Biometrics 62, 1037–1043
December 2006
DOI: 10.1111/j.1541-0420.2006.00570.x

Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited

Fushing Hsieh,[1] Yi-Kuan Tseng,[2] and Jane-Ling Wang[1,*]

[1]Department of Statistics, University of California, Davis, California 95616, U.S.A.
[2]Graduate Institute of Statistics, National Central University, Jhongli City, Taoyuan County 32001, Taiwan
[*]*email*: wang@wald.ucdavis.edu

Summary. The maximum likelihood approach to jointly model the survival time and its longitudinal covariates has been successful to model both processes in longitudinal studies. Random effects in the longitudinal process are often used to model the survival times through a proportional hazards model, and this invokes an EM algorithm to search for the maximum likelihood estimates (MLEs). Several intriguing issues are examined here, including the robustness of the MLEs against departure from the normal random effects assumption, and difficulties with the profile likelihood approach to provide reliable estimates for the standard error of the MLEs. We provide insights into the robustness property and suggest to overcome the difficulty of reliable estimates for the standard errors by using bootstrap procedures. Numerical studies and data analysis illustrate our points.

Key words: Joint modeling; Missing information principle; Nonparametric maximum likelihood; Posterior density; Profile likelihood.

1. Introduction

It has become increasingly common in survival studies to record the values of key longitudinal covariates until the occurrence of survival time (or event time) of a subject. This leads to informative missing/dropout of the longitudinal data, which also complicates the survival analysis. Furthermore, the longitudinal covariates may involve measurement error. All these difficulties can be circumvented by including random effects in the longitudinal covariates, for example, through a linear mixed effects model, and modeling the longitudinal and survival components jointly rather than separately. Such an approach is termed "joint modeling" and we refer the readers to the insightful surveys of Tsiatis and Davidian (2004) and Yu et al. (2004) and the references therein. Numerical studies suggest that the "joint maximum likelihood" (ML) method of Wulfsohn and Tsiatis (1997), hereafter abbreviated as WT, is among the most satisfactory approaches to combine information. We focus on this approach in this article and address several intriguing issues.

The approach described in WT is semiparametric in that no parametric assumptions are imposed on the baseline hazard function in the Cox model (Cox, 1972), while the random effects in the longitudinal component are assumed to be normally distributed. An attractive feature of this approach, as confirmed in simulations, is its robustness against departure from the normal random effects assumption. It is in fact as efficient as a semiparametric random effects model proposed by Song, Davidian, and Tsiatis (2002). These authors called for further investigation of this intriguing robustness feature. We provide a theoretical explanation in Section 2 and demonstrate it numerically in Section 4. A second finding of this article is to point out the theoretical challenges regarding efficiency and the asymptotic distributions of the parametric estimators in the joint modeling framework. No distributional or asymptotic theory is available to date, and even the standard errors (SE), defined as the standard deviations of the parametric estimators, are difficult to obtain. We explore these issues in Section 3, highlight the risks when adopting a heuristic proposal in the literature to estimate the SE, and provide numerical illustrations in Section 4. A bootstrap procedure is proposed instead to estimate the standard errors. Furthermore, in Section 5, a data example demonstrates the discussed issues as well as the effectiveness of the bootstrap procedure.

2. Robustness of the Joint Likelihood Approach

Without loss of generality, we assume a single time-dependent covariate, $X_i(t)$, for the ith individual with $i = 1, \ldots, \ldots, n$. The survival time L_i is subject to usual independent random censoring by C_i, and we observe (V_i, Δ_i) for the ith individual, where $V_i = \min(L_i, C_i)$ and $\Delta_i = 1(L_i \leq C_i)$. The longitudinal processes are scheduled to be measured at discrete time points t_{ij} for the ith individual, and are terminated at the endpoint V_i. Hence, the schedule that is actually observed is $\mathbf{t}_i, = \{t_{i1}, \ldots, t_{im_i}\}$ with $t_{im_i} \leq V_i < t_{im_i+1}$, and no longitudinal measurements are available after time V_i, prompting the need to incorporate the survival information and thus the joint modeling approach to combine information. We further assume a measurement error model for the longitudinal covariate, so in reality what is actually observed is

$$W_{ij} = X_i(t_{ij}) + e_{ij} = X_{ij} + e_{ij}, \quad j = 1, \ldots, m_i. \quad (1)$$

© 2006, The International Biometric Society

Here e_{ij} is measurement error that is independent of X_{ij} and has a parametric distribution, such as $N(0, \sigma_e^2)$. Hereafter, we use boldface vector symbols $\mathbf{W}_i = (W_{i1}, \ldots, W_{im_i})$ to denote the observed longitudinal data. To recover the unobserved $X_i(t)$, we assume a parametric random effects model with a k-dimensional random effects vector $\mathbf{b}_i \sim g_\alpha$, denoted by $X_i(t; \mathbf{b}_i) = X_i(t)$. The covariate history $\bar{X}_i(t; \mathbf{b}_i) = \{X_i(s; \mathbf{b}_i) \mid 0 \leq s \leq t\}$ is then related to the survival time through a time-dependent Cox proportional hazards model. The hazard function for the ith individual is thus

$$\lambda_i(t \mid \bar{X}_i(t; \mathbf{b}_i)) = \lambda_0(t) e^{\beta X_i(t; \mathbf{b}_i)}. \tag{2}$$

