\gamma}_0), \end{aligned}$$

where $\boldsymbol{\gamma}^*$ lies on the line segment connecting $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$. For an arbitrary function $H(x): R^{r+q} \to R$, denote its suprema in $\mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)$ by $\bigvee H(\boldsymbol{\gamma}) = \sup_{\boldsymbol{\gamma}\in\mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)} H(\boldsymbol{\gamma})$.

Then $\bigvee|M^{-1/2}\{\mathbb{S}(\boldsymbol{\gamma}) - \tilde{\mathbb{S}}(\boldsymbol{\gamma})\}| \le C_1C_2$, where

$$C_1 = \exp\left\{K\cdot\bigvee||M^{-1/2}\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})||\right\} \tag{B.1}$$

and $C_2 = E\{\bigvee||\mathbf{R}_M(\boldsymbol{\beta})||\,|\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\boldsymbol{\gamma}_0\}$.

Notice that

$$\begin{aligned} \bigvee||M^{-1/2}\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})|| &\le ||M^{-1/2}\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma}_0)|| \\ &\quad + K\cdot\bigvee||M^{-1}\frac{\partial}{\partial\boldsymbol{\gamma}}\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\boldsymbol{\gamma})||. \end{aligned}$$

The first term is $O_p(1)$ by applying the central limit theorem, while the second term on the right-hand side of the inequality above converges to $K||E\{\frac{\partial}{\partial \boldsymbol{\gamma}}\mathbf{U}_{b,\boldsymbol{\gamma}}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma})\}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}||$ by the UWLLN condition; hence $C_1 = O_p(1)$.

To estimate the magnitude of C_2, we consider its expected value, $E(C_2) = E(\bigvee ||\mathbf{R}_M(\boldsymbol{\beta})||)$. We observe that

$$\begin{aligned}
\bigvee ||R_M(\boldsymbol{\beta})|| &= \bigvee M^{-1/2}||\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) - \tilde{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})|| \\
&\quad + \bigvee M^{-1/2}||\tilde{\mathbf{S}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\eta}_0) - \tilde{\mathbf{S}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\eta})|| \\
&\leq ||\mathbf{R}_M(\boldsymbol{\beta}_0)|| + K \cdot \bigvee ||M^{-1}\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) - \mathbf{Q}(\boldsymbol{\gamma}_0)|| \\
&\quad + K \cdot \bigvee || - M^{-1}\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\beta}, \boldsymbol{\eta}_0) - \mathbf{Q}(\boldsymbol{\gamma}_0)||. \quad \text{(B.2)}
\end{aligned}$$

We first find a bound for $\mathbf{R}_M(\boldsymbol{\beta}_0)$. Applying the uniform laws of large numbers for the empirical processes (see Pollard 1990, p. 34), one may see

$$\sup_{t\in[0,\tau]} \left| \frac{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}_0, \mathbf{b})}{\mathcal{S}^{(0)}(t, \boldsymbol{\beta}_0, \mathbf{b})} - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)} \right| \xrightarrow{p} 0$$

as $M \to \infty$, where $s^{(l)}(t, \boldsymbol{\beta}, \boldsymbol{\eta}) = E\{\mathcal{S}^{(l)}(t, \boldsymbol{\beta}, \mathbf{b}); \boldsymbol{\beta}, \boldsymbol{\eta}\}$ for $l = 0, 1, 2$. Write $\bar{N}(t) = \frac{1}{Mn}\sum_{ij} N_{ij}(t)$ and $G(t) = E\{\bar{N}(t)\} = E\{N_{ij}(t)\}$. We may write $M^{-1/2}\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}_0)$ as

