Statistica Sinica **30** (2020), 1117-1134
doi:https://doi.org/10.5705/ss.202017.0156

REGRESSION ANALYSIS OF MULTIVARIATE CURRENT STATUS DATA WITH SEMIPARAMETRIC TRANSFORMATION FRAILTY MODELS

Shuwei Li, Tao Hu, Shishun Zhao and Jianguo Sun

Guangzhou University, Capital Normal University,
Jilin University and University of Missouri

Abstract: This study investigates regression analysis of multivariate current status data using a class of flexible semiparametric transformation frailty models. The maximum likelihood estimation procedure is derived for the problem. In particular, a novel EM algorithm, which is quite stable and can be easily implemented, is developed. In addition, the asymptotic properties of the resulting estimators are established, and a numerical study indicates that the proposed methodology works well in practical situations. An application is provided to illustrate the proposed method.

Key words and phrases: EM algorithm, multivariate current status data, semiparametric efficiency, transformation frailty models.

1. Introduction

Current status data, also known as case-1 interval-censored failure time data, occur frequently in many fields, such as demographic investigations, epidemiology studies, and tumorigenicity experiments (Huang (1996); Rossini and Tsiatis (1996); Lin, Oakes and Ying (1998); Martinussen and Scheike (2002); Jewell and van der Laan (2004); Xue, Lam and Li (2004); Sun (2006); Zeng, Mao and Lin (2016)). For such data, each subject in the study is observed only once, and we only know that the failure event of interest occurs either before or after the observation time. In other words, the failure time is either left- or right-censored, and can not be observed exactly. Multivariate current status data mean that the failure time study involves several correlated failure times of interest, and only current status data are available for each of the failure times of interest (Dunson and Dinse (2002); Jewell, van der Laan and Lei (2005); Chen, Tong and Sun (2009)).

A great deal of literature has been developed for regression analysis of univariate and multivariate current status data (Sun (2006)). For the latter, however, most of existing methods apply only to some restricted models or limited situations. For analysis of multivariate failure time data, one of the main challenges is how to deal with the correlation between the correlated failure times. Here, two general approaches are commonly used: marginal model-based methods, and frailty model-based methods. The former leaves the correlation as arbitrary, and treats the failure times of interest as independent, which is often referred to as the working independence assumption (Wei, Lin and Weissfeld (1989); Goggins and Finkelstein (2000); Chen, Tong and Sun (2007)). The advantage of these methods is their simplicity because, for example, the likelihood function and the corresponding estimation procedure can be relatively simple and easily derived. On the other hand, they may not be efficient compared with the frailty model-based methods (Guo and Rodriguez (1992)).

A frailty model-based approach usually tries to model the relationship between the correlated failure times of interest directly using the latent variable or frailty. Among others, Wen and Chen (2011) and Wang, Wang and McMahan (2015) recently proposed such methods for regression analysis of bivariate current status data under the gamma frailty proportional hazards model. The former developed a nonparametric maximum likelihood technique, and the latter employed a spline-based EM algorithm to estimate the parameters of the model. Note that the marginal approach aims to estimate the population-average covariate effect, whereas the frailty approach allows us to estimate subject-specific effects. Furthermore, it is well known that the proportional hazards model may not provide a proper fit. In the following, we develop a frailty model-based approach using a class of flexible semiparametric transformation frailty models. In addition to the differences discussed above between univariate and multivariate current status data, it is clear that the multivariate data also have much more complex structures.

The remainder of the paper is organized as follows. In Section 2, we introduce our notation and the assumptions that will be used throughout the paper. The semiparametric transformation frailty models are then described, along with the resulting likelihood function. Section 3 provides the nonparametric maximum likelihood estimation procedure, which is implemented through a novel EM algorithm involving some Poisson latent variables. In particular, the algorithm employs the probability integral transformation technique and the Gauss-Hermite quadrature method in the E-step. In Section 4, the asymptotic properties of the resulting estimators, including the consistency, asymptotical normality, and semiparametric efficiency, are established. Section 5 presents the results obtained from a simulation study, which suggest that the proposed methodology works well for practical situations. In Section 6, we illustrate the proposed method by means of a real-data example. Section 7 concludes the paper.

2. Notation, Assumptions, and the Likelihood Function

Consider a failure time study that involves n independent subjects, where each subject can experience K possibly correlated failure events of interest. For subject i, let T_{ik} denote the failure time of the kth event, and let X_{ik} be the corresponding d-dimensional vector of covariates, for $i = 1, \ldots, n$. Suppose that for each T_{ik}, only one observation is available at observation time C_{ik}, and we only know that the event occurs either before or after C_{ik}. In other words, T_{ik} is either left- or right-censored at C_{ik}, and the observed data have the form $O_{ik} = \{C_{ik}, \Delta_{ik} = I(T_{ik} \leq C_{ik}), X_{ik}\}$, with $I(\cdot)$ being the indicator function. In the following, we assume that T_{ik} and C_{ik} are conditionally independent, given the covariate X_{ik}.

To describe the covariate effects, we assume that there exists a latent variable b_i and, given X_{ik} and b_i, the cumulative hazard function of T_{ik} has the form

$$G_k\left\{\Lambda_k(t)e^{X_{ik}^T\beta}b_i\right\}, \tag{2.1}$$

where $\Lambda_k(t)$ denotes an unknown baseline cumulative hazard function, β is a d-dimensional vector of regression parameters, and G_k is a prespecified increasing function. Note that many authors, including Dabrowska and Doksum (1988) and Zeng and Lin (2007), have discussed the same or similar models, and it is easy to see that this class of models contains many commonly used models as special cases. For example, by letting $G_k(x) = x$, we obtain the proportional hazards frailty model and we obtain the proportional odds frailty model when $G_k(x) = \log(1 + x)$. Note that in the class of models (2.1), we have assumed that the covariate effects are the same for different failure times for simplicity of presentation. If they are different, we can still apply the methodology proposed below by simply defining a new, larger vector of covariates. In the following, we assume that, given b_i, $T_{i1}, \ldots, T_{iK}$ are independent of each other, and that the b_i follow a parametric model with mean one and density function $p(b_i|\gamma)$, where γ is an unknown parameter. Then, the likelihood function has the form

