Biometrika (2005), **92**, 3, *pp.* 587–603

Printed in Great Britain

Joint modelling of accelerated failure time and longitudinal data

BY YI-KUAN TSENG, FUSHING HSIEH AND JANE-LING WANG
Department of Statistics, University of California, Davis, California 95616, U.S.A.
yktseng@wald.ucdavis.edu fushing@wald.ucdavis.edu wang@wald.ucdavis.edu

SUMMARY

The accelerated failure time model is an attractive alternative to the Cox model when the proportionality assumption fails to capture the relationship between the survival time and longitudinal covariates. Several complications arise when the covariates are measured intermittently at different time points for different subjects, possibly with measurement errors, or measurements are not available after the failure time. Joint modelling of the failure time and longitudinal data offers a solution to such complications. We explore the joint modelling approach under the accelerated failure time assumption when covariates are assumed to follow a linear mixed effects model with measurement errors. The procedure is based on maximising the joint likelihood function with random effects treated as missing data. A Monte Carlo EM algorithm is used to estimate all the unknown parameters, including the unknown baseline hazard function. The performance of the proposed procedure is checked in simulation studies. A case study of reproductive egg-laying data for female Mediterranean fruit flies and their relationship to longevity demonstrate the effectiveness of the new procedure.

Some key words: EM algorithm; Measurement error; Missing data; Monte Carlo integration; Random effect; Survival data.

1. INTRODUCTION

In clinical trials or medical follow-up studies, it has become increasingly common to observe the event time of interest, called survival time or failure time, along with longitudinal covariates. An increasing popular approach is to model both processes simultaneously to explore their relationship and to borrow strength from each component in the model-building process. The longitudinal covariates are usually modelled parametrically with random effects, for example by a linear mixed effects model. Moreover, the longitudinal covariates may not be directly observed, because of an intermittent sampling schedule and/or measurement errors. Let $X(t)$ denote such a longitudinal covariate with additive measurement error, $e(t)$. Then what is actually observed is another process,

$$W(t) = X(t) + e(t), \tag{1}$$

at discrete time points. For simplicity we assume that there is only one longitudinal covariate; generalisation to the case of multiple longitudinal covariates and additional time-independent covariates is straightforward.

For the survival component, the Cox proportional hazards model has been used in the literature to describe the survival information through the hazard rate function

$$\lambda\{t|\bar{X}(t)\} = \lambda_0(t)\exp\{\beta X(t)\}, \tag{2}$$

where $\bar{X}(t) = \{X(s): 0 \leqslant s < t\}$ is the covariate history up to time t, β is the regression parameter, and $\lambda_0(t)$ is the unspecified baseline hazard rate function. The survival time is often subject to random censoring, and a well-known example is that of HIV clinical trials where time-dependent CD4 counts, or viral loads, and an event time, time to AIDS or death, are recorded. Finding associations between the time-varying CD4 count and the event time is an important goal of these experiments; see for instance Pawitan & Self (1993), Tsiatis et al. (1995), Wulfsohn & Tsiatis (1997) and Wang & Taylor (2001).

If there were no measurement error in (1) and the entire history of $X(t)$ were available, one could use Cox's partial likelihood to estimate the regression parameter β in (2). However, either or both assumptions may fail. Intuitively, one could overcome both difficulties by imputing the unobserved covariate process, $X(t)$, in the partial likelihood. Such an approach is called a two-stage procedure in the joint-modelling literature, and has been studied by Tsiatis et al. (1995) and Dafini & Tsiatis (1998) among others. This approach encounters bias when the observation of the longitudinal process is interrupted by the event time, that is when death strikes. In such situations, only measurements before death are available, which results in informatively missing longitudinal data. Bias will occur in both the longitudinal and survival components, if unmodified procedures for linear mixed effects models are employed to fit the longitudinal component. Various remedies have been proposed, and the most satisfactory approach is perhaps the joint likelihood approach in Wulfsohn & Tsiatis (1997), who constucted a joint likelihood of (1) and (2) under certain assumptions including that of normal random effects. The EM algorithm has been employed to estimate the missing random effects. The normality assumption for random effects was later relaxed in Tsiatis & Davidian (2001) through a conditional score approach, and was relaxed to a flexible parametric class of smooth density functions in Song et al. (2002). In addition to linear mixed effects, Henderson et al. (2000) added an extra Gaussian process in $X(t)$ to explain additional correlation in time-dependent covariates. Wang & Taylor (2001) consider a similar model to that of Henderson et al. (2000) and applied a Bayesian framework and Markov chain Monte Carlo methods to fit the joint model. For additional information about joint modelling, see the insightful reviews in Tsiatis & Davidian (2004) and Yu et al. (2004).