$$\begin{aligned}
&M^{-1/2}\sum_{i=1}^{M}\sum_{j=1}^{n}\int_0^\tau \mathbf{X}_{ij}dN_{ij}(t) - M^{1/2}n\int_0^\tau \frac{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}_0, \mathbf{b})}{\mathcal{S}^{(0)}(t, \boldsymbol{\beta}_0, \mathbf{b})}d\bar{N}(t) \\
&= M^{-1/2}\sum_{i=1}^{M}\sum_{j=1}^{n}\int_0^\tau \mathbf{X}_{ij}dN_{ij}(t) - M^{1/2}n\int_0^\tau \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}d\bar{N}(t) \\
&-M^{1/2}n\int_0^\tau \left\{\frac{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}_0, \mathbf{b})}{\mathcal{S}^{(0)}(t, \boldsymbol{\beta}_0, \mathbf{b})} - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\right\} dG(t) \\
&-M^{1/2}n\int_0^\tau \left\{\frac{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}_0, \mathbf{b})}{\mathcal{S}^{(0)}(t, \boldsymbol{\beta}_0, \mathbf{b})} - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\right\} d\{\bar{N}(t) - G(t)\}. \quad \text{(B.3)}
\end{aligned}$$

Because $M^{1/2}\{\bar{N}(t) - G(t)\}$ converges to a zero mean Gaussian process, the last term in (B.3) is $o_p(1)$. It can be shown the third term in (B.3) is equal to

$$M^{1/2}n\int_0^\tau \frac{1}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\left\{\mathcal{S}^{(1)}(t, \boldsymbol{\beta}_0, \mathbf{b}) - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\mathcal{S}^{(0)}(t, \boldsymbol{\beta}_0, \mathbf{b})\right\} dG(t) + o_p(1).$$

Then

$$\begin{aligned}
M^{-1/2}\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}_0) &= M^{-1/2}\sum_{i=1}^{M}\sum_{j=1}^{n}\left[\int_0^\tau \left\{\mathbf{X}_{ij} - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\right\} dN_{ij}(t)\right. \\
&\left. -e^{\mathbf{X}'_{ij}\boldsymbol{\beta}_0 + b_i}\int_0^\tau Y_{ij}(t)\left\{\mathbf{X}_{ij} - \frac{s^{(1)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\right\}\frac{dG(t)}{s^{(0)}(t, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)}\right] + o_p(1).
\end{aligned}$$

In other words, $\mathbf{R}_M(\boldsymbol{\beta}_0) = o_p(1)$. The above derivation resembles the decomposition of the partial likelihood score (Lin and Wei 1989) for independent survival data.

Recalling that $\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta})$ is the partial likelihood information (given frailties) and applying Theorems 3.2 and 4.2 of Andersen and Gill (1982), one may show that the second term in (B.2) is $o_p(1)$. With the UWLLN conditions on $\partial\boldsymbol{\phi}/\partial\boldsymbol{\beta}$, the third term in (B.2) is $o_p(1)$ as well. Hence we have shown that $E(C_2) \to 0$ and, furthermore, $C_2 = o_p(1)$, in view of the nonnegativity of C_2. Therefore, $C_1 C_2 = o_p(1)$, which proves the lemma. □

We next show in the following lemma that $\mathbf{A}_{ij}^M(\boldsymbol{\gamma})$, defined in (5.1), is a consistent estimator of $\mathbf{A}_{ij}$ in a small neighborhood of $\boldsymbol{\gamma}_0$.

Lemma 2. *As $M \to \infty$, for each $K > 0$,*

$$\sup_{\boldsymbol{\gamma} \in \mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)} ||\mathbf{A}_{ij}^M(\boldsymbol{\gamma}) - \mathbf{A}_{ij}|| = o_p(1)$$

Proof: As shown in the previous lemma,

$$\bigvee ||M^{-1}\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) - \mathbf{Q}(\boldsymbol{\gamma}_0)|| = o_p(1).$$

By a similar calculation as before,

$$\begin{aligned} Z_M &= \bigvee || \int \{M^{-1}\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) - \mathbf{Q}(\boldsymbol{\gamma}_0)\} dF(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\gamma})|| \\ &\leq E(D_M|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\gamma}_0) C_1, \end{aligned} \tag{B.4}$$