The joint modeling of the longitudinal and survival parts is specified by (1) and (2). The parameter that specifies the joint model is $\theta = (\beta,\ \lambda_0,\ \alpha,\ \sigma_e^2)$, where the baseline λ_0 is nonparametric.

Let $f(\mathbf{W}_i;\ \alpha,\ \sigma_e)$ and $f(\mathbf{W}_i \mid \mathbf{b}_i;\ \sigma_e^2)$ be, respectively, the marginal and conditional density of $\mathbf{W}_i$, and $f(V_i,\ \Delta_i \mid \mathbf{b}_i,\ \beta,\ \lambda_0)$ the conditional p.d.f. (probability density function) corresponding to (2). Under the assumptions of an "uninformative" time schedule as illuminated in Tsiatis and Davidian (2004), the contribution of the ith individual to the joint likelihood is

$$L_i(\theta) = f(\mathbf{W}_i; \alpha, \sigma_e) E_i^*\{f(V_i, \Delta_i \mid \mathbf{b}_i; \theta)\}, \tag{3}$$

where E_i^* denotes conditional expectation with respect to the posterior density of b_i given $\mathbf{W}_i$, which is $g(\mathbf{b}_i \mid \mathbf{W}_i;\ \alpha,\ \sigma_e^2) = f(\mathbf{W}_i \mid \mathbf{b}_i;\ \sigma_e^2) g_\alpha(\mathbf{b}_i) / f(\mathbf{W}_i;\ \alpha,\ \sigma_e^2)$.

Two facts emerge. First, $E_i^*\{f(V_i,\ \Delta_i \mid \mathbf{b}_i;\ \theta)\}$ carries information on the longitudinal data. If it is ignored, the marginal statistical inference based on $f(\mathbf{W}_i;\ \alpha,\ \sigma_e^2)$ alone would be inefficient and even biased (due to informative dropout). This sheds light on how a joint modeling approach eliminates the bias incurred by a marginal approach and also why it is more efficient. Second, the random effect structure (g_α) and $\mathbf{W}_i$ are relevant to the information for survival parameters (β and λ_0) only through the posterior density $g(\mathbf{b}_i \mid \mathbf{W}_i,\ \alpha,\ \sigma_e^2)$, which can be approximated well by a normal density through a Laplace approximation technique (Tierney and Kadane, 1986) when reasonably large numbers of longitudinal measurements are available per subject. This explains why WT's procedure is robust against departure from the normal prior assumption. Similar robust features were observed for other likelihood-based procedures (Solomon and Cox, 1992; Breslow and Clayton, 1993; Breslow and Lin, 1995). See Section 4 for further discussion.

3. Relation of EM Algorithm and Fisher Information

Direct maximization of the joint likelihood in (3) is impossible due to the nonparametric component λ_0. However, the nonparametric maximum likelihood estimate (MLE) of $\lambda_0(t)$, as defined in Kiefer and Wolfowitz (1956), can be used and it has discrete mass at each uncensored event time V_i. The semiparametric problem in (3) is thus converted to a parametric problem with the parameter representing λ_0 being a discrete probability measure of dimension in the order of n. The expectation maximization (EM) algorithm was employed successfully in WT to treat the unobserved random effects $\mathbf{b}_i$ as missing data, leading to a satisfactory nonparametric MLE approach; compare formulae (3.1)–(3.4), and the subsequent discussion on page 333 of WT. While there are some computational costs incurred by the imputation of the missing random effects in the EM algorithm, the real challenge lies in theory. A major challenge is the high-dimensional nature of the baseline hazards parameter so that standard asymptotic arguments for MLEs do not apply. A profile likelihood approach would be an alternative but it encounters difficulties as well, as elaborated below.