$$\begin{aligned} L(\beta, \gamma, \Lambda) = \prod_{i=1}^{n} \int_{b_i} \prod_{k=1}^{K} & \left\{1 - \exp\left[-G_k\left\{\Lambda_k(C_{ik})e^{X_{ik}^T\beta}b_i\right\}\right]\right\}^{\Delta_{ik}} \\ & \times \exp\left[-G_k\left\{\Lambda_k(C_{ik})e^{X_{ik}^T\beta}b_i\right\}\right]^{1-\Delta_{ik}} p(b_i|\gamma)\mathrm{d}b_i, \end{aligned}$$

with $\Lambda = (\Lambda_1, \ldots, \Lambda_K)$. Maximizing $L(\beta, \gamma, \Lambda)$ requires some numerical integration. Thus, a direct maximization is quite challenging and unstable, even under a Cox model setting (Wang, Wang and McMahan (2015)). More importantly, the resulting estimators have no closed forms, which naturally suggests the use of the following EM algorithm.

Note that, as discussed by Kosorok, Lee and Fine (2004), in the class of models (2.1), the transformation function $G_k(x)$ can be derived by, or written in the following Laplace transformation form:

$$\exp\{-G_k(x)\} = \int_0^\infty e^{-xt}\phi(t|r_k)\mathrm{d}t,$$

where $\phi(t|r_k)$ is a density function that depends on some constant r_k, with support $[0, \infty)$. An example of $\phi(t|r_k)$ is the gamma density function with mean one and variance r_k, which yields $G_k(x) = \log(1 + r_k x)/r_k$, the logarithmic transformation function. One advantage of the latter form is that we can convert the transformation frailty model into the proportional hazards model with two sets of random effects. Specifically, let μ_{ik} denote the latent variable following the density function $\phi(t|r_k)$. Then, the conditional survival function of T_{ik} can be expressed as

$$S_k(t|X_{ik}, b_i) = \int_{\mu_{ik}} \exp\left[-\mu_{ik}\left\{\Lambda_k(t)e^{X_{ik}^T\beta}b_i\right\}\right]\phi(\mu_{ik}|r_k)\mathrm{d}\mu_{ik},$$

given X_{ik} and b_i. It follows that the likelihood function $L(\beta, \gamma, \Lambda)$ can be rewritten as

$$\begin{aligned} L_1(\beta, \gamma, \Lambda) = \prod_{i=1}^n \int_{b_i} \prod_{k=1}^K \int_{\mu_{ik}} & \left\{1 - \exp\left[-\mu_{ik}\left\{\Lambda_k(C_{ik})e^{X_{ik}^T\beta}b_i\right\}\right]\right\}^{\Delta_{ik}} \\ & \times \exp\left[-\mu_{ik}\left\{\Lambda_k(C_{ik})e^{X_{ik}^T\beta}b_i\right\}\right]^{1-\Delta_{ik}} \phi(\mu_{ik}|r_k)\mathrm{d}\mu_{ik} p(b_i|\gamma)\mathrm{d}b_i. \end{aligned} \quad (2.2)$$

In the next section, we discuss the estimation of (β, γ, Λ), based on $L_1(\beta, \gamma, \Lambda)$ given in (2.2).

3. Maximum Likelihood Estimation

Now, we discuss the estimation of (β, γ, Λ). To do so, we derive the nonparametric maximum likelihood estimation procedure. For each k, let $t_{1k} < \cdots < t_{n_k k}$ denote the distinct ordered observation times of $\{C_{ik}; i = 1, \ldots, n\}$, and assume that Λ_k is a step function with nonnegative jump size λ_{lk} at t_{lk}, for $l = 1, \ldots, n_k$. In other words, we have $\Lambda_k(t) = \sum_{t_{lk} \leq t} \lambda_{lk}$.

$$L_2(\theta) = \prod_{i=1}^{n} \int_{b_i} \prod_{k=1}^{K} \int_{\mu_{ik}} \left\{ 1 - \exp\left[-\mu_{ik} \left(\sum_{t_{lk} \le C_{ik}} \lambda_{lk} \right) e^{X_{ik}^T \beta} b_i \right] \right\}^{\Delta_{ik}}$$
$$\times \exp\left[-\mu_{ik} \left(\sum_{t_{lk} \le C_{ik}} \lambda_{lk} \right) e^{X_{ik}^T \beta} b_i \right]^{1-\Delta_{ik}} \phi(\mu_{ik}|r_k) \mathrm{d}\mu_{ik} p(b_i|\gamma) \mathrm{d}b_i.$$

In the following, we develop an EM algorithm based on a two-stage data augmentation involving Poisson variables.

In the first stage, we assume that the latent variables b_i and μ_{ik} were known. In this case, the likelihood function has the form

$$L_3(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ 1 - \exp\left[-\mu_{ik} \left(\sum_{t_{lk} \le C_{ik}} \lambda_{lk} \right) e^{X_{ik}^T \beta} b_i \right] \right\}^{\Delta_{ik}}$$
$$\times \exp\left[-\mu_{ik} \left(\sum_{t_{lk} \le C_{ik}} \lambda_{lk} \right) e^{X_{ik}^T \beta} b_i \right]^{1-\Delta_{ik}} \phi(\mu_{ik}|r_k) p(b_i|\gamma).$$

In the second stage, define a mapping between Δ_{ik} and a new latent variable Z_{ik} by $\Delta_{ik} = I(Z_{ik} > 0)$, where $Z_{ik} = \sum_{t_{lk} \le C_{ik}} Z_{ilk}$, with Z_{ilk} being independent Poisson random variable with mean $\mu_{ik}\lambda_{lk}e^{X_{ik}^T\beta}b_i$ ($i = 1, \ldots, n$; $k = 1, \ldots, K$; $l = 1, \ldots, n_k$). Hence, if Z_{ilk} were known, we would have the following complete data likelihood function:

$$L_c(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{l=1}^{n_k} \psi(Z_{ilk}|\mu_{ik}\lambda_{lk}e^{X_{ik}^T\beta}b_i)\phi(\mu_{ik}|r_k)p(b_i|\gamma),$$

subject to constraints that $Z_{ik} = \sum_{t_{lk} \le C_{ik}} Z_{ilk} > 0$ if $\Delta_{ik} = 1$, and $Z_{ik} = \sum_{t_{lk} \le C_{ik}} Z_{ilk} = 0$ if $\Delta_{ik} = 0$. Here, $\psi(Z_{ilk}|\mu_{ik}\lambda_{lk}e^{X_{ik}^T\beta}b_i)$ is the probability mass function of Z_{ilk}, with parameter $\mu_{ik}\lambda_{lk}e^{X_{ik}^T\beta}b_i$. Of course, by integrating out the latent variables Z_{ilk}, $L_c(\theta)$ reduces back to $L_3(\theta)$.