So far the literature on joint modelling of survival and longitudinal data has concentrated on the use of the Cox proportional hazards model to characterise the relationship between the longitudinal covariates and the survival information. There are, however, many cases, such as the fecundity data in § 5, where the proportionality assumption in (2) fails. For such situations an accelerated failure time model is a viable alternative. The accelerated failure time model was introduced in Cox (1972) to model the effects of covariates directly on the length of survival time as

$$\log T = -\beta' X + e, \tag{3}$$

where T is the survival time, X is a time-independent covariate and e is the random error. Suppose that S_0 is the baseline survival function of T given $X = 0$. Then S_0 is also the survival function of $U = \exp(e)$.

For time-dependent covariates $X(t)$, Cox & Oakes (1984, Ch. 5, pp. 64–5) propose the following extension of the accelerated failure time model:

$$U \sim S_0, \tag{4a}$$

where

$$U = \psi\{\bar{X}(T); \beta\} = \int_0^T \exp\{\beta X(s)\} ds. \tag{4b}$$

With this transformation, the survival function for an individual with covariate history $\bar{X}(t)$, is $S\{t|\bar{X}(t)\} = S_0[\psi\{\bar{X}(t); \beta\}]$. This means that individuals age on an accelerated schedule, $\psi\{\bar{X}(t); \beta\}$, under a baseline survival function $S_0(.)$. Such a model is biologically meaningful and allows the entire covariate history to influence subject-specific risk. For an absolutely continuous S_0, the hazard rate function for an individual with covariate history $\bar{X}(t)$ can thus be expressed as

$$\lambda\{t|\bar{X}(t)\} = \lambda_0\left[\int_0^t \exp\{\beta X(s)\} ds\right] \exp\{\beta X(t)\} = \lambda_0[\psi\{\bar{X}(t); \beta\}]\dot{\psi}\{\bar{X}(t); \beta\}, \tag{5}$$

where $\lambda_0(.)$ is the hazard function for S_0 and $\dot{\psi}$ is the first derivative of ψ. Here, U plays the role of a baseline failure-time variable and we thus refer to $\lambda_0(.)$ as the baseline hazard function, which is usually left unspecified. Thus, (5) corresponds to a semiparametric model, first studied by Robins & Tsiatis (1992) using a certain class of rank estimating equations for β. These rank estimators were shown to be consistent and asymptotically normal by Lin & Ying (1995). Recently, F. Hsieh, in an unpublished manuscript, proposed an over-identified estimating equation approach for achieving semiparametric efficiency and for extending (5) to a heteroscedastic version. All this aforementioned work assumes, however, that the entire covariate process, $X(t)$, can be observed without measurement errors.

In the rest of the paper, we consider the joint accelerated failure time model as specified by (1) and (5), or equivalently (1) and (4), subject to the further complication that the observation of the longitudinal covariate process is truncated by the event time. Our goal is to provide effective estimators for the regression parameter β without assuming a parametric baseline hazard function $\lambda_0(.)$ in the survival components (4), or (5), and to obtain effective estimators for the model components of the longitudinal process. This is accomplished via the likelihood approach, so one could regard our proposal as the counterpart of the approach in Wulfsohn & Tsiatis (1997) for the proportional hazards mode.

As with the traditional time-independent accelerated failure time model, handling the accelerated failure time structure in the joint modelling setting is much harder than for the proportional hazards model. We assume that the baseline hazard function is a step function in § 2 when specifying the joint likelihood of T and $X(t)$. This differs from Wulfsohn & Tsiatis (1997), where the baseline hazard function is assumed to be discrete. The step-function structure is prompted by the continuous nature of the accelerated failure time model in (5), and it allows us to implement the EM algorithm in § 3.

Calculation of standard errors for the estimator for β turns out to be a difficult issue because of the missing information about random effects in the EM step. We propose a bootstrap method for estimating the standard error of $\hat{\beta}$ and illustrate it through a dataset in § 5.

2. THE JOINT MODEL

Consider n subjects and let T_i be the event time of subject i, which is subject to right censoring by C_i. The observed time is denoted by $V_i = \min(T_i, C_i)$, and Δ_i is the event time indicator, which is equal to 1, if $T_i \leqslant C_i$, and is 0 elsewhere. Without loss of generality, assume a single time-dependent covariate $X_i(t)$ for subject i, as the case of multiple covariates can be handled similarly. The covariate processes $X_i(.)$ are scheduled to be measured, with error, at times t_{ij}, but no measurement is available after the event time. Thus, the measurement schedule of subject i is $t_i = (t_{ij}, t_{ij} \leqslant V_i)$ and there are m_i repeated measurements for subject i, so that $j = 1, \ldots, m_i$. The measurements for subject i are $W_i = (W_{ij})$, with measurement errors $e_i = (e_{ij})$, for $j = 1, \ldots, m_i$, where $W_{ij} = X_i(t_{ij}) + e_{ij}$. Therefore, the observed data for each individual are $(V_i, \Delta_i, W_i, t_i)$, with all variables independent across i.