where C_1 is as in (B.1) and $D_M = \bigvee ||M^{-1}\mathcal{I}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}) - \mathbf{Q}(\boldsymbol{\gamma}_0)||$. Hence, $D_M = o_p(1)$. As $\mathbf{X}$ is bounded, for any $\mathbf{b}$, it is easy to show $\mathcal{S}^{(2)}(t, \boldsymbol{\beta}, \mathbf{b})/\mathcal{S}^{(0)}(t, \boldsymbol{\beta}, \mathbf{b})$ and $\mathcal{S}^{(1)}(t, \boldsymbol{\beta}, \mathbf{b})/\mathcal{S}^{(0)}(t, \boldsymbol{\beta}, \mathbf{b})$ are uniformly bounded. Therefore, D_M is bounded. By the dominated convergence theorem, $E(D_M) \to 0$. Hence, $Z_M = o_p(1)$. One can also establish that

$$\bigvee || \int \mathbf{R}_M(\boldsymbol{\beta}) \mathbf{U}_{b,\boldsymbol{\beta}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\eta}) dF(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\gamma})|| = o_p(1). \tag{B.5}$$

On the other hand,

$$\begin{aligned} &M^{-1} \int \sum_{i=1}^M \boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}) \mathbf{U}_{b,\boldsymbol{\beta}}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\eta}) dF(\mathbf{b}|\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}; \boldsymbol{\gamma}) \\ &= M^{-1} \sum_{i=1}^M \int \boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}) \mathbf{U}_{b_i,\boldsymbol{\beta}}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}) dF(b_i|\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i; \boldsymbol{\gamma}) \\ &\xrightarrow{p} E\{\boldsymbol{\phi}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma}) \mathbf{U}_{b_i,\boldsymbol{\beta}}(\mathbf{T}_i, \boldsymbol{\Delta}_i, \mathbf{X}_i, b_i; \boldsymbol{\gamma})\}, \end{aligned} \tag{B.6}$$

uniformly in $\mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)$ by the UWLLN conditions. Thus, combining (B.4)–(B.6), we finish the proof of the uniform convergence of the $\mathbf{A}_{11}^M$. Similarly, we can obtain the convergence for $\mathbf{A}_{12}^M$. Convergence of $\mathbf{A}_{21}^M$ and $\mathbf{A}_{22}^M$ follows from the standard maximum likelihood score argument and the UWLLN conditions. □

With Lemmas 1 and 2 established, we can prove consistency and asymptotic normality of the estimators.

Proof of Theorem 1: Let $P(\boldsymbol{\gamma}) = \{\mathbb{S}(\boldsymbol{\gamma}), \mathbb{U}(\boldsymbol{\gamma})\}$ and assume that $\mathbf{A}$ is positive definite, otherwise we can replace $P(\boldsymbol{\gamma})$ with $\mathbf{A}'P(\boldsymbol{\gamma})$. A standard Taylor expansion gives that

$$M^{-1/2}P(\boldsymbol{\gamma}) = M^{-1/2}P(\boldsymbol{\gamma}_0) - \mathbf{A}_M(\boldsymbol{\gamma}^*)M^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0),$$

where $\boldsymbol{\gamma}^*$ lies between $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}$.

By Lemma 1 and the central limit theorem, $M^{-1/2}P(\boldsymbol{\gamma}_0)$ converges to a mean 0 random normal variable. Hence $M^{-1/2}P(\boldsymbol{\gamma}_0) = O_p(1)$. Let $\epsilon > 0$ be arbitrary. Then for sufficiently large M_{01}, when $M > M_{01}$, on a set with probability $1 - \frac{1}{2}\epsilon$, $||M^{-1/2}P(\boldsymbol{\gamma}_0)|| < J$, where $J < \infty$. By Lemma 2, there exists an $M_{02} > 0$ such that when $M > M_{02}$, on a set with probability $1 - \frac{1}{2}\epsilon$, $\mathbf{A}_M(\boldsymbol{\gamma})$ converges uniformly to $\mathbf{A}$ in $\boldsymbol{\gamma} \in \mathcal{N}_{KM^{-\frac{1}{2}}}(\boldsymbol{\gamma}_0)$, where K is any positive numbers. Let $M_0 = \max(M_{01}, M_{02})$. We then work on the intersection of the two random sets (with probability at least $1 - \epsilon$). Now we fix any $M > M_0$. Denote by $\lambda_{\min}$ the minimum eigenvalue of $\mathbf{A}$. Then for $K_0 = 2J/\lambda_{\min}$, one can show $||(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)'P(\boldsymbol{\gamma})|| \geq M^{-\frac{1}{2}}(\lambda_{\min}K_0^2 - JK_0) > 0$ for $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = K_0M^{-\frac{1}{2}} = 2(J/\lambda_{\min})M^{-\frac{1}{2}}$. Because $P(\boldsymbol{\gamma})$ is continuous in $\boldsymbol{\gamma}$, by Lemma 2 of Aitchison and Silvey (1958) (an application of the fixed point theorem in continuous functions mapping from a closed unit ball to itself), $P(\boldsymbol{\gamma})$ has a solution in $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| < 2(J/\lambda_{\min})M^{-\frac{1}{2}}$. □