3.1 *Implicit Profile Estimates*

We begin with the nonparametric MLE denoted as $\hat{\lambda}_0(\alpha, \sigma_e, \beta)$, given the parameter $(\alpha,\ \sigma_e,\ \beta)$. Substituting this nonparametric MLE into (3) to produce a profile likelihood $L(\alpha, \sigma_e, \beta, \hat{\lambda}_0(\alpha, \sigma_e, \beta))$ and then maximizing this likelihood would yield profile estimates of $(\alpha,\ \sigma_e,\ \beta)$. However, for the profile approach to work as in the classic setting, the profile likelihood $L(\alpha, \sigma_e, \beta, \hat{\lambda}_0(\alpha, \sigma_e, \beta))$ with the nonparametric MLE $\hat{\lambda}_0$ in place should not involve λ_0. Unfortunately, this is not the case here because the nonparametric MLE $\hat{\lambda}_0(\alpha, \sigma_e, \beta)$ cannot be solved explicitly under the random effects structure. Instead, the EM algorithm is employed to update the profile likelihood estimate for $\lambda_0(t)$. The resulting estimate is

$$\hat{\lambda}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i 1_{(V_i = t)}}{\displaystyle\sum_{j=1}^{n} E_j[\exp\{\beta X_j(t; \mathbf{b}_j)\}] 1_{(V_j \geq t)}}. \tag{4}$$

Strictly speaking, this is not an estimate as it involves functions in the E-step, E_j, which are taken with respect to the posterior density involving $\lambda_0(t)$ itself. Specifically, the posterior density is

$$\begin{aligned} &h(\mathbf{b}_i \mid V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i; \theta) \\ &\quad = \frac{g\big(\mathbf{b}_i \mid \mathbf{W}_i, \mathbf{t}_i; \alpha, \sigma_e^2\big) f(V_i, \Delta_i \mid \mathbf{b}_i, \beta, \lambda_0)}{\displaystyle\int g\big(\mathbf{b}_i \mid \mathbf{W}_i, \mathbf{t}_i; \alpha, \sigma_e^2\big) f(V_i, \Delta_i \mid \mathbf{b}_i, \beta, \lambda_0)\, d\mathbf{b}_i}. \end{aligned} \tag{5}$$

We can now see that (4) yields an implicit profile estimate since E_j involves $\lambda_0(t)$. The implication is that although a point estimator of λ_0 can be derived via the EM algorithm (4), the usual profile approach to calculate the Fisher information cannot be employed. The asymptotic covariance matrix for λ_0 cannot be evaluated through derivatives of the implicit profile likelihood as in traditional situations, where an explicit form of profile likelihood exists. We will resume this information issue in Section 3.2 after we discuss the remaining profile estimates.

Likewise, the maximum "profile" likelihood estimates of α and σ_e^2 are both implicit because the posterior expectations E_j involve both parameters. The estimation of β is more complicated. To see this, note that only the second term, E_i^*, in (3) involves β and E_i^* involves the posterior density of $\mathbf{b}_i$, given $\mathbf{W}_i$. This posterior density is the g-function in (5) and not the posterior density h. Thus, E_i^* and E_i are different, but after some rearrangement the score equation for β can be expressed as $S_\beta = \sum_{i=1}^n E_i\{S^c(\beta; \lambda_0, \mathbf{b}_i)\}$, where $S^c(\beta; \lambda_0, \mathbf{b}_i) = \sum_{i=1}^n \frac{\partial}{\partial\beta} \log f(V_i, \Delta_i \mid \mathbf{b}_i, \beta, \lambda_0)$.

Substituting $\lambda_0(t)$ by $\hat{\lambda}_0(t)$ in equation (4), the maximum profile score of β is again in implicit form and is denoted as

$$S_\beta^{IP} = \sum_{i=1}^n E_i\{S^c(\beta; \hat{\lambda}_0(t), \mathbf{b}_i)\}$$
$$= \sum_{i=1}^n \Delta_i [E_i\{X_i(V_i; \mathbf{b}_i)\}]$$
$$- \frac{\sum_{j=1}^n E_j[X_j(V_i; \mathbf{b}_i)\exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}}{\sum_{j=1}^n E_j[\exp\{\beta X_j(V_i)\}]1_{(V_j\geq V_i)}}. \quad (6)$$

The EM algorithm then iterates between the E-step, to evaluate the conditional expectations E_i with parameter values obtained from the previous iteration $\hat{\theta}^{(k-1)}$, and the M-step, to update the estimated values via the above score equation. Since (6) has no closed-form solution, the Newton–Raphson method is applied: $\hat{\beta}_k = \hat{\beta}_{k-1} + I^{-1}_{\hat{\beta}_{k-1}} S^{IP}_{\hat{\beta}_{k-1}}$, where $S^{IP}_{\hat{\beta}_{k-1}}$ is the value of the incomplete profile score in (6) with $\beta = \hat{\beta}_{k-1}$, and the slope $I^{-1}_{\hat{\beta}_k}$ is obtained through the following working formula:

$$I_\beta^W$$
$$= \sum_{i=1}^n E_i\left\{-\frac{\partial}{\partial\beta}S^c(\beta; \lambda_0, \mathbf{b}_i)\right\}$$
$$= \sum_{i=1}^n \left\{ \frac{\sum_{j=1}^n E_j\left[X_j(V_i;\mathbf{b}_i)^2\exp\{\beta X_j(V_i;\mathbf{b}_i)\}\right]1_{(V_j\geq V_i)}}{\sum_{j=1}^n E_j[\exp\{\beta X_j(V_i;\mathbf{b}_i)\}]1_{(V_j\geq V_i)}} - \left(\frac{\sum_{j=1}^n E_j[X_j(V_i;\mathbf{b}_i)\exp\{\beta X_j(V_i;\mathbf{b}_i)\}]1_{(V_j\geq V_i)}}{\sum_{j=1}^n E_j[\exp\{\beta X_j(V_i;\mathbf{b}_i)\}]1_{(V_j\geq V_i)}}\right)^2 \right\}. \quad (7)$$

The above iterative procedure is implemented until a convergence criterion is met. We next examine what has been achieved by the EM algorithm at this point.