As with the practice for joint modelling, we restrict the longitudinal covariate to be from a Gaussian model specified via linear mixed effects,

$$X_i(t) = b_i' \rho(t), \tag{6}$$

where $\rho(t) = \{\rho_1(t), \ldots, \rho_p(t)\}'$ and the $\rho_j(t)$ are known functions; $b_i' = (b_{1i}, \ldots, b_{pi})$ are p-dimensional multivariate normal, $N_p(\mu, \Sigma)$, independent of the measurement errors e_i. The measurement errors, e_i, are also assumed to be multivariate normal, with independent and identically distributed components $e_{ij} \sim N(0, \sigma_e^2)$. The random effect vectors b_i, which are not observed and are treated as missing data in the likelihood approach to follow, are estimated by the EM algorithm. If $p = 2$ and $\{\rho_1(t), \rho_2(t)\} = (1, t)$, then (6) is the linear growth-curve model considered in the joint model literature. Higher-order polynomials $\{\rho_1(t), \ldots, \rho_p(t)\} = (1, \ldots, t^{p-1})$ can be used to include more complicated growth-curve models at high computational cost, as the EM steps involve evaluation of p-dimensional integrals. As a result, only a few random effects are employed in practice and different basis functions $\rho_k(t)$ may be called for if the trajectory of $X_i(t)$ is nonlinear over time. This occurs for the egg-laying trajectories of the medfly data in § 5, where we show that $\{\rho_1(t), \rho_2(t)\} = (\log t, t - 1)$ is a good choice. This dataset illustrates the flexibility of model (6). With a good choice of the basis functions $\rho_k(t)$, one can model effectively the longitudinal covariates jointly with the corresponding survival times.

Under the accelerated failure time assumption and the parametric longitudinal model (6), the hazard function in (5) now takes the form

$$\lambda\{t|\bar{X}(t)\} = \lambda(t|\beta, b_i) = \lambda_0\{\psi(t; \beta, b_i)\}\dot{\psi}(t; \beta, b_i), \tag{7}$$

where $\lambda_0(.)$ is the unspecified baseline hazard function, and

$$\psi(t; \beta, b_i) = \int_0^t \exp\{\beta X(s)\} ds = \int_0^t \exp\{\beta b_i' \rho(s)\} ds$$

corresponds to the transformation in (4) and (5) with derivative

$$\dot{\psi}(t; \beta, b_i) = \exp\{\beta X(t)\} = \exp\{\beta b_i' \rho(t)\}.$$

To construct the likelihood function, we assume noninformative censoring and measurement schedule t_{ij}, which is also independent of the future covariate history and random

effects b_i. Under these assumptions, the probability mechanisms of both censoring and the measurement schedule can be factorised out of the likelihood function, and the joint observed likelihood for the model made up of (1) and (7) can be expressed as

$$L(\theta) = L(\beta, \mu, \Sigma, \sigma_e^2, \lambda_0)$$
$$= \prod_{i=1}^{n} \left[\int \left\{ \prod_{j=1}^{m_i} f(W_{ij}|b_i, t_i, \sigma_e^2) \right\} f(V_i, \Delta_i|b_i, t_i, \lambda_0, \beta) f(b_i|\Sigma, \mu) db_i \right], \tag{8}$$

where $f(W_{ij}|b_i, t_i, \sigma_e^2)$ and $f(b_i|\Sigma, \mu)$ are the densities of $N\{b_i'\rho(t), \sigma_e^2\}$ and $N(\mu, \Sigma)$ respectively. The function, $f(V_i, \Delta_i|b_i, t_i, \lambda_0, \beta)$, from the survival component of the model is given as

$$f(V_i, \Delta_i|b_i, t_i, \lambda_0, \beta) = [\lambda_0\{\psi(V_i; \beta, b_i)\}\dot{\psi}(V_i; \beta, b_i)]^{\Delta_i} \exp\left\{ - \int_0^{\psi(V_i;\beta,b_i)} \lambda_0(t)dt \right\}. \tag{9}$$