Proof of Theorem 2: With $P(\hat{\boldsymbol{\gamma}}) = 0$, expanding it about $\boldsymbol{\gamma}_0$ gives that

$$\mathbf{A}_M(\boldsymbol{\gamma}^*)M^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = M^{-1/2}P(\boldsymbol{\gamma}_0), \tag{B.7}$$

where $\boldsymbol{\gamma}^*$ lies between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_0$. By the proof of Theorem 1, for any $\epsilon > 0$, there exists a $K_0 > 0$ and M such that the event $\{||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| < K_0M^{-1/2}\}$ has measure at least $1 - \epsilon$. Hence, by Lemma 2 $\mathbf{A}_M(\boldsymbol{\gamma}^*) \xrightarrow{p} \mathbf{A}$. Using Lemma 1 and a central limit theorem, one obtains that

$$M^{-1/2}P(\boldsymbol{\gamma}_0) \xrightarrow{d} N(0, \boldsymbol{\Psi}).$$

Hence, Theorem 2 follows from (B.7) by the Slutsky theorem. □

Proof of Theorem 3: Using the standard argument for the convergence of the Newton–Raphson iteration, one can show that there exists an $\epsilon_0 > 0$ such that when $||\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}}^*|| \leq \epsilon_0$, $\hat{\boldsymbol{\gamma}}_j \to \hat{\boldsymbol{\gamma}}^*$, where $\hat{\boldsymbol{\gamma}}^*$ is the solution to

$$\begin{aligned} \hat{\mathbb{S}}(\boldsymbol{\gamma}) &= 0, \\ \hat{\mathbb{U}}(\boldsymbol{\gamma}) &= 0, \end{aligned}$$

where $\hat{\mathbb{S}}(\boldsymbol{\gamma}) = \frac{1}{\bar{w}}\sum_{l=1}^m w^{(l)}\mathbf{S}(\mathbf{T}, \boldsymbol{\Delta}, \mathbf{X}, \mathbf{b}^{(l)}; \boldsymbol{\beta})$. Here,

$$w^{(l)} = L(\mathbf{T}, \boldsymbol{\Delta}|\mathbf{X}, \mathbf{b}^{(0,l)}; \boldsymbol{\gamma})\frac{f(\mathbf{b}^{(0,l)}; \theta)}{g(\mathbf{b}^{(0,l)})}$$

and $\bar{w} = \sum_{l=1}^{m} w^{(l)}$.

Because the supports of $g(\mathbf{b})$, $f(\mathbf{b};\theta)$ and $f(\mathbf{b}|\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\boldsymbol{\gamma})$ are identical, by Theorem 1 in Geweke (1989) $\hat{\mathbb{S}}(\boldsymbol{\gamma}) \rightarrow \mathbb{S}(\boldsymbol{\gamma})$ uniformly and almost surely in a compact set $\mathcal{B}$. Using the infinite smoothness of $L(\mathbf{T}_i,\boldsymbol{\Delta}_i|\mathbf{X}_i,b_i;\boldsymbol{\gamma})$ with respect to b_i and applying the Weierstrass' theorem (Lange 1999, p. 219), when the number of Gaussian abscissas, N_G, goes to ∞, $\hat{\mathbb{U}}(\boldsymbol{\gamma}) \rightarrow \mathbb{U}(\boldsymbol{\gamma})$ uniformly in $\mathcal{B}$. Since $\mathbb{S},\mathbb{U},\hat{\mathbb{S}}$, and $\hat{\mathbb{U}}$ are continuous in $\boldsymbol{\gamma}$, it follows immediately that $\hat{\boldsymbol{\gamma}}^* \overset{\text{almost surely}}{\rightarrow} \hat{\boldsymbol{\gamma}}$.