3.2 *Fisher Information*

The iterative layout of the EM algorithm is designed to achieve the MLEs and hence the consistency of parameters $(\alpha, \beta, \sigma_e^2)$ and the cumulative baseline hazard function. There is no complication on this front. However, the working formula I_β^W in (7) at the last step of the EM algorithm has been suggested in the literature to provide the precision estimate for the standard deviation (standard error) of the β-estimator via $(I_\beta^w)^{-1/2}$. This is invalid. It was derived by taking the partial derivative of the implicit profile score in equation (6) with respect to β, by treating the conditional expectation $E_i[\cdot]$ as if it does not involve β. There are two gaps. First, E_i does involve β through the posterior density, so the proper way to take the derivative of (6) should involve the multiplication rule. Second, since (6) is an implicit score, its partial derivative with respect to β does not yield the correct Fisher information. Instead, the projection method needs to be applied to the Hessian matrix, $-\frac{\partial^2}{\partial\theta^2}\log L(\theta)$, based on the likelihood $L(\theta) = \prod_1^n L_i(\theta)$ to properly pin down the correct Fisher information.

For illustration purposes we consider a simpler case and assume for the moment that α and σ_e^2 are known so that $\theta = (\beta,\ \lambda_0)$. For this θ denote $I_{\beta\beta} = -\frac{\partial^2}{\partial\beta^2}\log L(\theta), I_{\beta\lambda_0} = -\frac{\partial^2}{\partial\beta\partial\lambda_0}\log L(\theta)$, and define $I_{\lambda_0\lambda_0}$ and $I_{\lambda_0\beta}$ similarly. The correct projection to reach the Fisher information for β is $I_{\beta\beta} - I_{\beta\lambda_0}[I_{\lambda_0\lambda_0}]^{-1}I_{\lambda_0\beta}$, which involves inverting a high-dimensional matrix $I_{\lambda_0\lambda_0}$. In the general situation with four unknown parametric components the correct Fisher information would be even more difficult to compute, as this Hessian matrix has 4×4 block matrices corresponding to the four parametric components in θ, and the projection would be complicated. We recommend bootstrapping to estimate the standard deviation for all finite-dimensional parameters $(\alpha,\ \sigma_e,\ \beta)$ and illustrate its effectiveness in Section 4.

We close this section by further showing that the working SE, $(I_\beta^W)^{-1/2}$, is smaller than the real SE based on the sample Fisher information of β. Hence statistical inference based on $(I_\beta^w)^{-1/2}$ would be too optimistic. Note that the contribution of the ith individual to the true Hessian is

$$I_{i(\beta,\beta)} = -\frac{\partial}{\partial\beta}E_i[S^c(\beta;\lambda_0,\mathbf{b}_i)]$$
$$= E_i\left[-\frac{\partial}{\partial\beta}S^c(\beta;\lambda_0,\mathbf{b}_i)\right]$$
$$- E_i\left[S^c(\beta;\lambda_0,\mathbf{b}_i)S^h(\beta;\lambda_0,\mathbf{b}_i)\right], \quad (8)$$

with

$$S^h(\beta;\lambda_0,\mathbf{b}_i) = \frac{\partial}{\partial\beta}\log h(\mathbf{b}_i \mid V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i; \theta)$$
$$= S^c(\beta;\lambda_0,\mathbf{b}_i) - E_i\left\{S^c(\beta;\lambda_0,\mathbf{b}_i)\right\} \quad (9)$$

being the score function pertaining to the posterior density $h(\mathbf{b}_i \mid V_i,\ \Delta_i,\ \mathbf{W}_i,\ \mathbf{t}_i;\ \theta)$ in (5). It is interesting to note that the second term in (8) is equal to $\mathrm{Var}_i[S^c(\beta;\ \lambda_0,\ \mathbf{b}_i)]$, the conditional variance with respect to the posterior density, which is also equal to the amount of Fisher information of β contained in this posterior density. Recall from the first equality in (7) that I_β^W is the sample Fisher information of β, pertaining to the likelihood of $f(X_i, \Delta_i \mid b_i, \beta)$ in the parametric Cox model. It now follows from (7)–(9) that the true Hessian is

$$I_{\beta\beta} = I_\beta^W - \sum_{i=1}^n \mathrm{Var}_i\left\{S^c(\beta;\lambda_0,\mathbf{b}_i)\right\}. \quad (10)$$