In the remainder of this section we describe the difficulties in baseline estimation. The expression in (9), representing the contribution of the survival component to the joint likelihood, is much more complicated than its counterpart in the Cox proportional hazards model. Under the Cox model, the baseline hazard function does not involve other unknown quantities and is assumed in Wulfsohn & Tsiatis (1997) to take the form of its nonparametric maximum likelihood estimate, which is a point mass function with masses assigned to all uncensored V_i. The parameters representing the baseline hazards in the joint Cox and longitudinal model are thus the collection of all those masses, which has a dimension of the order of the number of subjects n. While this growing parameter size creates theoretical difficulties, it has no computational complication. However, the baseline function under the accelerated failure time model now becomes a computational challenge, as the accelerated failure time model in (5) or (9) excludes discrete survival times and hence the point mass approach for baseline hazards. Moreover, direct maximum likelihood estimation of the baseline hazard function fails for (9), as it involves a set of transformed variables, or baseline failure times, $U_i = \psi(V_i; \beta, b_i)$, which are not observed and involve both the random effects and the unknown parameters β. This makes it difficult to preassign a fixed set of parameters to represent the baseline function $\lambda_0(t)$ in a likelihood setting. In fact, even the issue of maximum likelihood estimation under the time-independent accelerated failure time model has not been resolved.

To circumvent this problem, we assume that $\lambda_0(.)$ is constant between two consecutive estimated baseline failure times, so that $\lambda_0(.)$ is a step function. This makes feasible the EM algorithm described in the next section for imputing the unobserved random effects b_i in (8) and (9). Such a step-function assumption about the baseline hazard function resembles the sieves-method approach to maximum likelihood estimation proposed in Grenander (1981, Ch. 8), and thus provides hope that the resulting procedures in this paper will be quite efficient. The simulation study and data application in §§ 4 and 5 later demonstrate the satisfactory performance of the proposed procedure and algorithm. The theoretical properties of the new procedure constitute a complicated problem and are currently under investigation. In fact, even the simpler procedure in Wulfsohn & Tsiatis (1997) poses theoretical challenges that remains unresolved because of the high-dimensional nature of the problem.

3. THE EM ALGORITHM

3·1. *Introduction*

The joint likelihood in (8) will be maximised via the EM algorithm. The complete data for the ith subject are $(V_i, \Delta_i, W_i, t_i, b_i)$ and the complete-data likelihood is

$$L^*(\theta) = \prod_{i=1}^{n} \left[\left\{ \prod_{j=1}^{m_i} f(W_{ij}|b_i, t_i, \sigma_e^2) \right\} f(V_i, \Delta_i|b_i, t_i, \lambda_0, \beta) f(b_i|\Sigma, \mu) \right]. \tag{10}$$

We will compute the expected loglikelihood of the complete data, conditioning on observed data and current parameter estimates in the E-step, and maximise the conditional expected loglikelihood to update estimates of current parameters in the M-step. This is repeated until the parameter estimates converge.

3·2. *The* M*-step*

For a function h of b_i, let $E\{h(b_i)|V_i, \Delta_i, W_i, t_i, \hat{\theta}\} = E_i\{h(b_i)\}$ be the conditional expected loglikelihood based on the current estimate $\hat{\theta} = (\hat{\mu}, \hat{\Sigma}, \hat{\sigma}_e^2, \hat{\lambda}_0, \hat{\beta})$. By differentiating $E_i\{\log L^*(\theta)\}$, we can derive the following maximum likelihood estimates:

$$\hat{\mu} = \sum_{i=1}^{n} E_i(b_i)/n, \tag{11}$$

$$\hat{\Sigma} = \sum_{i=1}^{n} E_i(b_i - \hat{\mu})(b_i - \hat{\mu})'/n, \tag{12}$$

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n} \sum_{j=1}^{m_i} E_i\{W_{ij} - b_i'\rho(t_{ij})\}^2 \Big/ \sum_{i=1}^{n} m_i. \tag{13}$$

To estimate the baseline hazard function, we need to parameterise λ_0, which is the hazard function of the baseline failure times, U, defined in (4). Ideally, we could approximate λ_0 by step functions, which leads to a natural parameterisation of the baseline hazard function. Since we cannot observe the baseline failure times, we estimate them through (4). Let $T_1, \ldots, T_d$ denote the d distinct observed failure times among the n subjects; that is, the T_i correspond to those distinct V_i with $\Delta_i = 1$. Then the baseline failure times, as specified by (4), for these d subjects are $u_k = \int_0^{T_k} \exp\{\beta b_k' \rho(s)\} ds$, for $k = 1, \ldots, d$. They can then be estimated by plugging in the current estimate of β and the current empirical Bayes estimate of b_k. Let $\hat{u}_k$ denote these estimates in ascending order. We have $0 = \hat{u}_{(0)} \leqslant \hat{u}_{(1)} \leqslant \ldots \leqslant \hat{u}_{(d)}$, and a natural parameterisation of the baseline hazard function as piecewise constants between two consecutive $\hat{u}_j$'s; that is, we restrict the baseline hazard function to take the form

$$\lambda_0(u) = \sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} \leqslant u < \hat{u}_{(j)}\}}. \tag{14}$$