Applying a one-term Taylor expansion to $\hat{\mathbb{S}}(\hat{\boldsymbol{\gamma}}^*)$ and $\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}^*)$ about $\hat{\boldsymbol{\gamma}}$ and using that $\mathbb{S}(\hat{\boldsymbol{\gamma}}) = 0$ and $U(\hat{\boldsymbol{\gamma}}) = 0$, one obtains

$$\begin{aligned} 0 &= \sqrt{m}\hat{\mathbb{S}}(\hat{\boldsymbol{\gamma}}^*) = \sqrt{m}\{\hat{\mathbb{S}}(\hat{\boldsymbol{\gamma}}) - \mathbb{S}(\hat{\boldsymbol{\gamma}})\} + \frac{\partial}{\partial\boldsymbol{\gamma}}\hat{\mathbb{S}}(\hat{\boldsymbol{\gamma}}^{**})\sqrt{m}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}), \\ 0 &= \sqrt{m}\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}^*) = \sqrt{m}\{\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}) - \mathbb{U}(\hat{\boldsymbol{\gamma}})\} + \frac{\partial}{\partial\boldsymbol{\gamma}}\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}^{**})\sqrt{m}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}), \end{aligned}$$

where $\hat{\boldsymbol{\gamma}}^{**}$ lies on the line segment joining $\hat{\boldsymbol{\gamma}}^*$ and $\hat{\boldsymbol{\gamma}}$.

Applying Theorem 2 in Geweke (1989), we have that $\sqrt{m}\{\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}) - \mathbb{U}(\hat{\boldsymbol{\gamma}})\} \rightarrow N(0,\mathcal{V}_{11})$, where

$$\mathcal{V}_{11} = \int \mathbf{S}(\mathbf{T},\boldsymbol{\Delta},\mathbf{X},\mathbf{b};\hat{\boldsymbol{\beta}})^2 w(\mathbf{b},\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\hat{\gamma})dF(\mathbf{b}|\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\hat{\gamma}),$$

and

$$w(\mathbf{b},\mathbf{T},\boldsymbol{\Delta},\mathbf{X};\hat{\gamma}) = \frac{L(\mathbf{T},\boldsymbol{\Delta}|\mathbf{X},\mathbf{b};\hat{\gamma})f(\mathbf{b},\hat{\theta})}{L(\mathbf{T},\boldsymbol{\Delta}|\mathbf{X};\hat{\gamma})g(\mathbf{b})}.$$

In addition, $\frac{\partial}{\partial\boldsymbol{\gamma}}\hat{\mathbb{U}}(\boldsymbol{\gamma}) \overset{\text{almost surely}}{\rightarrow} \frac{\partial}{\partial\boldsymbol{\gamma}}\mathbb{U}(\boldsymbol{\gamma})$ and $\frac{\partial}{\partial\boldsymbol{\gamma}}\hat{\mathbb{S}}(\boldsymbol{\gamma}) \overset{\text{almost surely}}{\rightarrow} \frac{\partial}{\partial\boldsymbol{\gamma}}\mathbb{S}(\boldsymbol{\gamma})$ uniformly in $\mathcal{B}$. Again using the Weierstrass' theorem, one can show that, for a fixed m, one can choose sufficently large number of Gaussian absissas to let $\sqrt{m}\{\hat{\mathbb{U}}(\hat{\boldsymbol{\gamma}}) - \mathbb{U}(\hat{\boldsymbol{\gamma}})\} = o(1)$.