The implication of this relation is that under the joint modeling with random effects structure on the covariate $\mathbf{X}_i$, some information is lost and that this loss is unrecoverable. Therefore, I_β^W is not the true Hessian, and would be too large when compared with the true Hessian, $I_{\beta\beta}$, in (10). The actual amount of information loss, $\sum_{i=1}^n \mathrm{Var}_i\{S^c(\beta;\lambda_0,\mathbf{b}_i)\}$ in (10),

Table 1
Simulated results for case (1) when the longitudinal data have normal random effects

	$(\frac{1}{4}\sigma_{22}, \sigma_e^2)$ (0.0025, 0.1)	$(4\sigma_{22}, \sigma_e^2)$ (0.04, 0.1)	$(\sigma_{22}, \sigma_e^2)$ (0.01, 0.1)	$(\sigma_{22}, 4\sigma_e^2)$ (0.01, 0.4)	$(\sigma_{22}, \frac{1}{4}\sigma_e^2)$ (0.01, 0.025)	$(\sigma_{22}, 0\sigma_e^2)$ (0.01, 0)
SD for $\hat{\beta}$	0.132	0.098	0.112	0.169	0.108	0.103
Mean of $(I_{\hat{\beta}}^W)^{-1/2}$	0.118	0.085	0.104	0.103	0.103	0.102
Mean of $\hat{\beta}$	1.008	0.998	1.004	1.007	0.984	0.987
Divergence	0%	0%	0%	4%	0%	0%

involves the posterior variances. Intuitively, we would expect these posterior variances to increase as the measurement error increases, and this is confirmed in the numerical study reported in Tables 1 and 2 in Section 4.

This missing information phenomenon, first introduced by Orchard and Woodbury (1972), applies to all the finite-dimensional parameters such as α and σ_e^2, as well as $\lambda_0(t)$ in the setting considered in this article. It is not unique to the joint modeling setting and would persist even for a fully parametric model whenever the EM algorithm is employed. Louis (1982) provided a specific way to calculate the observed Fisher information, but this would be a formidable task under our setting due to the large number of parameters involved.

4. Simulation Study

In this section, we examine the gap between the working SE formula $(I_{\hat{\beta}}^W)^{-1/2}$ for $\hat{\beta}$ and a reliable precision estimate from the Monte Carlo sample standard deviation. A second goal is to explore the robustness issue in Section 2. Three cases were considered in the simulations. In all cases, a constant baseline hazard function is used for λ_0 with measurement error $e_{ij} \sim N(0, \alpha\sigma_e^2)$, where $\alpha = 0, \frac{1}{4}, 1$, or 4. The longitudinal covariate $X_i(t) = b_{1i} + b_{2i}t$ is linear in t. The random effects $\mathbf{b}_i = (b_{1i}, b_{2i})$ were generated from either a bivariate normal distribution (with $E(b_{1i}) = \mu_1$, $E(b_{2i}) = \mu_2$, $\text{var}(b_{1i}) = \sigma_{11}$, $\text{var}(b_{2i}) = \gamma\sigma_{22}$ with $\gamma = \frac{1}{4}, 1$, or 4, and $\text{cov}(b_{1i}, b_{2i}) = \sigma_{12}$), or a truncated version where βb_{2i} is restricted to be nonnegative. The various values of α and γ allow us to examine the effects of the variance components on the accuracy of the working formula. The case of no measurement error ($\alpha = 0$) illustrates that the working formula is now correct as the random effects are determined from the observations $\mathbf{W}_i$ so that there is no information loss.

In each setting, the sample size is $n = 200$ and the number of replications is 100. The censoring time for each subject is generated from exponential distribution with mean 25 for the first setting and 110 for the other two settings, resulting in about 25% censoring. The EM algorithm procedure used here is the same as in WT except for the numerical integration to evaluate the conditional expectation in the E-step. Instead of using the Gauss–Hermite quadrature formula we applied Monte Carlo integration, as suggested by Henderson, Diggle, and Doboson (2000) to calculate conditional expectations. Parameters and time schedules for the longitudinal and survival parts under the three settings are specified below.

(a) $t_{ij} = (0, 1, \ldots, 12)$, $\beta = 1$, $\lambda_0 = 0.001$, $\sigma_e^2 = 0.1$, $(\mu_1, \mu_2) = (2, 0.5)$, and $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.5, -0.001, 0.01)$.
(b) $t_{ij} = (0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80)$, $\beta = -1$, $\sigma_e^2 = 0.6$, $\lambda_0 = 1$, $(\mu_1, \mu_2) = (4.173, -0.0103)$, and $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (4.96, -0.0456, 0.012)$.
(c) Same as (b) except that we reduce σ_e^2 to 0.15, and for each subject we only select the measurement at an initial time point and randomly one or two more points between the initial and event time.