Similarly, the cumulative baseline hazard function Λ_0 can be denoted by

$$\int_0^{\psi(V_i;\beta,b_i)} \lambda_0(s) ds = \int_0^{u_i} \lambda_0(s) ds = \sum_{j=1}^{d} C_j(\hat{u}_{(j)} - \hat{u}_{(j-1)}) 1_{\{\hat{u}_{(j)} \leqslant u_i\}}. \tag{15}$$

Differentiating $E_i\{\log L^*(\theta)\}$ with respect to C_k $(1 \leqslant k \leqslant d)$, we have

$$\frac{\partial}{\partial C_k} E_i\{\log L^*(\theta)\} = \frac{\partial}{\partial C_k} \sum_{i=1}^{n} E_i[\Delta_i \log \lambda_0(u_i) - \Lambda_0\{\psi(V_i; \beta, b_i)\}]$$
$$= 0. \tag{16}$$

If we substitute $\lambda_0(u_i)$ and $\int_0^{u_i} \lambda_0(t)dt$ in (16) by (14) and (15) respectively, (16) becomes

$$\frac{\partial}{\partial C_k} \sum_{i=1}^{n} E_i \left\{ \Delta_i \log \sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} < u_i \leqslant \hat{u}_{(j)}\}} - \sum_{j=1}^{d} C_j(\hat{u}_{(j)} - \hat{u}_{(j-1)}) 1_{\{\hat{u}_{(j)} \leqslant u_i\}} \right\}$$
$$= \sum_{i=1}^{n} E_i \left\{ \Delta_i \frac{1_{\{\hat{u}_{(k-1)} < u_i \leqslant \hat{u}_{(k)}\}}}{\sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} < u_i \leqslant \hat{u}_{(j)}\}}} - (\hat{u}_{(k)} - \hat{u}_{(k-1)}) 1_{\{\hat{u}_{(k)} \leqslant u_i\}} \right\}$$
$$= 0.$$

Therefore, the maximum likelihood estimate for C_k is

$$\hat{C}_k = \frac{\sum_{i=1}^{n} E_i(\Delta_i 1_{\{\hat{u}_{(k-1)} \leqslant u_i < \hat{u}_{(k)}\}})}{\sum_{i=1}^{n} E_i\{(\hat{u}_{(k)} - \hat{u}_{(k-1)}) 1_{\{\hat{u}_{(k)} \leqslant u_i\}}\}}. \tag{17}$$

Now that we have overcome the difficulty in estimating the baseline hazard function, we only have one task left, namely, the estimation of β. This turns out to be elusive as, under the assumption that $\lambda_0(.)$ is piecewise constant, $E_i\{\log L^*(\theta)\}$ is equal to

$$\sum_{i=1}^{n} E_i \left[\Delta_i \log \left(\sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} < u_i \leqslant \hat{u}_{(j)}\}} \right) + \Delta_i \beta\{b_i' \rho(V_i)\} - \sum_{j=1}^{d} C_j(\hat{u}_{(j)} - \hat{u}_{(j-1)}) 1_{\{\hat{u}_{(j)} \leqslant u_i\}} \right]$$
$$+ \sum_{i=1}^{n} E_i\{\log f(b_i|\Sigma, \mu)\} + \sum_{i=1}^{n} E_i \left\{ \sum_{j=1}^{m_i} \log f(W_{ij}|b_i, \sigma_e^2) \right\}. \tag{18}$$

There is no closed-form expression for the maximum likelihood estimate $\hat{\beta}$ in (18) since the u_i's involve β. Furthermore, it is not easy to derive the score for β because of the complexity of the u_i's and the indicator functions that are involved in β in (18). Therefore, instead of using the Newton–Raphson method to obtain the slope for $\hat{\beta}$, one can estimate β by directly maximising the likelihood when β is low-dimensional.