Combining all the pieces above, one can obtain that $\sqrt{m}(\hat{\boldsymbol{\gamma}}^* - \hat{\boldsymbol{\gamma}}) \overset{d}{\rightarrow} N(0,\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \tilde{\mathbf{A}}_M^{-1}(\hat{\boldsymbol{\gamma}})\mathcal{V}(\tilde{\mathbf{A}}_M^{-1})^T(\hat{\boldsymbol{\gamma}})$, $\tilde{\mathbf{A}}_M(\boldsymbol{\gamma}) = M \cdot \mathbf{A}_M(\boldsymbol{\gamma})$ and $\mathcal{V} = \begin{pmatrix} \mathcal{V}_{11} & 0 \\ 0 & 0 \end{pmatrix}$. □

ACKNOWLEDGMENTS

The authors thank the editor, an associate editor, and two anonymous referees for their helpful and constructive comments, which have greatly improved the presentation of this article. The authors thank Rosalind Wright for providing the East Boston Asthma data. The first author thanks David Harrington for constructive suggestions.

[Received September 2001. Revised June 2002.]

REFERENCES

Aitchison, J., and Silvey, S. D. (1958),"Maximum Likelihood Estimation of Parameters Subject to Constraints," *Annals of Mathematical Statistics*, 29, 813–828.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Process: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1992), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Clayton, D. G., and Cuzick, J. (1985), "Multivariate Generalizations of the Proportional Hazards Model" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 148, 82–117.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Datta, S. (1988), "Asymptotically Optimal Bayes Compound and Empirical Bayes Estimators in Exponential Families with Compact Parameter Spaces," PhD dissertation, Dept. of Statistics and Probability, Michigan State University.

Datta, S., Satten, G. A., and Williamson, J. M. (2000), " Consistency and Asymptotic Normality of Estimators in a Regression Model with Interval Censoring and Left Truncation," *Annals of the Institute of Statistical Mathematics*, 52, 160–172.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1981), "Nonparametric Estimates of Standard Error: The Jacknife, the Bootstrap and Other," *Biometrika*, 68, 589–599.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis,* New York: Wiley.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.

Gray, R. J. (1995), "Tests for Variation over Groups in Survival Data, " *Journal of the American Statistical Association*, 90, 198–203.

Lange, K. (1999), *Numerical Analysis for Statisticians*, New York: Springer-Verlag.

Li, Y., and Lin, X. (2000), "Covariate Measurement Errors in Frailty Models for Clustered Survival Data," *Biometrika*, 87, 849–866.

Lin, D. Y., and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 89, 1074–1078.

Lin, X. (1997), "Variance Component Testing in Generalized Linear Models with Random Intercepts," *Biometrika*, 84, 309–326.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Murphy, S. A. (1994), "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *The Annals of Statistics*, 22, 712–731.

——— (1995), "Asymptotic Theory for the Frailty Model," *The Annals of Statistics*, 23, 182–198.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992), "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics*, 19, 25–43.

Oakes, D., and Jeong, J. (1998), "Frailty Models and Rank Tests," *Lifetime Data Analysis*, 4, 209–228.

Parner, E. (1998), "Asymptotic Theory for the Correlated Gamma-frailty Model," *The Annals of Statistics*, 26, 183–214.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in Fortran: The Art of Scientific Computing* (2nd ed.), New York: Cambridge University Press.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Hayward, CA: Institute of Mathematical Statistics.

Ramlau–Hansen, H. (1983), "Smoothing Counting Process Intensities by Means of Kernel Functions," *The Annals of Statistics*, 11, 453–466.

Reid, N., and Crépeau, H. (1985), "Influence Functions for the Proportional Hazards Regression," *Biometrika*, 72, 1–9.

Satten, G. (1996), “Rank-Based Inference in the Proportional Hazards Model for Interval Censored Data,” *Biometrika*, 83, 355–370.

Satten, G., and Datta, S. (2000), “The S-U Algorithm for Missing Data Problems,” *Computational Statistics*, 15, 243–277.

Satten, G., Datta, S., and Williamson, J. M. (1998), “Inference Based on Imputed Failure Times for the Proportional Hazards Model with Interval Censored Data,” *Journal of the American Statistical Association*, 93, 318–327.

Shoukri, M. M., Attanasio, M., and Sargeant, J. M. (1998), “Parametric Versus Semi-parametric Models for the Analysis of Correlated Survival Data: A Case Study in Veterinary Epidemiology,” *Journal of Applied Statistics*, 25, 357–374.

Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.), New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82 , 528–540.

Wei, G. C., and Tanner, M. A. (1990), “A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithm,” *Journal of the American Statistical Association*, 85, 699–704.