Let $\theta^{(m)}$ denote the estimator of θ obtained in the mth iteration. To obtain $\theta^{(m+1)}$, in the E-step, we need to take the logarithm of the complete data likelihood function $L_c(\theta)$, and then calculate the following conditional expectations with respect to all latent variables:

$$\mathrm{E}(\mu_{ik}b_i) = \mathrm{E}_{b_i}\left\{ b_i \frac{\Delta_{ik} - \exp(-G_k(W_{ik}))G_k'(W_{ik})}{\Delta_{ik} - \exp(-G_k(W_{ik}))} \right\},$$
$$\mathrm{E}(Z_{ilk}) = \Delta_{ik}\lambda_{lk}e^{X_{ik}^T\beta}\mathrm{E}_{b_i}\left\{ \frac{b_i}{1 - \exp(-G_k(W_{ik}))} \right\} I(t_{lk} \le C_{ik})$$
$$+ \lambda_{lk}e^{X_{ik}^T\beta}\mathrm{E}(\mu_{ik}b_i)I(t_{lk} > C_{ik}),$$

and

$$
\begin{aligned}
&\mathrm{E}\{h(b_i)\}\\
&= \frac{\int_{b_i} h(b_i) \prod_{k=1}^{K} \{1-\exp(-G_k(W_{ik}))\}^{\Delta_{ik}} \exp\{-G_k(W_{ik})\}^{1-\Delta_{ik}} p(b_i|\gamma)\mathrm{d}b_i}{\int_{b_i} \prod_{k=1}^{K} \{1-\exp(-G_k(W_{ik}))\}^{\Delta_{ik}} \exp\{-G_k(W_{ik})\}^{1-\Delta_{ik}} p(b_i|\gamma)\mathrm{d}b_i}.
\end{aligned}
$$

In the above, $h(b_i)$ is an arbitrary function of b_i, $W_{ik} = \sum_{t_{lk} \le C_{ik}} \lambda_{lk} e^{X_{ik}^T \beta} b_i$, and

$$
G'(W_{ik}) = \frac{\int_{\mu_{ik}} \mu_{ik} e^{-W_{ik}\mu_{ik}} \phi(\mu_{ik}|r_k)\mathrm{d}\mu_{ik}}{\exp\{-G_k(W_{ik})\}}.
$$

For notational simplicity, in the above, we have suppressed the conditional arguments in all conditional expectations. In addition, note that if $\phi(\mu_{ik}|r_k)$ is the gamma density function, the integration above with respect to μ_{ik} has the following closed form:

$$
\int_{\mu_{ik}} \mu_{ik} e^{-W_{ik}\mu_{ik}} \phi(\mu_{ik}|r_k)\mathrm{d}\mu_{ik} = (r_k W_{ik} + 1)^{-r_k^{-1}-1}.
$$

Otherwise we suggest employing the Gauss-Laguerre quadrature technique to calculate the integration with respect to μ_{ik}. To determine $\mathrm{E}\{h(b_i)\}$, we suggest employing the probability integral transformation technique to transform b_i into a standard normal random variable, and then adopting the Gauss-Hermite quadrature method. The numerical study below suggests that the joint use of the probability integral transformation and the Gauss-Hermite quadrature performs well in practice. Nelson et al. (2006) provides a detailed discussion on the probability integral transformation when the random effects or frailties follow nonnormal distributions.

In the M-step, we need to maximize the following objective function with respect to θ:

$$
\begin{aligned}
Q(\theta,\theta^{(m)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{n_k} \Big\{ X_{ik}^T \beta \mathrm{E}(Z_{ilk}) + \log(\lambda_{lk})\mathrm{E}(Z_{ilk}) - \lambda_{lk} e^{X_{ik}^T\beta} \mathrm{E}(\mu_{ik} b_i) \Big\}\\
&+ \sum_{i=1}^{n} \mathrm{E}\{\log(p(b_i|\gamma))\}.
\end{aligned}
$$

Setting $\partial Q(\theta,\theta^{(m)})/\partial\lambda_{lk} = 0$, we can update λ_{lk} with the following closed-form expression

$$
\lambda_{lk} = \frac{\sum_{i=1}^{n} \mathrm{E}(Z_{ilk})}{\sum_{i=1}^{n} \mathrm{E}(\mu_{ik} b_i) e^{X_{ik}^T \beta}}, k = 1, \ldots, K, l = 1, \ldots, n_k. \tag{3.1}
$$

By substituting the estimators above into $Q(\theta,\theta^{(m)})$, we obtain the score equations for β as

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\left(\sum_{l=1}^{n_k}\mathrm{E}(Z_{ilk})\right)\left(X_{ik}-\frac{\sum_{i=1}^{n}\mathrm{E}(\mu_{ik}b_i)e^{X_{ik}^T\beta}X_{ik}}{\sum_{i=1}^{n}\mathrm{E}(\mu_{ik}b_i)e^{X_{ik}^T\beta}}\right)\right\}=0. \qquad (3.2)$$

Finally, by setting $\partial Q(\theta,\theta^{(m)})/\partial\gamma=0$, the estimator of γ can be obtained by solving the score equation

$$\sum_{i=1}^{n}\frac{\partial \mathrm{E}\{\log(p(b_i|\gamma))\}}{\partial\gamma}=0.$$

In summary, combining the above steps, the EM algorithm is given as follows:
Step 0. Choose an initial estimator $\theta^{(0)}$.
Step 1. At the $(m+1)$th iteration, first calculate the conditional expectations $\mathrm{E}(\mu_{ik}b_i)$, $\mathrm{E}(Z_{ilk})$, and $\mathrm{E}\{h(b_i)\}$ at $\theta^{(m)}$.
Step 2. Update $\beta^{(m+1)}$ by solving equation (3.2) using the one step Newton-Raphson method.
Step 3. Obtain $\lambda_{lk}^{(m+1)}$ from expression (3.1).
Step 4. Calculate $\gamma^{(m+1)}$ by solving $\sum_{i=1}^{n}\partial\mathrm{E}\{\log(p(b_i|\gamma))\}/\partial\gamma=0$.
Step 5. Repeat Steps 1-4 until convergence is achieved.