3·3. *The* E-*step*

The M-step above involved E_i, which requires knowledge of $f(b_i|V_i, \Delta_i, W_i, t_i, \hat{\theta})$. This can be obtained through the Bayes rule:

$$f(b_i|V_i, \Delta_i, W_i, t_i, \hat{\theta}) = \frac{f(b_i, V_i, \Delta_i|W_i, t_i, \hat{\theta})}{f(V_i, \Delta_i|W_i, t_i, \hat{\theta})}$$
$$= \frac{f(V_i, \Delta_i|b_i, t_i, \hat{\theta}) f(b_i|W_i, t_i, \hat{\theta})}{\int_{-\infty}^{\infty} f(V_i, \Delta_i|b_i, t_i, \hat{\theta}) f(b_i|W_i, t_i, \hat{\theta}) db_i},$$

where $f(V_i, \Delta_i|b_i, t_i, \hat{\theta})$ is the same as in (10) but with parameters replaced by current estimates and $f(b_i|W_i, t_i, \hat{\theta})$ is the density of a conditional multivariate normal distribution, whose exact form can be derived. To be more specific, let $\rho^* = \{\rho'(t_{i1})\mu, \ldots, \rho'(t_{im_i})\mu\}'$

and $A = \{\rho(t_{i1}), \ldots, \rho(t_{im_i})\}'$. Given t_i, we have

$$\begin{pmatrix} W_i \\ b_i \end{pmatrix} \sim N\left\{\begin{pmatrix} \rho^* \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right\},$$

where $\Sigma_{11} = A\Sigma A'$, $\Sigma_{12} = \Sigma_{21}' = A\Sigma$ and $\Sigma_{22} = \Sigma$. Hence

$$b_i|W_i, t_i, \hat{\theta} \sim N\{\mu + \Sigma_{21}\Sigma_{11}^{-1}(W_i - A\mu), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\}. \tag{19}$$

The empirical Bayes estimate for b_i is thus the estimated mean of (19). Moreover, Monte Carlo integration is used to derive all $E_i(.)$, similarly to Henderson et al. (2000), by generating a number, M, of multivariate normal sequences for $b_i|W_i, t_i, \hat{\theta}$, denoted by $N_i = (N_{i1}, \ldots, N_{iM})'$. Then, for any function $h(.)$ of b_i, we have

$$E_i\{h(b_i)\} = \frac{\int_{-\infty}^{\infty} h(b_i) f(V_i, \Delta_i|b_i, t_i, \hat{\theta}) f(b_i|W_i, t_i, \hat{\theta}) db_i}{\int_{-\infty}^{\infty} f(V_i, \Delta_i|b_i, t_i, \hat{\theta}) f(b_i|W_i, t_i, \hat{\theta}) db_i}$$

$$\simeq \frac{\sum_{j=1}^{M} h(N_{ij}) f(V_i, \Delta_i|N_{ij}, t_i, \hat{\theta})}{\sum_{j=1}^{M} f(V_i, \Delta_i|N_{ij}, t_i, \hat{\theta})},$$

when M is large. The accuracy of the Monte Carlo integration increases as M increases, at the cost of computation time. In order to have higher accuracy and less computing time, we may adopt the Monte Carlo EM method of Wei & Tanner (1990); that is, we use small values of M in the initial iterations of the algorithms, and increase M as the algorithm moves closer to convergence. This strategy is effective in the simulation studies.

3·4. *Summary and remarks*

The EM algorithm can be summarised as follows. Obtain reasonable initial values $\hat{\theta}^{(0)}$ for all parameters and, at the kth stage, carry out the following steps.

Step 1. Estimate b_i by the empirical Bayesian estimate as specified in (19), and then estimate the ordered baseline failure times $\{\hat{u}_{(1)}, \ldots, \hat{u}_{(d)}\}$.

Step 2. Compute (11), (12), (13) and (17) to obtain $\hat{\mu}^{(k)}$, $\hat{\Sigma}^{(k)}$, $\hat{\sigma}_e^{2(k)}$ and $\hat{\lambda}_0^{(k)}$, where E_i in those formulae are performed according to the E-step in § 3·3.

Step 3. Find the maximiser $\hat{\beta}^{(k)}$ of the conditional expected loglikelihood from all neighbouring grid points of the current $\hat{\beta}^{(k-1)}$.

Repeat Steps 1–3 until convergence.

Remark 1. The monotonicity property of the EM algorithm is lost because of the use of the Monte Carlo integration. However, as a result of a suggestion of Chan & Ledolter (1995), under suitable regularity conditions, the EM algorithm will approach the maximiser of the likelihood with high probability, and this probability increases as the Monte Carlo sample size increases.

