Statistica Sinica **18**(2008), 355-377

SEMIPARAMETRIC TRANSFORMATION MODELS WITH RANDOM EFFECTS FOR CLUSTERED FAILURE TIME DATA

Donglin Zeng[1], D. Y. Lin[1] and Xihong Lin[2]

[1] *University of North Carolina at Chapel Hill and* [2] *Harvard University*

Abstract: We propose a general class of semiparametric transformation models with random effects to formulate the effects of possibly time-dependent covariates on clustered or correlated failure times. This class encompasses all commonly used transformation models, including proportional hazards and proportional odds models, and it accommodates a variety of random-effects distributions, particularly Gaussian distributions. We show that the nonparametric maximum likelihood estimators of the model parameters are consistent, asymptotically normal and asymptotically efficient. We develop the corresponding likelihood-based inference procedures. Simulation studies demonstrate that the proposed methods perform well in practical situations. An illustration with a well-known diabetic retinopathy study is provided.

Key words and phrases: Correlated failure times, frailty model, nonparametric maximum likelihood estimation, proportional hazards, semiparametric efficiency, survival analysis.

1. Introduction

Clustered failure time data arise when the study subjects are sampled in clusters so that the failure times within the same cluster tend to be correlated. Medical examples include the onset of a genetic disease among family members, the appearance of tumors in littermates exposed to a carcinogen, the occurrence of visual loss in left and right eyes, and the initiation of cigarette smoking by classmates. Such failure times are inevitably subject to right censoring. The presence of censoring and intra-class dependence poses serious challenges in the semiparametric regression analysis of these data.

One approach to formulating the effects of covariates on the failure time while accounting for the intra-class dependence is the proportional hazards frailty model, under which the hazard function for the jth subject of the ith cluster associated with covariates $\mathbf{X}_{ij}(\cdot)$ takes the form

$$\lambda(t|\mathbf{X}_{ij},\xi_i)=\xi_i\lambda_0(t)e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}},\quad i=1,\ldots,n; j=1,\ldots,n_i, \tag{1.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, and ξ_i is an unobserved frailty for the ith cluster. Statistical inference under model (1.1) turns out to be an interesting and challenging problem. The consistency and asymptotic distribution of the nonparametric maximum likelihood estimator for this model have been rigorously studied by Murphy (1994, 1995) for the case of no covariates, and by Parner (1998) for the case with covariates. All the results are restricted to the special case of gamma frailty.

The proportional hazards model with gamma frailty, although very interesting and useful, has important limitations. First, the proportional hazards assumption on the effects of covariates may not be reasonable in certain applications. Secondly, gamma frailty induces a restrictive form of dependence.

To address the above concerns, we study a broad class of transformation models with random effects. For the jth subject of the ith cluster, let $\mathbf{X}_{ij}(\cdot)$ be a d_1-vector of (possibly time-dependent) covariates, and $\mathbf{Z}_{ij}(\cdot)$ be another set of covariates, which may contain 1 and part of $\mathbf{X}_{ij}(\cdot)$. Also, let $\overline{\mathbf{X}}_{ij}(t)$ and $\overline{\mathbf{Z}}_{ij}(t)$ denote the histories of $\mathbf{X}_{ij}(\cdot)$ and $\mathbf{Z}_{ij}(\cdot)$ over $[0,t]$. The cumulative hazard function of T_{ij}, the jth failure time of the ith cluster, is related to $\mathbf{X}_{ij}(\cdot)$ and $\mathbf{Z}_{ij}(\cdot)$ as follows:

$$\Lambda(t|\overline{\mathbf{X}}_{ij}(t),\overline{\mathbf{Z}}_{ij}(t),\mathbf{b}_i)=H_0\Big(\int_0^t e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(s)^T\mathbf{b}_i}d\Lambda(s)\Big),\ i=1,\ldots,n;\ j=1,\ldots,n_i,\quad(1.2)$$

where H_0 is a known increasing function with $H_0(0)=0$ and $H_0(\infty)=\infty$, $\Lambda(\cdot)$ is an unspecified increasing function, $\boldsymbol{\beta}$ is a set of unknown regression parameters, and $\mathbf{b}_i$ is a set of unobserved mean-zero random effects for the ith cluster with a density function $\psi(\mathbf{b}_i;\boldsymbol{\gamma})$ (with respect to a σ-finite measure $\mu(\mathbf{b}_i)$) indexed by a d_2-dimensional parameter $\boldsymbol{\gamma}$. Note that (1.2) allows covariate-specific or subject-specific random effects.

Let $G_0(x)=1-e^{-H_0(x)}$. We may rewrite (1.2) as

$$\int_0^{T_{ij}} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(s)^T\mathbf{b}_i}d\Lambda(s)=\epsilon_{ij},\quad i=1,\ldots,n;j=1,\ldots,n_i,\qquad(1.3)$$

where the ϵ_{ij} are i.i.d. random variables from a known distribution with the cumulative distribution function $G_0(\cdot)$. If both $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are time-independent, then (1.3) reduces to linear transformation models

$$H(T_{ij})=-\mathbf{X}_{ij}^T\boldsymbol{\beta}-\mathbf{Z}_{ij}^T\mathbf{b}_i+\log\epsilon_{ij},\quad i=1,\ldots,n;j=1,\ldots,n_i,\qquad(1.4)$$

where $H(x)=\log\Lambda(x)$. The choices of the extreme-value and standard logistic distributions for $\log\epsilon_{ij}$ or $G_0(x)=1-e^{-x}$ and $G_0(x)=1-(1+x)^{-1}$ correspond

to the proportional hazards model (Cox (1972)) and the proportional odds model (Bennett (1983) and Pettitt (1984)), respectively.

Equation (1.4) is reminiscent of the linear mixed-effects model (Laird and Ware (1982)) for longitudinal data. For the latter model, however, the transformation of the response variable is known, and there is no censoring. The presence of censoring and the involvement of an unknown transformation make the estimation of transformation models with random effects for correlated failure time data much harder. In view of the linear model representation given in (1.4), Gaussian random effects are the most natural choice even for the proportional hazards model. The focus of the existing literature on gamma frailty is due to its mathematical simplicity.

Linear transformation models for independent failure time data (i.e., in the absence of random effects) have been studied extensively. In particular, the proportional odds model was studied by Bennett (1983), Pettitt (1984), Cuzick (1988), Wu (1995), Murphy, Rossini and van der Vaart (1997), Shen (1998) and Lam and Leung (2001). Estimation for general linear transformation models was investigated by Bickel (1986), Dabrowska and Doksum (1988), Cheng, Wei and Ying (1995) and Chen, Jin and Ying (2002), among others. Recently, Kosorok, Lee and Fine (2004) considered a class of frailty models for independent observations which is a one-parameter extension of the proportional hazards model.

For clustered failure time data, Cai, Cheng and Wei (2002) considered the class of models given in (1.4) with a scalar random effect (i.e., $\mathbf{Z}_{ij} \equiv 1$). They proposed to estimate the parameters by minimizing the empirical sum of squares of the differences between certain observed quantities and their expected values. The estimators are not asymptotically efficient, and the variance estimation is computationally demanding. Furthermore, the censoring mechanism is required to be purely random and independent of covariates. Recently, Zeng, Lin and Yin (2005) studied efficient estimation of a special member of (1.4), namely, the proportional odds model with time-independent covariates and Gaussian random effects. They showed that the estimators of Cai et al. (2002) can be quite inefficient. Efficient estimation of (1.4), let alone (1.2), has not been studied in any generality.

In this paper, we study nonparametric maximum likelihood estimation of (1.2). Rather than focusing on specific models, we identify general conditions on the transformation $G_0(\cdot)$ and the distribution of random effects under which the nonparametric maximum likelihood estimators have desirable asymptotic properties. We show that many commonly used transformations, including the familiar Box-Cox transformations and the class of logarithmic transformations studied by Chen et al. (2002), and Gaussian distributions of random effects ensure that the nonparametric maximum likelihood estimators for the regression parameters are

asymptotically efficient. Important special cases include the Cox proportional hazards and proportional odds models with Gaussian random effects.

The structure of this paper is as follows. In Section 2, we describe the proposed methodology based on the nonparametric likelihood. In Section 3, we provide the asymptotic theory behind the proposed methodology. In Section 4, we report the results of our simulation studies. In Section 5, we provide an illustration with a medical study. In Section 6, we provide some concluding remarks. We relegate the proofs of the theoretical results to an appendix.