In the above estimation procedure, we have assumed that r_k is known, because it is usually unidentifiable without other assumptions or extra data (Zeng and Lin (2007)). In practice, a common way of determining it is to try different values, and then to select the best or optimal value based on some criterion, such as the maximum likelihood principle. This is discussed in further detail below.

4. Asymptotic Properties

Let $\zeta=(\beta^T,\gamma)^T$, and let $\zeta_0=(\beta_0^T,\gamma_0)^T$ and $\theta_0=(\zeta_0^T,\Lambda_{10},\ldots,\Lambda_{K0})$ denote the true values of ζ and θ, respectively. In addition, let $\hat{\theta}_n=(\hat{\zeta}_n^T,\hat{\Lambda}_{1n},\ldots,\hat{\Lambda}_{Kn})$ denote the maximum likelihood estimator of θ defined in the previous section. In the following, we establish the asymptotic properties of $\hat{\theta}_n$. To do so, we first present some needed regularity conditions.

(A1) The true value ζ_0 belongs to a known compact set $\mathcal{A}\otimes\mathcal{B}$ in R^{d+1}. In addition, for given covariates, each examination time C_k has a continuous conditional density function with support $[\tau_1,\tau_2]$, and the true value $\Lambda_{k0}(\cdot)$ is continuously differentiable with positive derivatives in $[\tau_1,\tau_2]$, with $M^{-1}<\Lambda_{k0}(\tau_1)<\Lambda_{k0}(\tau_2)<M$, for $k=1,\ldots,K$, where M is a large positive constant.

(A2) The covariate vector X_k is bounded.

(A3) The transformation function G_k is twice continuously differentiable on $[0,\infty)$, with $G_k(0)=0$, $G_k'(x)>0$, and $G_k(\infty)=\infty$.

(A4) For any smooth function $g(\cdot)$, we have $\sup_{\gamma\in\mathcal{C}}\int_b g(b)p^{(j)}(b|\gamma)\mathrm{d}b<\infty$, for $j=0,1,2$, where $p^{(j)}(b|\gamma)$ denotes the jth derivative of $p(b|\gamma)$ with respect to γ.

(A5) There exist $c_1,\ldots,c_K\in[\tau_1,\tau_2]$, for which there are $d+K+1$ values of $(\Delta_1,\ldots,\Delta_K,X_1,\ldots,X_K)$, such that if

$$\left(u^T\frac{\partial}{\partial\zeta}+\sum_{k=1}^{K}v_k\frac{\partial}{\partial y_k}\right)\Bigg|_{(\zeta,y_1,\ldots,y_K)=(\zeta_0,\Lambda_{10}(c_1),\ldots,\Lambda_{K0}(c_K))}$$
$$\log\int_b\prod_{k=1}^{K}\left\{\Delta_k+(-1)^{\Delta_k}\exp\left[-G_k\left(y_ke^{X_k^T\beta}b\right)\right]\right\}p(b|\gamma)\mathrm{d}b=0$$

for each of these values, then $u=0_{(d+1)\times1}$ and $v_k=0$. Here, $0_{(d+1)\times1}$ denotes a $(d+1)$-dimensional vector of zeros.

The conditions above are mild and can be satisfied in practical situations. Conditions (A1) and (A2) are standard conditions in survival analysis. Condition (A3) pertains to the transformation function, and it is easy to check that it holds for the logarithmic transformation function $G_r(x)=r^{-1}\log(1+rx)$ $(r\geq 0)$ among others. In addition, condition (A4) is often required for modeling multivariate data with frailty models, and condition (A5) is needed for the identifiability of the model (Chang, Wen and Wu (2007)). Now, we are ready to present the asymptotic properties of $\hat{\theta}_n$. In the following, let $\|\cdot\|$ be the Euclidean norm, and for a function f and a random variable X with distribution P, define $\mathbb{P}f=\int f(x)\mathrm{d}P(x)$ and $\mathbb{P}_nf=n^{-1}\sum_{i=1}^{n}f(X_i)$.

Theorem 1. *Suppose that conditions* (A1)–(A5) *hold. Then, we have that as* $n\to\infty$, $\|\hat{\zeta}_n-\zeta_0\|\to 0$ *and* $\sum_{k=1}^{K}\sup_{t\in[\tau_1,\tau_2]}|\hat{\Lambda}_{kn}(t)-\Lambda_{k0}(t)|\to 0$ *in probability.*

Theorem 2. *Suppose that conditions* (A1)–(A5) *hold. Then, we have that as* $n\to\infty$, $d(\hat{\theta}_n,\theta_0)=\left\{\|\hat{\zeta}_n-\zeta_0\|^2+\sum_{k=1}^{K}\int[\hat{\Lambda}_{kn}(c)-\Lambda_{k0}(c)]^2f_k(c)\mathrm{d}c\right\}^{1/2}=O_p(n^{-1/3})$, *where* $f_k(c)$ *denotes the density of* C_k.

Theorem 3. *Suppose that conditions* (A1)–(A5) *hold. Then, we have that as* $n\to\infty$, $\sqrt{n}(\hat{\zeta}_n-\zeta_0)\xrightarrow{d}N(0,I_0^{-1})$, *where* $I_0=\mathbb{P}\{\tilde{l}(\theta_0)\tilde{l}(\theta_0)^T\}$ *with* $\tilde{l}(\theta_0)$, *given in the online supplementary Material, denoting the efficient score for* ζ *at* θ_0.