The first setting yields normally distributed random effects while the second one yields very skewed truncated-normal random effects (46% of the positive b_{2i} were discarded), but with a sufficient amount of longitudinal measurements to invoke the robustness feature as discussed in Section 2. The third case also yields truncated-normal random effects, but with at most three longitudinal measurements per subject. This sparse design for the longitudinal measurements confirms our remark in Section 2 that it has an adverse effect to the robustness of a likelihood-based procedure. The choice of a much smaller σ_e^2 for the sparse design is due to the high rates of nonconvergence observed for the EM algorithm (over 40% when the same σ_e^2 as in case [b] was used). A smaller nonconvergence rate as reported in Table 3 provides a stable background for illustrating the breakdown of the procedure against model violations.

The simulation results are summarized in Tables 1–3 corresponding to the three settings described above. We focus on the parameter β to save space. Nonconvergence rates of the

Table 2
Simulated results for case (2) with nonsparse longitudinal data and nonnormal random effects

	$(\frac{1}{4}\sigma_{22}, \sigma_e^2)$ (0.003, 0.6)	$(4\sigma_{22}, \sigma_e^2)$ (0.048, 0.6)	$(\sigma_{22}, \sigma_e^2)$ (0.012, 0.6)	$(\sigma_{22}, 4\sigma_e^2)$ (0.012, 2.4)	$(\sigma_{22}, \frac{1}{4}\sigma_e^2)$ (0.012, 0.15)	$(\sigma_{22}, 0\sigma_e^2)$ (0.012, 0)
SD for $\hat{\beta}$	0.130	0.111	0.119	0.196	0.114	0.103
Mean of $(I_{\hat{\beta}}^W)^{-1/2}$	0.113	0.090	0.100	0.098	0.103	0.103
Mean of $\hat{\beta}$	−0.977	−0.981	−0.995	−0.987	−0.992	−1.002
Divergence	0%	1%	0%	3%	0%	0%

Table 3
Simulated results for case (3) with sparse longitudinal data and nonnormal random effects

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Simulated value	−1	4.173	−0.0103	4.96	−0.0456	0.012	0.15
Empirical target	−1	4.5032	−0.0913	4.7950	−0.0175	0.0046	0.1533
Mean	−1.3498	4.4172	−0.2366	3.4356	0.0130	0.0067	3.2903
SD	0.1647	0.1573	0.0169	0.4413	0.0246	0.0015	0.3574

EM algorithm are in the fourth row of Tables 1 and 2, and mean of the 100 β-estimates in the third row. This suggests that the EM algorithm works well in these settings with low nonconvergence rates, and the likelihood-based joint modeling approach indeed provides unbiased estimates under the true random effects distribution setting in Table 1. When the normal random effects assumption is violated (as in Table 2), it provides roughly unbiased estimates, a confirmation of the robustness of the procedure. The same robustness phenomenon was observed in Table 1 of Tsiatis and Davidian (2004) where the random effects were a mixture of two bivariate normal distributions leading to symmetric but bimodal distributions. Our setting in (b) has a skewed (truncated-normal) random effects distribution and thus complements their findings. Other simulations not reported here with heavier tail distributions, such as multivariate t-distributions, enjoy the same robustness property as long as the longitudinal measurements are dense enough. Note that there are at most 13, and often much fewer, longitudinal measurements for each individual in Table 2, so this is a rather encouraging finding in favor of the normal likelihood-based approach.

The SD reported in the first row of Tables 1 and 2 corresponds to the sample standard deviation of the β-estimates from 100 Monte Carlo samples, and the working SE reported in the second row is the mean of the 100 estimated $(I_{\hat{\beta}}^{W})^{-1/2}$ in equation (7), with all unknown quantities estimated at the last step of the EM algorithm. From these two tables, we can see that the working SE formula underestimates the standard deviation of $\hat{\beta}$ in all cases except in the last column, corresponding to no measurement error. The size of observed differences is directly related to the size of the measurement errors (see columns 3–6). The discrepancy ranges from almost none (when measurement error is small or 0) to as large as 40% in Table 1 and 50% in Table 2. This confirms empirically that the working slope should not be used for statistical inference of β, especially in the case of large σ_e^2. We re-emphasize that this applies equally to all other parameters among $(\alpha, \sigma_e, \lambda_0)$.

An interesting finding from the first three columns of Tables 1 and 2 is that the discrepancies are noticeable but they do not vary much when one changes only the variance component of b_{2i}. Larger σ_{22}, however, gives more precise estimates for β as the associated SD is smallest for the second column in both tables. This is perhaps not surprising and is consistent with usual experimental design principles in cross-sectional studies. Typically, larger variations in covariates lead to more precise regression coefficients, which would correspond to β in the joint modeling setting. So our simulations suggest that the same design principle continues to hold in this setting. Increased variation among individual longitudinal processes provides more information on how the longitudinal covariate of a subject relates to its survival time.