2. Likelihood and Inference

Suppose that there are n independent clusters with potentially different sizes. The relationship between T_{ij} and $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ is given in equation (1.2) or (1.3). Let C_{ij} be the censoring time on T_{ij}. The data consist of $(Y_{ij}, \Delta_{ij}, \overline{\mathbf{X}}_{ij}(Y_{ij}), \overline{\mathbf{Z}}_{ij}(Y_{ij}))$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$, where $Y_{ij} = T_{ij} \wedge C_{ij}$ and $\Delta_{ij} = I(T_{ij} \le C_{ij})$. Here and in the sequel, $a \wedge b = \min(a, b)$, and $I(\cdot)$ is the indicator function. Our goal is to make inference about the regression parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the function $\Lambda(\cdot)$.

We make the coarsening at random assumption that, conditional on $\overline{\mathbf{X}}_{ij}(\cdot)$, $\overline{\mathbf{Z}}_{ij}(\cdot)$, T_{ij} and $\mathbf{b}_i$, the hazard function of C_{ij} at time t is only a function of $\overline{\mathbf{X}}_{ij}(t)$ and $\overline{\mathbf{Z}}_{ij}(t)$. Then under (1.2), the likelihood function for the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$ is proportional to

$$\prod_{i=1}^{n}\Bigg[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\Bigg\{G_0'\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\Big)e^{\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}}\Lambda'(Y_{ij})\Bigg\}^{\Delta_{ij}}$$
$$\times\Bigg\{1-G_0\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\Big)\Bigg\}^{1-\Delta_{ij}}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b})\Bigg]$$

where, for any function g, $g'(x)$ is the derivative of $g(x)$.

It would seem natural to calculate the maximum likelihood estimators of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$ by maximizing the above likelihood function. The maximum of this function, however, is infinity since we can always choose some function $\Lambda(t)$ with fixed values at each Y_{ij} while letting $\Lambda'(Y_{ij})$ go to infinity for some Y_{ij} with $\Delta_{ij} = 1$. Thus, we relax $\Lambda(t)$ to be right-continuous and allow $\Lambda(t)$ to have jumps at the Y_{ij}. We then propose to maximize

$$L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$$
$$\equiv\prod_{i=1}^{n}\Bigg[\int_{\mathbf{b}}\prod_{j=1}^{n_i}\Bigg\{G_0'\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\Big)e^{\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}}\Lambda\{Y_{ij}\}\Bigg\}^{\Delta_{ij}}$$
$$\times\Bigg\{1-G_0\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\Big)\Bigg\}^{1-\Delta_{ij}}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b})\Bigg], \tag{2.1}$$

where $\Lambda\{t\}$ denotes the jump size of $\Lambda(\cdot)$ at t. To be specific, we maximize $L_n(\boldsymbol{\beta}, \gamma, \Lambda)$ over the parameter space

$$\{(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda) : (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta,\ \Lambda(t) \text{ is an increasing step function in } [0, \tau] \text{ with jumps at the observed failure times and } \Lambda(0) = 0\}.$$

The resulting estimators, denoted by $\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n$ and $\widehat{\Lambda}_n$, correspond to the Kiefer-Wolfowitz nonparametric maximum likelihood estimators (NPMLEs).

We show later that the maximum of (2.1) exists and that the jump sizes of $\widehat{\Lambda}_n$ are finite. Thus, the NPMLEs for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$ can be obtained by maximizing $L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$ over the parameter space $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta$ and the jump sizes of Λ at the Y_{ij} for which $\Delta_{ij} = 1$. Computationally, to ensure the positiveness of the jump size, we can use the transformed parameter $\log(\Lambda\{Y_{ij}\})$ instead of $\Lambda\{Y_{ij}\}$ in the maximization. For a general transformation $G_0(\cdot)$, the maximization can be realized via optimization algorithms which consist of optimum search based on the interior-reflective Newton method (Coleman and Li (1994, 1996)). These algorithms are available in the optimization toolbox of MATLAB. In the numerical calculation, the integration over $\mathbf{b}$ is replaced by numerical summation, such as the Gaussian quadrature approximation for Gaussian $\mathbf{b}$. In each iteration of the search, a large linear system is approximately solved by using the method of preconditioned conjugate gradients (Coleman and Li (1994, 1996)). This search works very well in our setting. In the special case when the transformation $G_0(\cdot)$ induces the proportional hazards model, the maximization can be carried out efficiently by the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin (1977)). In the EM algorithm, random effects are treated as missing data and efficient computation takes advantage of the explicit solution for estimating $\Lambda(\cdot)$ in the M-step.

It is desirable to estimate the asymptotic covariance matrices of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\gamma}}_n$. When the nuisance parameter is of high dimension, i.e., the number of jumps in $\widehat{\Lambda}_n$ is large, the profile likelihood method (Murphy and van der Vaart (2000)) is particularly useful in estimating the variances. We define the profile log-likelihood function for $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\gamma})$ as

$$\mathrm{pl}_n(\boldsymbol{\theta}) = \max_{\Lambda} l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda),$$

where $l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda) = \log L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$, and Λ is any right-continuous and increasing function in $[0, \tau]$ with $\Lambda(0) = 0$. Theorem 3 of Section 3 states that the asymptotic covariance for $(\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\gamma}}_n^T)^T$ can be estimated by the negative inverse of the curvature of $\mathrm{pl}_n(\boldsymbol{\theta})$ around $(\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\gamma}}_n^T)^T$. Specifically, to estimate the (s, l)th element of the asymptotic covariance matrix for $(\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\gamma}}_n^T)^T$, we choose a constant

h_n of the order $1/\sqrt{n}$, and let $\mathbf{e}_s$ and $\mathbf{e}_l$ be the canonical bases which are one at the sth and the lth coordinates, respectively, and are zero elsewhere. Then the (s,l)th element of the inverse of the asymptotic covariance matrix can be estimated by

$$-\frac{1}{h_n^2}\Big\{\mathrm{pl}_n(\widehat{\boldsymbol{\theta}}_n + h_n\mathbf{e}_s + h_n\mathbf{e}_l) - \mathrm{pl}_n(\widehat{\boldsymbol{\theta}}_n - h_n\mathbf{e}_s + h_n\mathbf{e}_l) \\ -\mathrm{pl}_n(\widehat{\boldsymbol{\theta}}_n + h_n\mathbf{e}_s - h_n\mathbf{e}_l) + \mathrm{pl}_n(\widehat{\boldsymbol{\theta}}_n)\Big\}.$$

Thus, we need to evaluate the profile likelihood function $\mathrm{pl}_n(\boldsymbol{\theta})$ in a neighborhood of $\widehat{\boldsymbol{\theta}}_n$. Computationally, for a general transformation $G_0(\cdot)$, the profile likelihood function can be calculated by using the optimization search for fixed $\boldsymbol{\theta}$ close to $\widehat{\boldsymbol{\theta}}_n$. When $G_0(\cdot)$ is the transformation corresponding to the proportional hazards model, the profile likelihood function can be calculated via the EM algorithm in which $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ is held constant in both the E-step and M-step, so that the only updated parameters are the jump sizes of $\Lambda(\cdot)$ at the observed failure times. Our experiences showed that the EM algorithm is more efficient than direct optimization.

When the number of observed failure times is not large, an alternative way of estimating the asymptotic variance is simply to invert the observed information matrix for all the parameters including $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and the jump sizes of $\widehat{\Lambda}_n$. That is, we treat the likelihood function (2.1) as a likelihood function from a parametric model. One benefit of this approach is that we can estimate the asymptotic variance for $\widehat{\Lambda}_n$. The validity of inverting the observed information matrix is ensured by Theorem 4. Our numerical studies revealed that this approach works very well in practical situations.

3. Asymptotic Theory

We impose the following regularity conditions.

C.1. There exists some positive constant δ_0 such that $P(C_{ij} \geq \tau | \overline{\mathbf{X}}_{ij}(\tau), \overline{\mathbf{Z}}_{ij}(\tau)) = P(C_{ij} = \tau | \overline{\mathbf{X}}_{ij}(\tau), \overline{\mathbf{Z}}_{ij}(\tau)) \geq \delta_0$ almost surely, where τ is a constant denoting the end of the study.

C.2. With probability one, $\mathbf{X}_{ij}(\cdot)$ and $\mathbf{Z}_{ij}(\cdot)$ have right-continuous sample paths in $[0,\tau]$ and their right derivatives exist. In addition, there exists a constant M_0 such that

$$P\Big(\max_{1\leq j\leq n_i}\sup_{t\in[0,\tau]}\big\{|\mathbf{X}_{ij}(t)| + |\mathbf{X}_{ij}{}'_+(t)| + |\mathbf{Z}_{ij}(t)| + |\mathbf{Z}_{ij}{}'_+(t)|\big\} \leq M_0\Big) = 1.,$$

where $\mathbf{X}_{ij}{}'_+$ and $\mathbf{Z}_{ij}{}'_+$ denote the right derivatives.