Remark 2. Since the likelihood function may have multiple modes, it is necessary to choose a variety of initial values to ensure that the global maximum likelihood estimates are obtained. A reasonable initial value is needed to speed up convergence. A simple two-stage procedure can be used to obtain the initial value, with the procedure in F. Hsieh's unpublished manuscript providing the initial estimate for β at the second stage. One could

also apply the last-value-carry-forward technique to implement Hsieh's procedure at the second stage. Of course, any two-stage approach is likely to induce bias because of truncation at the time of death, but can still be a useful source of initial estimates.

Remark 3. Even with all the above precautionary measures, the EM algorithm may still converge slowly, especially if a large number of basis functions is used in (6). It is thus very important to find good but few basis functions, as illustrated in § 5.

3·5. *Bootstrap estimate of the standard errors*

When estimating the standard error of $\hat{\beta}$, we encounter two difficulties. The first is that the exact information matrix of parameters of interest cannot be obtained directly in the EM algorithm. Remedies proposed in Louis (1982) and McLachlan & Krishnan (1997, Ch. 4) approximate the observed Fisher information matrix. These approximations are asymptotically valid for a finite-dimensional parameter space, but we consider the baseline hazard to be unspecified, and the asymptotic validity of such approximations is dubious for an infinite-dimensional parameter space.

The second difficulty is that a promising way of deriving the information matrix is provided by profile likelihood. However, the mixture structure of the joint accelerated failure time model does not allow an explicit profile likelihood. Hence we need to project on to all other parameters, including the infinite-dimensional parameter, λ_0, in order to derive estimated standard errors for $\hat{\beta}$. It is very difficult to derive the projection, which involves the infinite-dimensional parameter λ_0.

In view of the above difficulties, we suggest the use of Efron's (1994) bootstrap technique for missing data to derive the standard error estimates. The following is an outline of the procedure.

Step 4. Generate a bootstrap sample w^* from the original observed data w.

Step 5. The EM algorithm is applied to the bootstrap sample w^* to derive the maximum likelihood estimate $\hat{\theta}^*$.

Step 6. Repeat Steps 4 and 5 B times.

Step 7. Compute $\text{cov}(\hat{\theta}^*) = (B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}_b)(\hat{\theta}_b^* - \bar{\theta}_b)'$, where $\bar{\theta}_b = \sum_{b=1}^{B} \hat{\theta}_b^*/B$.

The example in § 5 supports the use of such bootstrap estimates for standard errors.

4. Simulation studies

We study the performance of the EM procedures in § 3 through simulations with $n = 100$ subjects and 100 simulated samples. In the survival model (5), the baseline function is set to be constant with $\lambda_0 \equiv 0{\cdot}01$, and $\beta = 1$. For the longitudinal component, we consider the linear growth model (6) with $\rho_1(t) = 1$ and $\rho_2 = t$, normal random effects with mean $\mu = (1, 0{\cdot}5)'$, and measurement errors with $\sigma_e^2 = 0{\cdot}25$ in (1). The preliminary scheduled measure times for each subject are $(0, 1, \ldots, 7)$, but no measurement is available after death or censoring time. Three different settings are considered for the variance components, Σ and censoring schemes: (i) $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0{\cdot}01, -0{\cdot}001, 0{\cdot}001)$ and no censoring on scheduled measure times; (ii) the σ_{ij} take the same values as in (i), but the lifetime is subject to censoring according to the exponential distribution with mean 25, which resulted in about 20% censoring among all subjects; (iii) the same setting

as in (ii) except that $\sigma_{22} = 0{\cdot}3$. As a result of the large variation, b_{2i} may become negative in (iii), leading to improper survival distributions with positive point mass at ∞. While this causes no problem as the data would be censored at the censoring time in such a case, they are unnatural in that this assumes infinite survival time, as in the cure model setting. We choose to discard the negative values and the resulting b_i is thus actually generated from a truncated bivariate normal distribution with 35% of the bivariate vectors truncated. This deviation from the normality assumption allows us to check the robustness of our procedure, which assumes a normal random effect.

These three different settings allow us to examine the impact of censoring and violations of the Gaussian random effects model on the performance of the proposed joint accelerated failure time procedure. In the first and second settings the random effects are normally distributed, as assumed, but in the third setting the random effects depart from the normality assumption.