These theorems states that the maximum likelihood estimator $\hat{\zeta}_n$ is asymptotically efficient, and that the estimator $\hat{\Lambda}_{kn}$ only has a $n^{-1/3}$ convergence rate. The proofs for these results are provided in the online Supplementary Material. To make inferences about $\hat{\beta}_n$ and $\hat{\gamma}_n$ based on the theorems above, we need to estimate the asymptotic covariance matrix of the corresponding estimators. Because it would be very difficult to derive a consistent estimator of I_0^{-1}, we suggest employing the nonparametric bootstrap method (Efron (1981); Su and Wang (2016)), as follows. Let Q be an integer, and for $1 \leq q \leq Q$, draw a new data set, denoted by of $Q^{(q)}$, of sample size n, with replacement, from the original observed data $O = (O_i = (O_{i1}, \ldots, O_{iK}); i = 1, \ldots, n)$. Let $\hat{\beta}_n^{(q)}$ and $\hat{\gamma}_n^{(q)}$ denote the maximum likelihood estimators of β_0 and γ_0, respectively, defined above based on the bootstrap sample $Q^{(q)}$. Then, we can estimate the covariance matrix and variance of $\hat{\beta}_n$ and $\hat{\gamma}_n$ using the sample covariance matrix and variance of $(\hat{\beta}_n^{(1)}, \ldots, \hat{\beta}_n^{(Q)})$ and $(\hat{\gamma}_n^{(1)}, \ldots, \hat{\gamma}_n^{(Q)})$, respectively. The numerical studies below indicate that this method works well for practical situations.

5. A Simulation Study

In this section, we report the results obtained from an extensive simulation study, performed to investigate the finite-sample performance of the proposed method. In the study, we considered $K = 2$ correlated failure times and, for simplicity, we assumed that $C_{i1} = C_{i2}$, with observation times generated from the uniform distribution over $(3, 5)$. We further assumed that $X_{i1} = X_{i2}$, and that there exist two covariates, with the first covariate generated from a Bernoulli distribution with the success probability of 0.5, and the second covariate following the uniform distribution over $(0, 1)$. To generate the failure times, we took G_k as the logarithmic transformation function, and let $\Lambda_k(t) = 0.05t^2$. Then, we supposed that the latent variable b_i follows a log-normal distribution with mean one and variance γ_0^2, or a gamma distribution with mean one and variance γ_0. The results given below are based on 1,000 replications, with $Q = 50$ and $n = 200$ or 400.

Table 1 presents the results for the estimation of β and γ with $(\beta_{10}, \beta_{20}) = (0, 0.5)$ or $(0.5, -0.5)$ and $\gamma_0 = 1$. They include the estimated bias (Bias) given by the average of the estimates minus the true value, sample standard error (SSE) of the estimates, average of the standard error estimates (SEE), and 95% empirical coverage probability (CP). The table shows that the proposed maximum likelihood estimators seem to be unbiased and the bootstrap variance estimates are appropriate. In addition, the normal approximation to the distribution of the

Table 1. Estimation of regression and variance parameters with the log-normal latent variable distribution.

n	(β_{10}, β_{20})	Par	Bias	SSE	SEE	CP	(β_{10}, β_{20})	Par	Bias	SSE	SEE	CP
$G_1(x) = x$ & $G_2(x) = x$												
200	(0, 0.5)	β_1	-0.011	0.197	0.203	97.8	(0.5, -0.5)	β_1	-0.014	0.224	0.216	94.8
		β_2	-0.012	0.346	0.346	95.2		β_2	-0.016	0.366	0.363	94.9
		γ	0.022	0.174	0.170	94.9		γ	0.021	0.186	0.173	94.0
400	(0, 0.5)	β_1	0.006	0.150	0.150	95.0	(0.5, -0.5)	β_1	0.007	0.157	0.168	96.0
		β_2	-0.010	0.284	0.265	94.6		β_2	-0.012	0.286	0.285	95.0
		γ	0.012	0.136	0.130	95.0		γ	0.005	0.133	0.132	95.0
$G_1(x) = 2\log(1 + x/2)$ & $G_2(x) = 2\log(1 + x/2)$												
200	(0, 0.5)	β_1	-0.005	0.245	0.250	95.8	(0.5, -0.5)	β_1	0.009	0.273	0.269	95.0
		β_2	0.021	0.441	0.446	95.1		β_2	0.024	0.419	0.432	96.8
		γ	0.030	0.245	0.260	95.2		γ	0.025	0.244	0.248	95.2
400	(0, 0.5)	β_1	-0.004	0.169	0.172	95.2	(0.5, -0.5)	β_1	0.006	0.185	0.182	94.9
		β_2	0.009	0.310	0.311	95.2		β_2	0.014	0.305	0.300	95.2
		γ	0.023	0.177	0.177	95.0		γ	0.017	0.170	0.169	94.9
$G_1(x) = 2\log(1 + x/2)$ & $G_2(x) = \log(1 + x)$												
200	(0, 0.5)	β_1	-0.010	0.246	0.260	96.4	(0.5, -0.5)	β_1	0.012	0.269	0.270	95.1
		β_2	0.018	0.452	0.452	95.6		β_2	0.016	0.457	0.458	95.1
		γ	0.033	0.262	0.281	95.8		γ	0.024	0.264	0.266	95.2
400	(0, 0.5)	β_1	-0.009	0.175	0.176	95.1	(0.5, -0.5)	β_1	0.011	0.188	0.185	94.4
		β_2	0.011	0.311	0.317	95.2		β_2	-0.013	0.300	0.303	95.0
		γ	0.021	0.180	0.185	95.6		γ	0.016	0.184	0.188	95.2
$G_1(x) = \log(1 + x)$ & $G_2(x) = \log(1 + x)$												
200	(0, 0.5)	β_1	0.005	0.279	0.280	95.3	(0.5, -0.5)	β_1	-0.010	0.276	0.287	95.0
		β_2	0.022	0.486	0.490	96.2		β_2	-0.023	0.486	0.485	95.2
		γ	0.028	0.275	0.299	96.1		γ	0.034	0.255	0.256	94.6
400	(0, 0.5)	β_1	-0.003	0.188	0.187	95.6	(0.5, -0.5)	β_1	0.009	0.196	0.196	95.0
		β_2	0.001	0.341	0.342	94.0		β_2	0.014	0.346	0.347	95.1
		γ	0.026	0.197	0.200	94.8		γ	0.023	0.191	0.190	94.8

estimators appears to be reasonable and, as expected, the results become better when the sample size increases. Furthermore, the estimation procedure seems to give similar performance for different G_k. We also considered other setups, including different assumed functions for $\Lambda_1(t)$ and $\Lambda_2(t)$, other types of G_k, and different distribution functions for b_i. The results remained similar.