C.3. The true value $\Lambda_0(t)$ of $\Lambda(t)$ is a strictly increasing function in $[0,\tau]$ and is continuously differentiable. In addition, $\Lambda_0(0)=0$ and $\Lambda_0'(0)>0$.

C.4. The true values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, denoted by $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$, belong to a known compact set

$$\Theta=\Big\{(\boldsymbol{\beta},\boldsymbol{\gamma}):\|\boldsymbol{\beta}\|\le B_0 \text{ for some constant } B_0 \text{ and } \boldsymbol{\gamma} \text{ is in a known compact set } \boldsymbol{\Gamma}_0\Big\}.$$

C.5. The size of the cluster is independent of the survival and censoring variables, and $\max_{1\le i\le n}|n_i|\le n_0$ for a constant n_0, almost surely.

C.6. The function $G_0(x):[0,\infty)\to[0,1]$ is four times-continuously differentiable in $[0,\infty)$ with $G_0(0)=0$, $G_0'(x)>0$, and $\sup_{x\ge0}\{|G_0^{(k)}(x)|\}<\infty$ for $k=1,2,3,4$, where $G_0^{(k)}(x)$ denotes the kth derivative of $G_0(x)$. The function $\psi(\mathbf{b};\boldsymbol{\gamma})$ is thrice-differentiable with respect to $\boldsymbol{\gamma}$, and for $k=1,2,3$, $\int_{\mathbf{b}}|\psi^{(k)}(\mathbf{b};\boldsymbol{\gamma})|d\mu(\mathbf{b})$ is uniformly bounded for $\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_0$.

C.7. There exists a positive constant ρ_0 such that

$$\limsup_{x\to\infty}(1+x)^{\rho_0}(1-G_0(x))<\infty,\quad \limsup_{x\to\infty}(1+x)^{1+\rho_0}G_0'(x)<\infty. \tag{3.1}$$

C.8. For any fixed constant K,

$$\sup_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_0}\int_{\mathbf{b}}e^{K\|\mathbf{b}\|}\Big\{\sum_{k=0}^{3}|\psi^{(k)}(\mathbf{b};\boldsymbol{\gamma})|\Big\}d\mu(\mathbf{b})<\infty. \tag{3.2}$$

C.9. For any pair of parameters $(\boldsymbol{\beta}_1,\boldsymbol{\gamma}_1,\Lambda_1)$ and $(\boldsymbol{\beta}_2,\boldsymbol{\gamma}_2,\Lambda_2)$, if with probability one,

$$\int_{\mathbf{b}}\prod_{j=1}^{k}G_0\Big(\int_0^{t_j}e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_1+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_1\Big)\psi(\mathbf{b};\boldsymbol{\gamma}_1)d\mathbf{b}$$
$$=\int_{\mathbf{b}}\prod_{j=1}^{k}G_0\Big(\int_0^{t_j}e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_2+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_2\Big)\psi(\mathbf{b};\boldsymbol{\gamma}_2)d\mathbf{b}$$

for any $k\in\{1,\dots,n_i\}$ and any $t_1,\dots,t_k\in[0,\tau]$, then $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}_1=\boldsymbol{\gamma}_2$ and $\Lambda_1(t)=\Lambda_2(t)$ for $t\in[0,\tau]$.

C.10. If $\mathbf{X}_{ij}(t)^T\mathbf{h}_1+h(t)=0$ with probability one for some vector $\mathbf{h}_1$ and a function $h(t)$, then $\mathbf{h}_1=\mathbf{0}$ and $h(t)=0$. In addition, if there exist a vector $\mathbf{h}_2$ and functions $A_j(t,\mathbf{b})$, $j=1,\dots,n_i$ such that with probability one,

$$\int_{\mathbf{b}}\prod_{j=1}^{k}G_0\Big(\int_0^{t_j}e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_0\Big)\Big\{\sum_{j=1}^{k}A_j(t_j,\mathbf{b})+\frac{\psi'(\mathbf{b};\boldsymbol{\gamma}_0)^T\mathbf{h}_2}{\psi(\mathbf{b};\boldsymbol{\gamma}_0)}\Big\}d\mathbf{b}=0$$

for any $k \in \{1, \ldots, n_i\}$ and any $t_1, \ldots, t_k \in [0, \tau]$, then $\mathbf{h}_2 = \mathbf{0}$ and $A_j(t, \mathbf{b}) = 0$, $j = 1, \ldots, n_i$.

Remark 1. C.1−C.5 are standard conditions for clustered failure time data. Conditions C.6 and C.7 are satisfied by all common transformations, including the Box-Cox transformations $H_0(x) = ((1 + x)^\rho - 1)/\rho$, and the logarithmic transformations $H_0(x) = r^{-1} \log(1 + rx)$ (Chen et al. (2002)). Condition C.8 pertains to the random-effects distribution. This condition is clearly satisfied by the Gaussian distribution and any distribution with tails less heavy than $e^{-\|\mathbf{b}\|^{1+\epsilon_0}}$ with $\epsilon_0 > 0$ (e.g., log-inverse Gaussian). Condition C.9 pertains to parameter identifiability, while C.10 entails that the Fisher information along any submodel at the true parameters is nonsingular. If $\mathbf{X}$ and $\mathbf{Z}$ are time-independent, then C.9 and C.10 reduce to C.9' and C.10':

C.9' For any $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, if there exist two constant vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ such that with probability one,

$$\int_{\mathbf{b}} \prod_{j=1}^{k} e^{[1, \mathbf{X}_{ij}^T]\boldsymbol{\phi}_1 + \mathbf{Z}_{ij}^T \mathbf{b}} \psi(\mathbf{b}; \boldsymbol{\gamma}_1) d\mathbf{b} = \int_{\mathbf{b}} \prod_{j=1}^{k} e^{[1, \mathbf{X}_{ij}^T]\boldsymbol{\phi}_2 + \mathbf{Z}_{ij}^T \mathbf{b}} \psi(\mathbf{b}; \boldsymbol{\gamma}_2) d\mathbf{b}$$

for any $k \in \{1, \ldots, n_i\}$, then $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ and $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$.

C.10' If there exist two vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ such that with probability one,

$$\int_{\mathbf{b}} \prod_{j=1}^{k} e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}} \Big\{ \sum_{j=1}^{k} [1, \mathbf{X}_{ij}^T] \mathbf{h}_1 + \frac{\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{h}_2}{\psi(\mathbf{b}; \boldsymbol{\gamma}_0)} \Big\} \psi(\mathbf{b}; \boldsymbol{\gamma}_0) d\mathbf{b} = 0, \ k = 1, \ldots, n_i,$$

then $\mathbf{h}_1 = \mathbf{0}$ and $\mathbf{h}_2 = \mathbf{0}$. When the random effects are Gaussian and the $\mathbf{Z}_{ij}$ are the same within each cluster, C.9 and C.10 are implied by the linear independence of the covariates.

The following lemma holds under conditions C.7 and C.8.

Lemma 1. *With probability one,*

$$\begin{aligned} &\int_{\mathbf{b}} \prod_{j=1}^{n_i} \Big\{ G_0' \Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T \boldsymbol{\beta} + \mathbf{Z}_{ij}(t)^T \mathbf{b}} d\Lambda(t) \Big) e^{\mathbf{X}_{ij}(Y_{ij})^T \boldsymbol{\beta} + \mathbf{Z}_{ij}(Y_{ij})^T \mathbf{b}} \Big\}^{\Delta_{ij}} \\ &\qquad \times \Big\{ 1 - G_0 \Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T \boldsymbol{\beta} + \mathbf{Z}_{ij}(t)^T \mathbf{b}} d\Lambda(t) \Big) \Big\}^{1 - \Delta_{ij}} \psi(\mathbf{b}; \boldsymbol{\gamma}) d\mu(\mathbf{b}) \\ &\quad \le c_0 \prod_{j=1}^{n_i} \{1 + \Lambda(Y_{ij})\}^{-(\Delta_{ij} + \rho_0)}, \end{aligned} \tag{3.3}$$

where c_0 is a constant independent of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and Λ.