The robustness enjoyed in situations with rich longitudinal data information is in doubt for the sparse data situation in case (c). Table 3 shows 35% bias in estimating β for this case. The mean of the working SE is 0.1163 while the simulated SD is 0.1647, and a large discrepancy is exhibited here. In addition to β, we were also curious about the robustness of other parameters. Since the original simulated values (reported in row 1) are no longer the target because of truncation, the target values are estimated empirically in Table 3 and reported in row 2. This then should be the comparison base for the bias of other parameters, which reveals significant, and sometimes serious, biases as well.

4.1 *Bootstrap Estimate for the Standard Error*

As there is no reliable SE formula available due to the theoretical difficulties mentioned in Section 3, we propose a bootstrap method to estimate the SE. The bootstrap sample $\{Y_i^*,\ i = 1, \ldots, n\}$, where $Y_i = (V_i,\ \Delta_i,\ \mathbf{W}_i,\ \mathbf{t}_i)$, is generated from the original observed data as defined in Section 2. The EM algorithm is then applied to the bootstrap sample, Y_i^*, $i = 1, \ldots, n$, to derive the MLE $\hat{\theta}^*$, and this is repeated B times. The bootstrapped SD estimates for θ come from the diagonal elements of the covariance matrix, $\mathrm{Cov}(\hat{\theta}^*) = 1/(B-1)\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}_b)(\hat{\theta}_b^* - \bar{\theta}_b)^T$, where $\bar{\theta}_b = \sum_{b=1}^{B}\hat{\theta}_b^*/B$.

To check the stability of the above bootstrap procedure we conducted a small simulation study, using the design in the first setting as reported in the third column of Table 1. Of the 100 Monte Carlo samples, we resampled $B = 50$ bootstrap samples for each single Monte Carlo sample. The SD for $\hat{\beta}$ based on the 100 Monte Carlo samples is 0.1193 and the mean of the 50 bootstrap standard deviation estimates for β is 0.1210 with a standard deviation of 0.0145. This suggests that the bootstrap SD is quite stable and reliable as it is close to the Monte Carlo SD. Moreover, the EM algorithm has a low nonconvergence rate of 0.44%, which is 22 out of the 5000 Monte Carlo bootstrap samples.

5. Egg-Laying Data Analysis

The lifetime (in days) and complete reproductive history, in terms of number of eggs laid daily until death, of 1000 female medflies (Mediterranean fruit fly) were recorded by Carey et al. (1998) to explore the relation between reproduction and longevity. Since we have the complete egg-laying history without missing or censored data, we could use standard software to check the proportional hazards assumption without relying on the last-value-carry-forward procedure, known to be prone

Table 4

Analysis of lifetime and log-fecundity of medflies with lifetime reproduction less than 300 eggs. The bootstrap SD and the working SE, $(I_{\hat{\theta}}^{W})^{-1/2}$, are reported in the last two rows.

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
EM estimate	0.5597	0.6950	−0.0542	0.7656	−0.1123	0.0196	0.9883
Bootstrap mean	0.5676	0.6932	−0.0541	0.7621	−0.1121	0.0198	0.9858
Bootstrap SD	0.0921	0.0736	0.0156	0.1056	0.0207	0.0043	0.0606
$(I_{\hat{\theta}}^{W})^{-1/2}$	0.0801	0.0663	0.0141	0.0988	0.0176	0.0039	0.0495

to biases. The proportional hazards assumption failed for flies that are highly productive, so we restrict our attention to the less productive half of the flies. This includes female medflies that produced less than a total of 749 eggs during their lifetimes. To demonstrate the effects of wrongly employing the working SE formula to estimate the standard errors of the estimating procedures in Section 3, we further divide these flies into two groups. The first group includes 249 female medflies that have produced eggs but with a lifetime reproduction of less than 300 eggs. The second group includes the 247 female medflies producing between 300 and 749 eggs in their lifetimes.

As daily egg production is subject to random fluctuations but the underlying reproductive process can be reasonably assumed to be a smooth process, the measurement error model (1) is a good prescription to link the underlying reproductive process to longevity. Because we are dealing with count data, it is common to take the log transformation, so we take $W(t) = \log\{m(t) + 1\}$, to avoid days when no eggs were laid. An examination of the individual egg-laying trajectories in Carey et al. (1998) suggests the following parametric longitudinal process for $X(t)$:

$$X(t) = b_1 \log(t) + b_2(t - 1),$$

where the prior distribution of $(b_1,\ b_2)$ is assumed to be bivariate normally distributed with mean $(\mu_1,\ \mu_2)$, and $\sigma_{11} = \mathrm{var}(b_1)$, $\sigma_{12} = \mathrm{cov}(b_1,\ b_2)$, $\sigma_{22} = \mathrm{var}(b_2)$. Note here that the longitudinal data are very dense as daily observations were available for all flies, so the model fitting should not be sensitive to this normality assumption as supported by the robustness property discussed in previous sections.