Note that the proposed method assumes that the distribution of b_i is known. Thus, a question of interest is how robust the estimation procedure is to the misspecification of the latent variable distribution. To assess this, we repeated the study in which we generated b_i from a gamma distribution with mean one and variance one, but assumed that they followed a log-normal distribution. Table

Table 2. Estimation of regression parameters with the misspecified latent variable distribution.

(β_{10}, β_{20})	Par	Bias	SSE	SEE	CP	(β_{10}, β_{20})	Par	Bias	SSE	SEE	CP
					$G_1(x) = x$ & $G_2(x) = x$						
(0, 0.5)	β_1	-0.015	0.224	0.226	95.6	(0.5, -0.5)	β_1	0.035	0.253	0.252	95.8
	β_2	-0.016	0.396	0.395	95.2		β_2	0.031	0.436	0.428	94.8
					$G_1(x) = 2\log(1 + x/2)$ & $G_2(x) = 2\log(1 + x/2)$						
(0, 0.5)	β_1	-0.006	0.280	0.278	95.2	(0.5, -0.5)	β_1	0.028	0.288	0.290	94.8
	β_2	0.015	0.501	0.522	96.8		β_2	-0.019	0.476	0.480	96.8
					$G_1(x) = 2\log(1 + x/2)$ & $G_2(x) = \log(1 + x)$						
(0, 0.5)	β_1	0.014	0.281	0.286	95.2	(0.5, -0.5)	β_1	0.020	0.304	0.304	96.0
	β_2	0.022	0.525	0.541	96.6		β_2	-0.018	0.476	0.497	96.0
					$G_1(x) = \log(1 + x)$ & $G_2(x) = \log(1 + x)$						
(0, 0.5)	β_1	0.011	0.282	0.297	95.4	(0.5, -0.5)	β_1	0.019	0.304	0.309	95.0
	β_2	0.022	0.562	0.569	96.2		β_2	-0.025	0.512	0.517	94.8

2 gives the estimation results on β with $n = 200$; the other model specifications are the same as those in Table 1. As in Table 1, the proposed estimators seem to perform well, and the results suggest that the estimation procedure is robust with respect to the latent variable distribution. For the question posed above, we also studied situations in which b_i was generated from a log-normal distribution, but wrongly assumed to be from a gamma distribution; again, the results were similar.

In the simulation study, we also compared the proposed method to that of Wang, Wang and McMahan (2015), who discussed regression analysis of bivariate current status data, with a special case of transformation models $G_k\{\Lambda_k(t)e^{X_{ik}^T\beta^{(k)}} b_i\}$, with $G_1(x) = G_2(x) = x$ and b_i following a gamma distribution. For comparison, we repeated the study shown in Table 1, but with b_i generated from a gamma distribution with both mean and variance one, and one common covariate following a Bernoulli distribution. The results are shown in Table 3. Note that we only considered the estimated bias and sample standard error for the estimation of $\beta = (\beta^{(1)}, \beta^{(2)})^T$ with $n = 200$, various combinations of true values for the regression parameters $(\beta_0^{(1)}, \beta_0^{(2)})^T$, and $\gamma_0 = 0.5$, 1, or 1.5. In addition, we assumed that the covariate effect may be different for two different failure times and, as mentioned above, the proposed method can be applied to this situation. It is apparent that the two methods give similar results and, in particular, exhibit similar efficiency.

Table 3. Comparison of the proposed estimator with that of Wang, Wang and McMahan (2015).

			Proposed method		Wang, Wang and McMahan (2015)	
γ_0	$(\beta_0^{(1)}, \beta_0^{(2)})$	Par	Bias	SSE	Bias	SSE
0.5	(0, 0.5)	$\beta^{(1)}$	−0.003	0.250	−0.002	0.250
		$\beta^{(2)}$	0.022	0.256	0.025	0.257
		γ	0.009	0.215	0.026	0.232
	(0.5, −0.5)	$\beta^{(1)}$	0.004	0.263	0.006	0.263
		$\beta^{(2)}$	−0.026	0.278	−0.025	0.277
		γ	0.022	0.236	0.042	0.253
1	(0, 0.5)	$\beta^{(1)}$	0.001	0.300	−0.002	0.299
		$\beta^{(2)}$	0.027	0.310	0.029	0.313
		γ	0.035	0.292	0.045	0.328
	(0.5, −0.5)	$\beta^{(1)}$	−0.009	0.320	−0.011	0.322
		$\beta^{(2)}$	0.018	0.291	0.019	0.293
		γ	0.042	0.397	0.040	0.420
1.5	(0, 0.5)	$\beta^{(1)}$	0.004	0.343	0.005	0.345
		$\beta^{(2)}$	0.017	0.355	0.019	0.354
		γ	0.032	0.459	0.100	0.472
	(0.5, −0.5)	$\beta^{(1)}$	0.003	0.351	0.001	0.343
		$\beta^{(2)}$	−0.029	0.345	−0.035	0.351
		γ	0.024	0.392	0.083	0.461

6. An Illustration

In this section, we apply the proposed methodology to a set of real bivariate current status data from the Infertility Prevention Project, which was designed as screening test for subjects at risk, in order to asses the prevalence of chlamydia and gonorrhea throughout the United States. Chlamydia and gonorrhea are sexually transmitted diseases that can frequently coexist and can lead to complicated clinical syndromes if left untreated. The data set consists of 5,879 subjects in Nebraska whose urine specimens were collected during the individuals' visits to health clinics in 2008, and then sent to the Nebraska Public Health Laboratory (NPHL) to test the infection status for both diseases. For the data, the overall prevalence of chlamydia and gonorrhea is approximately 0.083 and 0.017, respectively, and the factors or covariates of interest include the patient's gender, whether the patient is Caucasian, and whether the patient presents any symptoms at the time of test. For the analysis, as in many epidemiological surveys, we focus on the ages of the subjects at infection, with the age at the test serving as the observation time. In other words, we have $C_{i1} = C_{i2}$ for the data.