Remark 2. Inequality (3.3) is essential to the consistency of the NPMLEs. In fact, C.7 and C.8 can be replaced by (3.3) in proving consistency. We impose C.7 and C.8 because they are easier to verify. Although the popular Cox proportional hazard model with gamma frailty does not satisfy C.8, we now show that (3.3) still holds for this model. Under this model, the left-hand side of (3.3) is

$$\exp\Big\{\sum_{j=1}^{n_i}\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}\Big\}\frac{\Gamma(\gamma+\sum_{j=1}^{n_i}\Delta_{ij})}{\Gamma(\gamma)\gamma^{\sum_{j=1}^{n_i}\Delta_{ij}}}\Big[1+\frac{1}{\gamma}\Big\{\sum_{j=1}^{n_i}\int_0^{Y_{ij}}e^{\mathbf{X}_{ij}^T(t)\boldsymbol{\beta}}d\Lambda(t)\Big\}\Big]^{-\sum_{j=1}^{n_i}\Delta_{ij}-\gamma}$$
$$\le O(1)\Big\{\sum_{j=1}^{n_i}\Big(1+\int_0^{Y_{ij}}e^{\mathbf{X}_{ij}^T(t)\boldsymbol{\beta}}d\Lambda(t)\Big)\Big\}^{-\sum_{j=1}^{n_i}(\Delta_{ij}+\frac{\gamma}{n_i})},$$

where $O(1)$ denotes some positive constant. Since

$$\frac{\Delta_{ij}+\frac{\gamma}{n_i}}{\sum_{j=1}^{n_i}(\Delta_{ij}+\frac{\gamma}{n_i})}\le\frac{1+\frac{\gamma}{n_i}}{\gamma},$$

the right-hand side of the above inequality is bounded by

$$O(1)\Big\{\sum_{j=1}^{n_i}\frac{\Delta_{ij}+\frac{\gamma}{n_i}}{\sum_{j=1}^{n_i}(\Delta_{ij}+\frac{\gamma}{n_i})}\Big(1+\int_0^{Y_{ij}}e^{\mathbf{X}_{ij}^T(t)\boldsymbol{\beta}}d\Lambda(t)\Big)\Big\}^{-\sum_{j=1}^{n_i}(\Delta_{ij}+\frac{\gamma}{n_i})}.$$

By the concavity of $\log(x)$, we obtain the upper bound

$$O(1)\prod_{j=1}^{n_i}\Big(1+\int_0^{Y_{ij}}e^{\mathbf{X}_{ij}^T(t)\boldsymbol{\beta}}d\Lambda(t)\Big)^{-(\Delta_{ij}+\frac{\gamma}{n_i})}.$$

This gives the inequality (3.3) in which $\rho_0=\gamma/n_0$.

As stated in the next lemma, C.9 and C.10 ensure identifiability of parameters and non-singularity of information matrix.

Lemma 2. *Under C.*9 *and C.*10*, the parameters in* (1.2) *are identifiable. Furthermore, the Fisher information matrix along any one-dimensional submodel is non-singular.*

Our last lemma pertains to the existence of the NPMLEs.

Lemma 3. *Under C.*1 $\sim$*C.*8*, the maximum likelihood estimators* $(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n)$ *exist almost surely.*

The following two theorems state our main results about the asymptotic properties of the proposed maximum likelihood estimators.

Theorem 1. *Under C.1 ∼C.10,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$, $\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\| \to 0$ *and* $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_n(t) - \Lambda_0(t)| \to 0$ *almost surely, where* $\|\cdot\|$ *is the Euclidean norm.*

Theorem 2. *Under C.1 ∼C.10,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^T - \boldsymbol{\beta}_0^T, \widehat{\boldsymbol{\gamma}}_n^T - \boldsymbol{\gamma}_0^T, \widehat{\Lambda}_n - \Lambda_0)^T$ *weakly converges to a zero-mean Gaussian process in the metric space* $R^{d_1} \times R^{d_2} \times l^\infty[0,\tau]$, *where* $l^\infty[0,\tau]$ *is the linear space consisting of all the bounded functions in* $[0,\tau]$ *and is equipped with the supremum norm. Furthermore,* $\widehat{\boldsymbol{\beta}}_n$ *and* $\widehat{\boldsymbol{\gamma}}_n$ *are asymptotically efficient.*

Remark 3. Theorem 1 states the consistency of the maximum likelihood estimators. In C.1 to C.10, $\Lambda(\cdot)$ is not assumed to be a bounded function, which means that the weak-compactness of the parameter $\Lambda(\cdot)$ is not imposed. Thus, obtaining a bound for $\widehat{\Lambda}_n(\cdot)$ is a key to the proof of Theorem 1. The consistency proof is based on the essential inequality (3.3), and it adopts the partitioning idea from Murphy's (1994) proof of the consistency in the gamma frailty model. This partitioning idea was also used by Parner (1998), Kosorok et al. (2004) and Zeng et al. (2005). However, we provide a novel justification to avoid the concavity of G_0 assumed in all previous papers. Once the consistency is established, the asymptotic distributions of the maximum likelihood estimators stated in Theorem 2 can be proved by verifying the conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Our verification of the continuous invertibility of the information operator is specific to model (1.2) and is based on Lemma 2. Moreover, the Donsker property of some new classes of functions is proven. In the statement of Theorem 2, asymptotically efficient estimators mean that the asymptotic variances attain the semiparametric efficiency bounds as defined in Bickel et al. (1993, Chap. 3).

The next two theorems justify the validity of the proposed approach to estimating the asymptotic covariance.

Theorem 3. *Under C.1 ∼C.10,*

$$-\frac{pl_n(\widehat{\boldsymbol{\theta}}_n + h_n\mathbf{e}) - 2pl_n(\widehat{\boldsymbol{\theta}}_n) + pl_n(\widehat{\boldsymbol{\theta}}_n - h_n\mathbf{e})}{nh_n^2} \to^p \mathbf{e}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{e},$$

where $h_n = O_p(n^{-1/2})$, $\mathbf{e}$ *is any vector in* $R^{d_1+d_2}$ *with norm* 1, *and* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is the efficient information matrix for* $\boldsymbol{\theta}_0 \equiv (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$.

Theorem 3 does not deal with the estimation of the asymptotic variance of $\widehat{\Lambda}_n$, which is often desirable when one wishes to make prediction on future survival experience. Theorem 2 suggests that the parameter $\Lambda(\cdot)$, although infinite-dimensional, can be treated in the same way as the finite-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Thus, the asymptotic covariance matrix can be estimated by the inverse of the observed information matrix. Specifically, for any constant vector

$(\mathbf{h}_1, \mathbf{h}_2) \in R^{d_1} \times R^{d_2}$ and any bounded function h_3, the asymptotic variance of

$$\mathbf{h}_1^T \widehat{\boldsymbol{\beta}}_n + \mathbf{h}_2^T \widehat{\boldsymbol{\gamma}}_n + \int_0^\tau h_3(t) d\widehat{\Lambda}_n(t) \equiv \mathbf{h}_1^T \widehat{\boldsymbol{\beta}}_n + \mathbf{h}_2^T \widehat{\boldsymbol{\gamma}}_n + \sum_{\Delta_{ij}=1} h_3(Y_{ij}) \widehat{\Lambda}_n\{Y_{ij}\}$$

can be estimated by $\mathbf{h}_n^T \mathbf{J}_n^{-1} \mathbf{h}_n$, where $\mathbf{h}_n$ is the vector comprising of $\mathbf{h}_1$, $\mathbf{h}_2$ and the $h_3(Y_{ij})$ for which $\Delta_{ij} = 1$, and $\mathbf{J}_n$ is the negative Hessian matrix of $\log L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda)$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the jump sizes of Λ at the Y_{ij} for which $\Delta_{ij} = 1$, evaluated at $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, \widehat{\Lambda}_n)$. The next theorem formalizes this approximation.

Theorem 4. *Let* $V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ *be the asymptotic variance of* $n^{1/2}\{\mathbf{h}_1^T(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{h}_2^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) + \int_0^\tau h_3(t) d(\widehat{\Lambda}_n(t) - \Lambda_0(t))\}$. *Under C.1∼C.10,* $n\mathbf{h}_n^T \mathbf{J}_n^{-1} \mathbf{h}_n \to^p V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ *uniformly in* $(\mathbf{h}_1, \mathbf{h}_2, h_3)$ *such that* $\|\mathbf{h}_1\| \le 1, \|\mathbf{h}_2\| \le 1$ *and* $\|h_3\|_V \le 1$, *where* $\|h\|_V$ *denotes the total variation of* $h(t)$ *in* $[0, \tau]$.