The Cox proportional hazard regression model assumptions were checked via martingale residuals for both data sets and were reasonably satisfied with p-values 0.9 and 0.6, respectively. We thus proceeded with fitting the joint models in (1) and (2). Tables 4 and 5 summarize the EM algorithm estimates of θ together with their bootstrap means and standard deviations based on 100 bootstrap replications. The corresponding working SE, $(I_{\hat{\theta}}^{W})^{-1/2}$, is also reported on the last row of each table for a contrast.

From Tables 4 and 5, we can see that the sample bootstrap means are close to the corresponding EM estimates, suggesting the feasibility of the bootstrap approach. On the other hand, the working SE corresponding to the last step of the EM algorithm produced estimates that are noticeably smaller than the bootstrap SD estimate. For β, it yielded an estimate of 0.0801 under the setting of Table 4, which is about a 15% departure from the bootstrap SD (0.0921). The discrepancy increases to 25% (working SE is 0.0672) for the second data set in Table 5 due to a higher measurement error (1.4344 in Table 5 vs. 0.9883 in Table 4), consistent with the simulation findings in Section 4. Similar discrepancies can also be seen between Tables 4 and 5 for all other parameters.

6. Conclusions

We have accomplished several goals in this article: (i) Reinforcing the merit of the joint modeling approach in WT by providing a theoretical explanation of the robustness features observed in the literature. This suggests that the likelihood-based procedure with normal random effects can be very efficient and robust as long as there is rich enough information available from the longitudinal data. Generally, this means that the longitudinal data should not be too sparse or carry too large measurement errors. (ii) Demonstrating the missing information and implicit profile features in joint modeling both theoretically and empirically to alert practitioners. The efficiency loss due to missing random effects can be quite substantial, as observed in numerical studies. (iii) Recommending until further theoretical advances afford us reliable standard error estimates to use the bootstrap SD estimate for estimating the standard errors of the EM estimators.

However, theoretical gaps remain to validate the asymptotic properties of the estimates in WT and the validity of the bootstrap SE estimates. The theory of profile likelihood

Table 5

Analysis of lifetime and log-fecundity of medflies with lifetime reproduction between 300 and 749 eggs. The bootstrap SD and the working SE, $(I_{\hat{\theta}}^{W})^{-1/2}$, are reported in the last two rows.

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
EM estimate	−0.1997	1.7618	−0.1730	1.0540	−0.1734	0.0301	1.4344
Bootstrap mean	−0.2036	1.7659	−0.1739	1.0487	−0.1766	0.0309	1.4333
Bootstrap SD	0.0836	0.0800	0.0145	0.0784	0.0151	0.0033	0.0356
$(I_{\hat{\theta}}^{W})^{-1/2}$	0.0672	0.0576	0.0120	0.0572	0.0112	0.0023	0.0282

is well developed for parametric settings (Patefield, 1977), but not immediately applicable in semiparametric settings (Murphy and van der Vaart, 2000; Fan and Wong, 2000). The complexity caused by profiling nonparametric parameters with only implicit structure creates additional difficulties in theoretical and computational developments for joint modeling. Further efforts will be required in future work to resolve these issues and to provide reliable precision estimates for statistical inference under the joint modeling setting.

Acknowledgements

The research of Jane-Ling Wang was supported in part by the National Science Foundation and National Institutes of Health. The authors greatly appreciate the suggestions from a referee and an associate editor.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component dispersion. *Biometrika* **82,** 81–91.

Carey, J. R., Liedo, P., Müller, H. G., Wang, J. L., and Chiou, J. M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *Journal of Gerontology—Biological Sciences* **53,** 245–251.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Fan, J. and Wong, W. H. (2000). Comment on "On profile likelihood" by Murphy and van der Vaart. *Journal of the American Statistical Association* **95,** 468–471.

Henderson, R., Diggle, P., and Doboson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **4,** 465–480.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27,** 887–906.

Louis, T. A. (1982). Finding the observed Fisher information when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Murphy, S. and van der Vaart, A. W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association* **95,** 449–465.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, p. 697–715. Berkeley: University of California Press.

Patefield, W. M. (1977). On the maximized likelihood function. *Sankhya, Series B* **39,** 92–96.

Solomon, P. J. and Cox, D. R. (1992). Nonlinear component of variance models. *Biometrika* **79,** 1–11.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Tierney, L. and Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *Journal of the American Statistical Association* **81,** 82–86.

Tsiatis, A. A. and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14,** 809–834.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53,** 330–339.

Yu, M., Law, N. J., Taylor, J. M. G., and Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14,** 835–862.

Received July 2004. *Revised December* 2005.
Accepted January 2006.