Table 4. Results on the analysis of the chlamydia and gonorrhea data, assuming the same covariate effects.

Frailty		PH frailty model			PO frailty model			Optimal model		
distribution		Est	Std	p-value	Est	Std	p-value	Est	Std	p-value
Log-normal	Gender	0.097	0.096	0.311	0.107	0.111	0.334	0.119	0.085	0.165
	C-America	−0.749	0.096	<0.001	−0.786	0.094	<0.001	−0.826	0.105	<0.001
	Symptoms	0.507	0.121	<0.001	0.549	0.144	<0.001	0.593	0.123	<0.001
Gamma	Gender	0.101	0.104	0.332	0.112	0.107	0.293	0.115	0.083	0.166
	C-America	−0.756	0.099	<0.001	−0.797	0.101	<0.001	−0.810	0.097	<0.001
	Symptoms	0.513	0.128	<0.001	0.560	0.127	<0.001	0.574	0.147	<0.001

Table 5. Results on the analysis of the chlamydia and gonorrhea data, assuming different covariate effects.

Frailty		PH frailty model			PO frailty model			Optimal model		
distribution		Est	Std	p-value	Est	Std	p-value	Est	Std	p-value
Log-normal					Chlamydia					
	Gender	0.060	0.117	0.611	0.065	0.088	0.460	0.073	0.117	0.532
	C-America	−0.619	0.108	<0.001	−0.650	0.097	<0.001	−0.692	0.113	<0.001
	Symptoms	0.296	0.125	0.017	0.314	0.111	0.005	0.340	0.148	0.021
					Gonorrhea					
	Gender	0.318	0.189	0.091	0.335	0.190	0.078	0.363	0.206	0.079
	C-America	−1.577	0.303	<0.001	−1.599	0.346	<0.001	−1.636	0.298	<0.001
	Symptoms	0.500	0.170	<0.001	1.346	0.203	<0.001	1.402	0.243	<0.001
Gamma					Chlamydia					
	Gender	0.063	0.088	0.476	0.062	0.093	0.505	0.061	0.101	0.545
	C-America	−0.626	0.104	<0.001	−0.622	0.112	<0.001	−0.618	0.082	<0.001
	Symptoms	0.301	0.130	0.020	0.296	0.130	0.023	0.293	0.142	0.038
					Gonorrhea					
	Gender	0.328	0.223	0.143	0.325	0.214	0.129	0.323	0.217	0.137
	C-America	−1.582	0.261	<0.001	−1.578	0.266	<0.001	−1.574	0.304	<0.001
	Symptoms	1.322	0.217	<0.001	1.316	0.217	<0.001	1.313	0.195	<0.001

To apply the proposed method, let T_1 denote the age of chlamydia infection and T_2 be the age of gonorrhea infection. In addition, let Gender (1 for male, and 0 for female), C-America (1 for yes, and 0 for no) and Symptoms (1 for yes, and 0 for no) represent the three covariates described above. As in the simulation study, we employed both log-normal and gamma frailty distributions for the latent variable b_i and the logarithmic transformation function. We considered equally spaced grid points of r_1 and r_2, ranging from 0 to 3 with increments of 0.1 for the transformation functions. Then the maximum likelihood principle was used to select the optimal model. According to the analysis, under the log-normal and gamma frailty distributions, the optimal models were given by $(r_1, r_2) = (2.1, 2.6)$ and $(r_1, r_2) = (1.3, 1.5)$, respectively. Table 4 gives the esti-

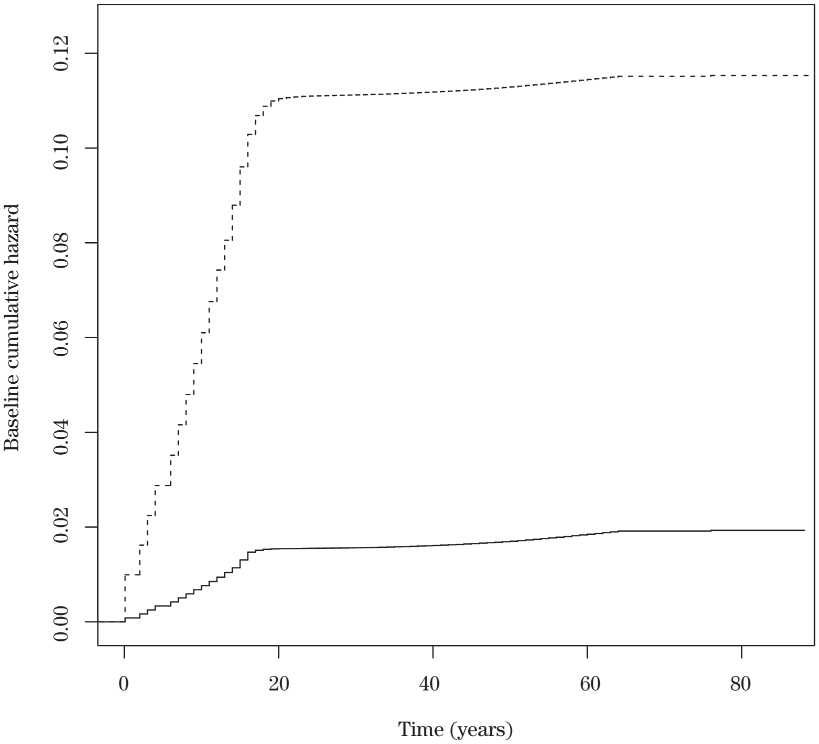

Figure 1. The estimated baseline cumulative hazard functions for the chlamydia infection (upper step function) versus the gonorrhea infection.

mated covariate effects obtained under the optimal model. For comparison, we also include the corresponding results obtained under the proportional hazards (PH) frailty model and the proportional odds (PO) frailty model. The results include the point estimates, estimated standard errors, and p-values for testing the no-covariate effect for each of the three covariates.