4. Simulation Studies

We carried out simulation studies to assess the performance of the proposed inference procedures in finite samples. We set the cluster size to two, and generated failure times from the proportional hazards model with a Gaussian random effect

$$\Lambda(t|\mathbf{X}_{ij}, b_i) = \Lambda_0(t) \exp(\beta_1 X_{1ij} + \beta_2 X_{2ij} + b_i), \quad j = 1, 2; \ i = 1, \ldots, n,$$

where $\Lambda_0(t) = t$, $\beta_1 = 1$, $\beta_2 = -1$, $X_{1i1} \equiv X_{1i2}$ is a dichotomous variable with half of the subjects taking the value 1, X_{ij} is an independent uniform$(0, 1)$ variable, and b_i is normal with mean zero and variance σ^2. The censoring time was set to be the minimum of 3 and a uniform$(0, 4)$ variable, corresponding to an approximate 35% censoring rate. The MLEs of β and σ^2 were obtained via the EM algorithm. The standard error estimates were based on the profile likelihood function at some fixed parameter values around the MLEs. This calculation was done through the the EM algorithm, where these parameters were held fixed in both the E-step and M-step. Then the variance of the MLE was computed by using the numerical difference of the profile log-likelihood function as stated in Theorem 3 for an appropriate choice of h_n. Specifically, we chose $h_n = 5/\sqrt{n}$. The confidence intervals for $\boldsymbol{\beta}$ and σ^2 were based on the normal approximations to $\widehat{\boldsymbol{\beta}}_n$ and $\log \widehat{\sigma}_n^2$, respectively. We considered variance estimation with h_n ranging from $0.1/\sqrt{n}$ to $10/\sqrt{n}$. It turned out that the variance estimation for $\widehat{\beta}_n$ is fairly robust to the choice of h_n, whereas that of $\log \widehat{\sigma}_n^2$ is more sensitive. Murphy et al. (1997) suggested a rule of thumb of $h_n = 1/\sqrt{n}$ or $h_n = |\widehat{\theta}_n| \mathrm{sgn}(\widehat{\theta}_n)/\sqrt{n}$, where $\widehat{\theta}_n$ is the maximum likelihood estimate.

The simulation results with $n = 200$ are summarized in Table 1. These results demonstrate that the proposed methods work well in that the parameter estimators have little bias, the variance estimators are reasonably accurate and the confidence intervals have proper coverage probabilities. Additional simulation studies (results not shown) revealed that the efficiency gains of the proposed MLEs over the estimators of Cai et al. (2002) can be substantial in realistic situations.

Table 1. Simulation results for the proportional hazards model with Gaussian random effects.

		Bias	SE	SEE	95% CP
$\sigma^2 = 1$	β_1	-0.003	0.210	0.201	0.942
	β_2	-0.015	0.267	0.285	0.953
	σ	-0.018	0.151	0.144	0.960
$\sigma^2 = 3$	β_1	-0.010	0.300	0.287	0.946
	β_2	-0.012	0.313	0.328	0.958
	σ	-0.025	0.190	0.180	0.935

Note: Bias and SE are the bias and standard error of the estimator. SEE is the mean of the standard error estimator, and 95% CP is the coverage probability of the 95% confidence interval. Each entry is based on 1,000 simulated data sets.

In related simulation studies, Zeng et al. (2005) generated failure times from proportional odds models with Gaussian random effects. The MLEs were calculated via the optimization search method and the variance estimates were calculated by inverting the observed information matrix. The conclusions are similar.

5. An Example

We now consider the well-known Diabetic Retinopathy Study (Huster, Brookmeyer and Self (1989)). This study was conducted to assess the ability of laser photocoagulation in delaying visual loss among patients with diabetic retinopathy. The subset of the data that has been analyzed extensively in the statistical literature pertains to 197 high-risk patients. For each patient, one eye was randomly selected to receive the laser treatment while the other eye was observed without treatment. The failure time of interest is the time to visual loss as measured by visual acuity less than 5/200. As in the existing literature, we consider three covariates: X_{1ij} indicates, by the values 1 versus 0, whether or not the jth eye ($j = 1$ for the left eye and $j = 2$ for the right eye) of the ith patient was treated with laser photocoagulation, $X_{2i1} \equiv X_{2i2}$ indicates, by the values

1 versus 0, whether the ith patient had adult-onset or juvenile-onset diabetics, and $X_{3ij} \equiv X_{1ij} * X_{2ij}$ is the interaction between X_{1ij} and X_{2ij}. We fit model (1.2) with these three covariates, along with a Gaussian random effect b_i to account for the dependence between the two eyes of the same patient. We consider the transformation $G_0(\cdot)$ from the following class: $\left\{1-(1+\xi x)^{-1/\xi};\ \xi \in [0,1]\right\}$, where $\xi = 0$ corresponds to the proportional hazards model and $\xi = 1$ to the proportional odds model.

We vary the value of ξ from 0 to 1 in 0.1 increments and maximize the corresponding likelihood. It turns out that $\xi = 0.3$ is the best choice in that it yields the maximal value of the observed-data likelihood function. Table 2 summarizes the results under the selected transformation model, as well as the proportional hazards and proportional odds models. There is a high degree of dependence between the two eyes of the same patient in time to visual loss. The treated eye is less likely to suffer visual loss than the untreated eyes, and treatment is more effective for adult-onset diabetics than for juvenile-onset diabetics.

Table 2. Parameter estimates under random-effect transformation models for the diabetic retinopathy study.

	Model		
Parameter	$\xi = 0$	$\xi = 0.3$	$\xi = 1$
β_1	-0.523 (0.231)	-0.564 (0.250)	-0.659 (0.295)
β_2	0.421 (0.264)	0.447 (0.288)	0.496 (0.345)
β_3	-0.999 (0.369)	-1.073 (0.398)	-1.234 (0.466)
σ	1.038 (0.191)	1.114 (0.207)	1.296 (0.251)

Note: Standard error estimates are shown in parentheses.

6. Conclusion

The proposed likelihood-based methods have several advantages over the estimating-equations methods of Cai et al. (2002). First, the proposed estimators are more efficient. Second, it is less time-consuming to evaluate the variances of the proposed estimators than those of Cai et al.'s estimators. Third, the assumption on the independence of the censoring time and failure time required in the Cai et al. approach is avoided in the likelihood approach. Finally, the likelihood approach allows one to use AIC and other likelihood-based criteria for model selection, as demonstrated in the example.

Our experience shows that the algorithms described in Section 2 perform very well when the initial values are chosen appropriately. We recommend setting $\boldsymbol{\beta} = \mathbf{0}$, $\sigma^2 = 1$ and the jump sizes of Λ to $1/n$. The algorithms are quite fast. It took about 10 hours on an IBM BladeCenter HS20 machine to complete all the

simulation studies reported in Table 1. No convergence problem was encountered in any simulation run.

We have found that the estimation of regression parameters is not sensitive to the misspecification of the the random-effects distribution. For example, when we simulated failure times from the proportional hazards gamma frailty model but fitted the data using the proportional hazards model with normal random effect, the estimators of the regression parameters have very little bias and the confidence intervals have reasonable coverage probabilities.

An alternative approach to random-effects models is marginal models. Indeed, Cai, Wei and Wilcox (2000) studied marginal linear transformation models for clustered failure time data. There are several reasons for using random-effects models. First, these models allow one to predict survival experience of a subject given the event history of other members of the same cluster. Second, efficient estimation is possible under these models. Third, the dependence structures can be of scientific interest, especially in genetic studies.

Acknowledgements

This research was supported by the National Institutes of Health Grants R01 CA82659 (D. Zeng and D. Y. Lin) and R01 CA76404 (X. Lin). The authors thank the an associate editor and the referees for their helpful comments.

Appendix

In this appendix, we outline the proofs of the lemmas and theorems. The detailed proofs are given in a supplementary technical report. We introduce some notation. Let $\mathbf{O}_i$ denote the observations in the ith cluster consisting of n_i and $(Y_{ij}, \Delta_{ij}, \overline{\mathbf{X}}_{ij}(Y_{ij}), \overline{\mathbf{Z}}_{ij}(Y_{ij}))$, $j = 1, \ldots, n_i$. Let $\mathcal{P}_n$ and $\mathcal{P}$ be the empirical measure and the expectation of n i.i.d observations $\mathbf{O}_1, \ldots, \mathbf{O}_n$. That is, for any measurable function $g(\mathbf{O})$, $\mathcal{P}_n[g(\mathbf{O})] = n^{-1}\sum_{i=1}^n g(\mathbf{O}_i)$ and $\mathcal{P}[g(\mathbf{O})] = E[g(\mathbf{O})]$.

Proof of Lemma 1. Under C.7, $G_0'(x)^{\Delta_{ij}}(1-G_0(x))^{1-\Delta_{ij}} \le c_1(1+x)^{-(\Delta_{ij}+\rho_0)}$ for some constant c_1. Therefore, the left-hand side of (3.3) is bounded by

$$c_1^{n_i}\exp\Big\{\sum_{j=1}^{n_i}\Delta_{ij}\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}\Big\}\int_{\mathbf{b}}\prod_{j=1}^{n_i}\Big\{1+\int_0^{Y_{ij}}e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t)\Big\}^{-(\Delta_{ij}+\rho_0)}$$
$$\times\exp\Big\{\sum_{j=1}^{n_i}\Delta_{ij}\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}\Big\}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b}).$$

Let M be a constant larger than 1 such that $M^{-1} \le e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}} \le M$ and $M^{-1} \le$

$\|Z_{ij}(t)\| \le M$. Then

$$1 + \int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t) \ge 1 + M^{-1}e^{-M\|\mathbf{b}\|}\Lambda(Y_{ij}) \ge e^{-M\|\mathbf{b}\|}M^{-1}\left\{1+\Lambda(Y_{ij})\right\}.$$

Thus,

$$\begin{aligned}
&\int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{1 + \int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}} d\Lambda(t)\right\}^{-(\Delta_{ij}+\rho_0)} \\
&\quad \times \exp\Big\{\sum_{j=1}^{n_i} \Delta_{ij}\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}\Big\}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b}) \\
&\le \int_{\mathbf{b}} \prod_{j=1}^{n_i} \{1+\Lambda(Y_{ij})\}^{-(\Delta_{ij}+\rho_0)} \\
&\quad \times \exp\Big\{(M\|\mathbf{b}\| + \log M)\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij}) + \sum_{j=1}^{n_i}\Delta_{ij}\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}\Big\}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b}).
\end{aligned}$$

Since $\mathbf{Z}_{ij}$ and $\mathbf{X}_{ij}$ are bounded and $\psi(\mathbf{b};\boldsymbol{\gamma})$ satisfies C.8, (3.3) in Lemma 1 holds for some constant c_0.

Proof of Lemma 2. Suppose that the parameters $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \Lambda^*)$ and $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \Lambda_0)$ yield the same joint density of the data. That is, almost surely,

$$\begin{aligned}
&\int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{\Lambda^{*\prime}(Y_{ij}) e^{\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}^*+\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}} G_0'\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}^*+\mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda^*(s)\Big)\right\}^{\Delta_{ij}} \\
&\qquad \times \left\{1 - G_0\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}^*+\mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda^*(s)\Big)\right\}^{1-\Delta_{ij}} \psi(\mathbf{b};\boldsymbol{\gamma}^*)d\mu(\mathbf{b}) \\
&= \int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{\Lambda_0{}'(Y_{ij}) e^{\mathbf{X}_{ij}(Y_{ij})^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(Y_{ij})^T\mathbf{b}} G_0'\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda_0(s)\Big)\right\}^{\Delta_{ij}} \\
&\qquad \times \left\{1 - G_0\Big(\int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda_0(s)\Big)\right\}^{1-\Delta_{ij}} \psi(\mathbf{b};\boldsymbol{\gamma}_0)d\mu(\mathbf{b}).
\end{aligned}$$

We wish to show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0, \boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda^* = \Lambda_0$. For any fixed $k \le n_i$, we perform the following actions on both sides of the above equality: for $j \le k$, we let $\Delta_{ij} = 1$ and integrate Y_{ij} from 0 to t_j; for $j > k$, if $\Delta_{ij} = 1$, we integrate Y_{ij} from 0 to τ; otherwise, we let $Y_{ij} = \tau$. We then sum over the equalities for all

possible $\{\Delta_{ij} : j = k+1, \ldots, n_i\}$ to obtain

$$\int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ G_0\Big(\int_0^{t_j} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}^* + \mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda^*(s) \Big) \right\} \psi(\mathbf{b};\boldsymbol{\gamma}^*) d\mu(\mathbf{b})$$
$$= \int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ G_0\Big(\int_0^{t_j} e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0 + \mathbf{Z}_{ij}(s)^T\mathbf{b}} d\Lambda_0(s) \Big) \right\} \psi(\mathbf{b};\boldsymbol{\gamma}_0) d\mu(\mathbf{b}). \tag{A.1}$$

It follows from C.9 that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0, \boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda^* = \Lambda_0$.

To prove the second half of the lemma, we suppose that there exists a one-dimensional submodel at the true parameters, denoted by $(\boldsymbol{\beta}_0 + \epsilon\mathbf{h}_1, \boldsymbol{\gamma}_0 + \epsilon\mathbf{h}_2, \Lambda_0 + \int h_3(s) d\Lambda_0(s)), \epsilon \in R$, for which the Fisher information is zero, or equivalently, the score function along this path is zero almost surely. Simple algebraic manipulations yield

$$\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \prod_{j=1}^{n_i} \left\{ \Lambda_0'(Y_{ij}) \right\}^{\Delta_{ij}}$$
$$\times \left[\sum_{j=1}^{n_i} M_{ij}(\mathbf{h}_1, h_3, \mathbf{b}) + \frac{\psi'(\mathbf{b};\boldsymbol{\gamma}_0)^T\mathbf{h}_2}{\psi(\mathbf{b};\boldsymbol{\gamma}_0)} \right] \psi(\mathbf{b};\boldsymbol{\gamma}_0) d\mu(\mathbf{b}) = 0, \tag{A.2}$$

where

$$R_{1k}(\boldsymbol{\beta}, \Lambda, \mathbf{b}) = \prod_{l=1}^{n_k} \left\{ G_0'\Big(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(t)^T\boldsymbol{\beta} + \mathbf{Z}_{kl}(t)^T\mathbf{b}} d\Lambda(s) \Big) e^{\mathbf{X}_{kl}(Y_{kl})^T\boldsymbol{\beta} + \mathbf{Z}_{kl}(Y_{kl})^T\mathbf{b}} \right\}^{\Delta_{kl}}$$
$$\times \left\{ 1 - G_0\Big(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(t)^T\boldsymbol{\beta} + \mathbf{Z}_{kl}(t)^T\mathbf{b}} d\Lambda(t) \Big) \right\}^{1-\Delta_{kl}},$$

$$M_{ij}(\mathbf{h}_1, h_3, \mathbf{b}) = Q_{ij}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \left\{ \int_0^{Y_{ij}} e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}_0 + \mathbf{Z}_{ij}(t)^T\mathbf{b}} (\mathbf{X}_{ij}(t)^T\mathbf{h}_1 + h_3(t)) d\Lambda_0(t) \right\}$$
$$+ \Delta_{ij} \left\{ \mathbf{X}_{ij}(Y_{ij})^T\mathbf{h}_1 + h_3(Y_{ij}) \right\}.$$

We show that (A.2) yields $\mathbf{h}_1 = \mathbf{0}, \mathbf{h}_2 = \mathbf{0}$ and $h_3 = 0$. Fix a k such that $1 \le k \le n_i$ and, for any function of the type $g_1(\Delta_{i1}, Y_{i1}) \ldots g_{n_i}(\Delta_{in_i}, Y_{in_i})$, perform the following action. Partition $\{(\Delta_{ij}, Y_{ij}) : j = 1, \ldots, n_i\}$ into three subsets: for $j \le k$, let $\Delta_{ij} = 1$ and integrate Y_{ij} from 0 to t_j; for $j > k$ and $\Delta_{ij} = 0$, let $Y_{ij} = \tau$; for $j > k$ and $\Delta_{ij} = 1$, integrate Y_{ij} from 0 to τ. Apply this action to the integrand on the left-hand side of (A.2). Then sum over all possible choices of $\Delta_{ij} \in \{0, 1\}$ for $j > k$ and let $t_1 = \ldots = t_{ni} = t$. These calculations yield

$$\int_{\mathbf{b}} \left\{ \prod_{m=1}^{k} a_{im}(t_m; \mathbf{b}) \sum_{m=1}^{k} b_{im}(t_m; \mathbf{b}) + \prod_{m=1}^{k} a_{im}(t_m; \mathbf{b}) \frac{\psi'(\mathbf{b};\boldsymbol{\gamma}_0)^T\mathbf{h}_2}{\psi(\mathbf{b};\boldsymbol{\gamma}_0)} \right\} \psi(\mathbf{b};\boldsymbol{\gamma}_0) d\mu(\mathbf{b})$$
$$= 0,$$

where

$$a_{im}(t;\mathbf{b}) = G_0\Big(\int_0^t e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_0(s)\Big),$$

$$b_{im}(t;\mathbf{b}) = \frac{G_0'(\int_0^t e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_0(s))}{G_0(\int_0^t e^{\mathbf{X}_{ij}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}(s)^T\mathbf{b}}d\Lambda_0(s))} \times \int_0^t (\mathbf{X}_{im}(s)^T\mathbf{h}_1 + h_3(s))e^{\mathbf{X}_{im}(s)^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}(s)^T\mathbf{b}}d\Lambda_0(s).$$

It then follows from C.10 that $\mathbf{h}_2 = \mathbf{0}$ and $b_{im}(t,\mathbf{b}) = 0$. The latter implies that $\mathbf{X}_{ij}(t)^T\mathbf{h}_1 + h_3(t) = 0$. Thus, C.10 yields $\mathbf{h}_1 = \mathbf{0}$ and $h_3 = 0$.