First, Table 4 shows that the results are quite consistent across the three models and the two frailty distributions. They all suggest that Caucasians have a significantly lower risk of being infected by chlamydia or gonorrhea than other races. Furthermore, the patients with symptoms are more likely to develop infections than those without the symptoms. However, the risk of developing chlamydia and gonorrhea infections does not seem to be significantly related to the gender of the patients. In addition, under the optimal model with the log-normal frailty distribution, we obtain $\hat{\gamma}_n = 0.518$, indicating that some correlation may exist between the two failure events. Similar results were obtained under the other models and frailty distribution.

Note that, as mentioned above, with the class of models (2.1), we assumed that the three covariates have the same effects on the two risks. As suggested by a reviewer, we repeated the analysis by assuming that the effects may vary; the results are shown in Table 5. It is apparent that the overall results and conclusions are similar to those given above. On the other hand, it is clear that the covariates Caucasian and Symptoms have much stronger effects on the risk of gonorrhea infection than on the risk of chlamydia infection. To further see this, Figure 1 shows the estimates of the baseline cumulative hazard functions for chlamydia and gonorrhea infections under the optimal model with the log-normal frailty. The results seem to indicate that the two baseline hazards are quite different. Note that in the above analysis, we set the bootstrap sample size to $Q = 50$, as in the simulation study. We also considered other values for Q, and obtained similar results.

7. Conclusion

This study examines regression analysis of multivariate current status data under a class of flexible semiparametric transformation frailty models, which includes many existing models as special cases. For inference purpose, a non-parametric maximum likelihood procedure is derived, and a novel EM algorithm is developed using some Poisson random variables to implement the procedure in an easy way. In addition, the asymptotic properties of the resulting estimators are established and, in particular, the estimators of regression parameters are shown to be efficient. In addition, a numerical study shows that the proposed methodology works well in practical situations.

Note that one of the distinct features of the proposed EM algorithm is the joint use of the probability integral transformation technique and the Gauss–Hermite quadrature method, which allows us to easily calculate the conditional expectations for various frailty distributions. The use of Poisson variables allows us to calculate the high-dimensional parameters for the cumulative hazard function explicitly, and to update the low-dimensional parameters β and γ using one-step Newton-Raphson method, separately. As a result, this avoids the inversion of the possibly high-dimensional matrix, making the estimating procedure computationally stable.

The focus of the discussion presented here has been on time-independent covariates. However, it is apparent that there may exist time-dependent covariates. It is straightforward to generalize the idea and method discussed here to the latter situation, although we would need to reformulate the class of models (2.1). In addition, we have assumed that r_k is known, thus, it would be helpful to develop simultaneous estimation procedures. However, as mentioned above, this is usually not possible without additional assumptions or information, and we can employ some selection criterion for their determination in practice. Note that, according to our simulation study, a misspecification of r_k could cause mild bias, especially for the estimation of γ.

Supplementary Material

The online Supplementary Material contains the proofs for Theorems 1, 2, and 3.

Acknowledgements

The authors wish to thank the Co-Editor, Dr. Zhiliang Ying, an Associate Editor, and two reviewers for their helpful comments and suggestions. This work was partly supported by the National Nature Science Foundation of China, grants 11671274, 11731011, and 11671168, the Support Project of High-level Teachers in Beijing Municipal Universities in the Period of the 13th Five-year Plan, grant CIT & TCD 201804078, the Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds, grant 025185305000/204, the Youth Innovative Research Team of Capital Normal University, and the Science and Technology Developing Plan of Jilin Province, grant 20170101061JC.

References

Chang, I. -S., Wen, C. -C. and Wu, Y. -J. (2007). A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica* **17**, 1023–1046.

Chen, M. H., Tong, X. W. and Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statistics in Medicine* **26**, 5147–5161.

Chen, M. H., Tong, X. W. and Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine* **28**, 3424–3436.

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1–23.

Dunson, D. B. and Dinse, G. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics* **58**, 79–88.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.

Guo, G. and Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala.

Journal of the American Statistical Association **87**, 969–976.

Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56**, 940–943.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *The Annals of Statistics* **24**, 540–568.

Jewell, N. P. and van der Laan, M. J. (2004). Current status data: review, recent developments and open problems. *Advances in Survival Analysis* **23**, 625–642.

Jewell, N. P., van der Laan, M. J. and Lei, X. (2005). Bivariate current status data with univariate monitoring times. *Biometrika* **92**, 847–862.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics* **32**, 1448–1491.

Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–298.

Martinussen, J. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* **89**, 649–658.

Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M., and Strawderman, R.(2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics* **15**, 39–57.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91**, 713–721.

Su, Y.-R. and Wang, J. -L. (2016). Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data. *The Annals of Statistics* **44**, 1298–1331.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer: New York.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge Univ. Press: New York.

Wang, N., Wang, L. and McMahan, C. S. (2015). Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm. *Computational Statistics and Data Analysis* **83**, 140–150.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

Wen, C. -C. and Chen, Y. -H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty Cox model. *Computational Statistics and Data Analysis* **55**, 1053–1060.

Xue, H., Lam, K. F. and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association* **99**, 346–356.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 507–564.

Zeng, D., Mao, L., and Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103**, 253–271.

School of Economics and Statistics, Guangzhou University, Guangzhou, China.
E-mail: lishuwstat@163.com
School of Mathematical Sciences, Capital Normal University, Beijing, China.
E-mail: hutaomath@foxmail.com
Center for Applied Statistical Research, School of Mathematics, Jilin University, Changchun, China.
E-mail: zhaoss@jlu.edu.cn
Department of Statistics, University of Missouri, Columbia, MO 65211, USA.
E-mail: sunj@missouri.edu

(Received April 2017; accepted July 2018)