Proof of Lemma 3. Under C.7, $\{G_0'(x)x\}^{\Delta_{ij}}\{1-G_0(x)\}^{1-\Delta_{ij}}$ is bounded by some constant c_1. Thus, (2.1) is bounded from above by

$$c_2\prod_{i=1}^n\int_{\mathbf{b}}\prod_{j=1}^{n_i}\Big\{1-G_0(\int_0^\tau e^{\mathbf{X}_{ij}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{ij}(t)^T\mathbf{b}}d\Lambda(t))\Big\}^{I(Y_{ij}=\tau)}\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b}),$$

where c_2 is some number depending on the observations. On the other hand, C.1 implies that there exists some (i,j) such that $Y_{ij} = \tau$ with probability tending to one. Therefore, at least one integral in the above expression is present, and such an integral is zero if Λ has an infinite jump size for some failure time. Thus the NPMLE exists and $\widehat{\Lambda}_n$ has finite jump sizes.

Proof of Theorem 1. Let Ω be the measurable set in the probability space such that all the conditions hold for any fixed $\omega \in \Omega$. Clearly, $P(\Omega) = 1$. Thus, the following arguments pertain to fixed $\omega \in \Omega$. We use $O(1)$ to denote some positive constant, which may depend on ω but is independent of parameters and sample size. The proof consists of two steps.

Step 1. We prove that $\widehat{\Lambda}_n(t)$ has an upper bound in $[0,\tau]$ with probability one. Write $l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\Lambda) = \log L_n(\boldsymbol{\beta},\boldsymbol{\gamma},\Lambda)$. We prove the boundedness of $\widehat{\Lambda}_n(\cdot)$ by contradiction. Suppose that $\widehat{\Lambda}_n(\tau) \to \infty$. From the compactness of Θ, we also assume that $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\gamma}}_n \to \boldsymbol{\gamma}^*$. The idea of obtaining a contradiction is the following: we first construct a step function $\overline{\Lambda}_n$ with jumps only at the Y_{ij} for which $\Delta_{ij} = 1$ such that $\overline{\Lambda}_n$ is close to the true function Λ_0; then since $(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n)$ maximizes $l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\Lambda)$, it holds that $0 \le \{l_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n) - l_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\overline{\Lambda}_n)\}/n$; finally, we show that if $\widehat{\Lambda}_n(\tau) \to \infty$, the right-hand side of the foregoing inequality will eventually be negative, which yields the contradiction.

By differentiating $l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\Lambda)$ with respect to $\Lambda\{Y_{ij}\}$ and setting it to zero, we see that $\widehat{\Lambda}_n\{Y_{ij}\}$ satisfies

$$\frac{\Delta_{ij}}{\Lambda\{Y_{ij}\}} = \sum_{k=1}^n\left\{\frac{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n,\Lambda,\mathbf{b})R_{2k}(Y_{ij};\widehat{\boldsymbol{\beta}}_n,\Lambda,\mathbf{b})\psi(\mathbf{b};\widehat{\boldsymbol{\gamma}}_n)d\mu(\mathbf{b})}{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n,\Lambda,\mathbf{b})\psi(\mathbf{b};\widehat{\boldsymbol{\gamma}}_n)d\mu(\mathbf{b})}\right\}, \qquad (A.3)$$

where $R_{1k}(\cdot)$ is defined in the proof of Lemma 2, and

$$R_{2k}(t;\boldsymbol{\beta},\Lambda,\mathbf{b}) = \sum_{l=1}^{n_k} I(Y_{kl}\geq t)e^{\mathbf{X}_{kl}(t)^T\boldsymbol{\beta}+\mathbf{Z}_{kl}(t)^T\mathbf{b}} \times\Bigg\{ -\frac{\Delta_{kl}G_0''(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{kl}(s)^T\mathbf{b}}d\Lambda(s))}{G_0'(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{kl}(s)^T\mathbf{b}}d\Lambda(s))} +\frac{(1-\Delta_{kl})G_0'(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{kl}(s)^T\mathbf{b}}d\Lambda(s))}{1-G_0(\int_0^{Y_{kl}} e^{\mathbf{X}_{kl}(s)^T\boldsymbol{\beta}+\mathbf{Z}_{kl}(s)^T\mathbf{b}}d\Lambda(s))}\Bigg\}.$$

In view of (A.3), we construct a step function $\overline{\Lambda}_n(t)$ with jumps only at the Y_{ij} with jump size $\overline{\Lambda}_n\{Y_{ij}\}$ satisfying

$$\frac{\Delta_{ij}}{\overline{\Lambda}_n\{Y_{ij}\}} = \sum_{k=1}^{n}\left\{\frac{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0,\Lambda_0,\mathbf{b})R_{2k}(Y_{ij};\boldsymbol{\beta}_0,\Lambda_0,\mathbf{b})\psi(\mathbf{b};\boldsymbol{\gamma}_0)d\mu(\mathbf{b})}{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0,\Lambda_0,\mathbf{b})\psi(\mathbf{b};\boldsymbol{\gamma}_0)d\mu(\mathbf{b})}\right\}. \tag{A.4}$$

Thus, $\overline{\Lambda}_n(t)=\sum_{i=1}^n\sum_{j=1}^{n_i} I(Y_{ij}\leq t)\overline{\Lambda}_n\{Y_{ij}\}$. By the Glivenko-Cantelli property of the classes R_{1k} and R_{2k} (proved in the appendix of our technical report), we can show that $\overline{\Lambda}_n(t)$ converges uniformly in $[0,\tau]$ to $\Lambda_0(t)$.

Clearly, $n^{-1}l_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n)-n^{-1}l_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\overline{\Lambda}_n)\geq 0$. From the construction of $\overline{\Lambda}_n$ and according to (3.3), this inequality is equivalent to

$$0\leq O(1)+\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{n_i}\log\left\{n\widehat{\Lambda}_n\{Y_{ij}\}\right\}-\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})\log(1+\widehat{\Lambda}_n(Y_{ij})). \tag{A.5}$$

We show that if $\widehat{\Lambda}_n(\tau)\to\infty$, the right-hand side of (A.5) is eventually negative. The proof of the divergence of the right-hand side mimics the arguments of Murphy (1994). Specifically, we consider a partition of $[0,\tau]$ which consists of a sequence $\tau=s_0>\cdots>s_N=0$. Then the right-hand side of (A.5) can be bounded from above by

$$\begin{aligned} &O(1)+\sum_{q=0}^{N}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{n_i} I(Y_{ij}\in[s_{q+1},s_q))\log\left\{n\widehat{\Lambda}\{Y_{ij}\}\right\}\\ &\quad-\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{n_i} I(Y_{ij}=\tau)\rho_0\log(1+\widehat{\Lambda}(\tau))\\ &\quad-\sum_{q=0}^{N}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{n_i} I(Y_{ij}\in[s_{q+1},s_q))\log(1+\widehat{\Lambda}(Y_{ij})), \end{aligned}$$

which is further bounded by

$$
\begin{aligned}
&-\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I(Y_{ij}=\tau)\log(1+\widehat{\Lambda}_n(\tau))\\
&-\Bigg\{\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I(Y_{ij}=\tau)\log(1+\widehat{\Lambda}_n(\tau))\\
&\qquad -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_1,s_0))\log(1+\widehat{\Lambda}_n(\tau))\Bigg\}\\
&-\sum_{q=1}^{N}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\rho_0+\Delta_{ij})I(Y_{ij}\in[s_q,s_{q-1}))\log(1+\widehat{\Lambda}_n(s_q))\\
&\qquad -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{q+1},s_q))\log(1+\widehat{\Lambda}_n(s_q))\Bigg\}+O(1). \qquad \text{(A.6)}
\end{aligned}
$$

Using Murphy's (1994) idea of constructing the partition, we can choose $s_0 > s_1 > s_2 > \ldots > s_N$ such that the first term on the right-hand side of (A.6) diverges to $-\infty$ as $\widehat{\Lambda}_n(\tau) \to \infty$ and the second term and the third term are negative for large n. This contradicts the fact that (A.6) should be non-negative.

Thus we have shown that, with probability one, $\widehat{\Lambda}_n(\tau)$ has an upper bound. By Helly's Selection Theorem, we can assume that $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$, $\widehat{\boldsymbol{\gamma}}_n \to \boldsymbol{\gamma}^*$, and $\widehat{\Lambda}_n$ converges pointwise to some increasing function Λ^*.

Step 2. We show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0, \boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda^*(t) = \Lambda_0(t)$. We consider

$$
\begin{aligned}
0 &\le \frac{1}{n}l_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n)-\frac{1}{n}l_n(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\overline{\Lambda}_n)\\
&=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\int_b R_{1i}(\widehat{\boldsymbol{\beta}}_n,\widehat{\Lambda}_n,\mathbf{b})\psi(\mathbf{b};\widehat{\boldsymbol{\gamma}}_n)d\mu(\mathbf{b})\right\}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\int_b R_{1i}(\boldsymbol{\beta}_0,\overline{\Lambda}_n,\mathbf{b})\psi(\mathbf{b};\boldsymbol{\gamma}_0)d\mu(\mathbf{b})\right\}\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}\log\left[\frac{\widehat{\Lambda}_n\{Y_{ij}\}}{\overline{\Lambda}_n\{Y_{ij}\}}\right]. \qquad \text{(A.7)}
\end{aligned}
$$

Using equations (A.3) and (A.4), we can easily see that $\widehat{\Lambda}_n(t)$ is absolutely continuous with respect to $\overline{\Lambda}_n(t)$, and

$$
\widehat{\Lambda}_n(t)=\int_0^t\frac{\mathcal{P}_n\left[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)\right]}{\left|\mathcal{P}_n\left[\nu(\mathbf{O};\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n,t)\right]\right|}d\overline{\Lambda}_n(t), \qquad \text{(A.8)}
$$

where

$$\nu(\mathbf{O}_k;\boldsymbol{\beta},\boldsymbol{\gamma},\Lambda,t)=\frac{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta},\Lambda,\mathbf{b})R_{2k}(t;\boldsymbol{\beta},\Lambda,\mathbf{b})\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b})}{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta},\Lambda,\mathbf{b})\psi(\mathbf{b};\boldsymbol{\gamma})d\mu(\mathbf{b})}.$$

It follows from the Donsker property proved in the appendix of our technical report that

$$\sup_{t\in[0,\tau]}\left|\mathcal{P}_n[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)]-\mathcal{P}[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)]\right|\to 0,\quad a.s.,$$

$$\sup_{t\in[0,\tau]}\left|\mathcal{P}_n[\nu(\mathbf{O};\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\gamma}}_n,\widehat{\Lambda}_n,t)]-\mathcal{P}[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)]\right|\to 0,\quad a.s..$$

We wish to take limits on both sides of (A.8). We first show that the denominator of the integrand is uniformly bounded away from zero. From (A.8), for any $\epsilon>0$,

$$\limsup_n \widehat{\Lambda}_n(\tau)\geq\int_0^\tau\frac{\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)\right]}{\epsilon+\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|}d\Lambda_0(t).$$

Let $\epsilon\to 0$ and use the Monotone Convergence Theorem to obtain

$$\int_0^\tau\frac{\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)\right]}{\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|}\lambda_0(t)dt<\infty. \tag{A.9}$$

We claim that $\min_{t\in[0,\tau]}\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|>0$. If this inequality does not hold, then there exists some $t^*\in[0,\tau]$ such that $\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t^*)\right]=0$. The function $\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]$ is right-differentiable almost everywhere provided that $\lambda_c(t|\overline{\mathbf{X}}_{ij}(t),\overline{\mathbf{Z}}_{ij}(t))$ exists and is uniformly bounded almost everywhere. Thus, there exists a $\delta>0$ such that for $t\in(t^*,t^*+\delta)$,

$$\begin{aligned}\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|&=\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]-\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|\\&\leq O(1)|t-t^*|\end{aligned}$$

almost everywhere. Thus (A.9) implies $\int_{t^*}^{t^*+\delta}\lambda_0(t)/|t-t^*|dt<\infty$. This is a contradiction.

We can now take the limits on both sides of (A.8) to obtain

$$\Lambda^*(t)=\int_0^t\frac{\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\Lambda_0,t)\right]}{\left|\mathcal{P}\left[\nu(\mathbf{O};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,\Lambda^*,t)\right]\right|}d\Lambda_0(t).$$

We conclude that $\Lambda^*(t)$ is absolutely continuous with respect to $\Lambda_0(t)$, so that $\Lambda^*(t)$ is differentiable with respect to t. In addition, $d\widehat{\Lambda}_n(t)/d\overline{\Lambda}_n(t)$ converges to $d\Lambda^*(t)/d\Lambda_0(t)$ uniformly in t. Let $n \to \infty$ in (A.7). Then we have

$$0 \leq \frac{1}{n} l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, \widehat{\Lambda}_n) - \frac{1}{n} l_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \overline{\Lambda}_n)$$
$$\to E\left\{ \log \frac{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}^*, \Lambda^*, \mathbf{b}) \psi(\mathbf{b}; \boldsymbol{\gamma}^*) d\mu(\mathbf{b}) \prod_{j=1}^{n_i} \Lambda^{*\prime}(Y_{ij})^{\Delta_{ij}}}{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}_0, \Lambda_0, \mathbf{b}) \psi(\mathbf{b}; \boldsymbol{\gamma}_0) d\mu(\mathbf{b}) \prod_{j=1}^{n_i} \Lambda_0{}'(Y_{ij})^{\Delta_{ij}}} \right\},$$

which is the negative Kullback-Leibler information. The identifiability result in Lemma 2 implies that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$, and $\Lambda^* = \Lambda_0$.

Combining the results from Step 1 and Step 2, we conclude that, almost surely,

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0, \quad \|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\| \to 0, \quad |\widehat{\Lambda}_n(y) - \Lambda_0(y)| \to 0, \quad y \in [0, \tau].$$

The uniform convergence of $\widehat{\Lambda}_n$ to Λ_0 follows from the fact that Λ_0 is a continuous function.

Proofs of Theorems 2-4. The proof of the weak convergence of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^T - \boldsymbol{\beta}_0^T, \widehat{\boldsymbol{\gamma}}_n^T - \boldsymbol{\gamma}_0^T, \widehat{\Lambda}_n - \Lambda_0)$ in Theorem 2 makes use of Theorem 3.3.1 of van der Vaart and Wellner (1996). The most difficult part is to verify that the information operator at the true parameters is invertible. This can be done by showing that the information operator is the summation of an invertible operator and a compact operator and that the the information operator is one to one. The former is derived from the explicit expression of the information operator and the latter follows from the fact, shown in Lemma 2, that any submodel has non-singular information. The details can be found in our technical report.

The proof of Theorem 3 proceeds by verifying the conditions of Murphy and van der Vaart (2000). In particular, we can construct an approximate least favorable submodel using the invertibility of the information operator. The no-bias condition along the least favorable submodel follows from the arguments used in proving Theorem 1. The other regularity conditions follow from the Donsker property of appropriate functional classes proved in our technical report.

The proof of Theorem 4 is essentially the same as the that of Theorem 3 of Parner (1998). The main idea is that the empirical information operator based on $\mathbf{J}_n$ approximates the true information operator, so that it is invertible; see our technical report.

References

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Medicine* **2**, 273-277.

Bickel, P. J. (1986). Efficient testing in a class of transformation models. *Papers on Semiparametric Models at the ISI Centenary Session* Amsterdam, 63-81.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Cai, T, Cheng, S. C. and Wei, L. J. (2002). Semiparametric mixed-effects models for clustered failure time data. *J. Amer. Statist. Assoc.* **97**, 514-522.

Cai, T., Wei, L. J. and Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika* **87**, 867-878.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659-668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835-845.

Coleman, T. F. and Li, Y. (1994). On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Math. Program.* **67**, 189-224.

Coleman, T. F. and Li, Y. (1996). An Interior, Trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* **6**, 418-445.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cuzick, J. (1988). Rank regression. *Ann. Statist.* **16**, 1369-1389.

Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in the two-sample generalized odds-rate model. *J. Amer. Statist. Assoc.* **83**, 744-749.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Huster, W. J, Brookmeyer, R. and Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, **45**, 145-156.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004). Robust inference for proportional hazards univariate frailty regression models. *Ann. Statist.* **32**, 1448-1491.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lam, K. F. and Leung, T. L. (2001). Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis* **7**, 39-54.

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

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23**, 182-198.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximal likelihood estimate in the proportional odds model. *J. Amer. Statist. Assoc.* **92**, 968-976.

Murphy, S. A. and van der Vaart, A. W. (2000). On the profile likelihood. *J. Amer. Statist. Assoc.* **95**, 449-465.

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

Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. *Appl. Statist.* **33**, 169-175.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165-177.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, Berlin.

Wu, C. O. (1995). Estimating the real parameter in a two-sample proportional odds model. *Ann. Statist.* **23**, 376-395.

Zeng, D., Lin, D. Y. and Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *J. Amer. Statist. Assoc.* **100**, 470-483.

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420, U.S.A.

E-mail: dzeng@bios.unc.edu

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420, U.S.A.

E-mail: lin@bios.unc.edu

Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.

E-mail: xlin@hsph.harvard.edu

(Received March 2006; accepted August 2006)