For the first and second settings the results in Table 1 show that the proposed joint accelerated failure time procedure provides approximately unbiased estimators, and that censoring mainly affects the variances of the estimators but not the biases. With setting (iii), the target values are no longer the actual model parameters because of the truncation of the normal random effects. The actual targets were estimated empirically and reported in the 'empirical target' row. These should be the values with which the 'mean' estimates should be compared, and in this case also our procedure provides good estimates for all parameters. Although the estimators for μ_2, σ_{12} and σ_{22} now have much larger standard deviations than their counterparts in settings (i) and (ii), this is probably caused by the increase in the target variance components rather than the stability of the procedures. If we compare the results for settings (ii) and (iii), violation of the normality assumption on the random effects has little impact on the biases of the procedures, and

Table 1: *Simulation study. Results for settings* (i), (ii) *and* (iii)

(*a*) *No censoring, normal random effects*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0·5	0·01	−0·001	0·001	0·25
Mean	1·0075	0·9955	0·5013	0·0087	−0·0011	0·0009	0·2528
SD	0·0945	0·0163	0·0055	0·0015	0·0002	0·0002	0·0135

(*b*) 20% *censoring, normal random effects*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0·5	0·01	−0·001	0·001	0·25
Mean	0·9918	0·9944	0·5015	0·0083	−0·0011	0·0009	0·2516
SD	0·1272	0·0249	0·0056	0·0023	0·0004	0·0002	0·0198

(*c*) 20% *censoring, nonnormal random effects*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0·5	0·01	−0·001	0·3	0·25
Empirical target	1	0·9993	0·6758	0·0104	−0·0058	0·1358	0·2753
Mean	0·9950	1·0007	0·6682	0·0099	−0·0006	0·1627	0·2500
SD	0·1091	0·0140	0·0535	0·0004	0·0036	0·0318	0·0223

Mean, average of 100 Monte Carlo estimates; SD, standard deviation of 100 Monte Carlo estimates.

yet the standard deviation of $\hat{\beta}$ is smaller when the target values of the variance components are bigger. This is intriguing but can be explained by the design feature that larger variance components on the random effects may offer larger information about $\hat{\beta}$ and hence a smaller standard error for $\hat{\beta}$.

The robustness property exemplified with setting (iii) was also observed in Song et al. (2002) and Tsiatis & Davidian (2004) for the joint Cox model setting when the true random effects have bimodal or skew distributions. This is probably because, when there are enough repeated measurements on the longitudinal data, the posterior density of b_i, given the W_i, μ and Σ, has a mode near the true parameters regardless of the random effects distribution. Thus, one could comfortably apply the accelerated failure time procedure in this paper by assuming normal random effects, whenever there are enough measurements on the longitudinal data. However, caution must be exercised when the data are sparse, as departure from the normal random effects assumption may have effects on the estimating procedures.

In Fig. 1 we plot the average estimated cumulative baseline hazard function together with the true one for each simulation setting. All curves ended at the 95% percentile of the true survival distribution. In each setting the estimated function is similar to the true one. Pointwise 95% confidence bands based on the Monte Carlo simulations are also reported in Fig. 1, and all of them include the true function.

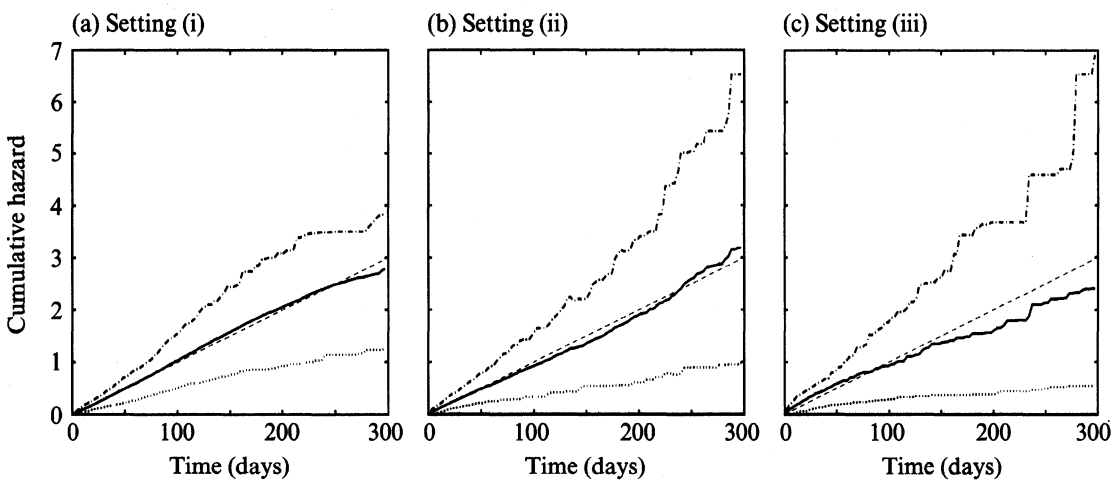

Fig. 1: Simulation study. Estimated cumulative baseline hazard function for settings (i), (ii) and (iii). Each panel shows the true cumulative baseline hazard function, dashed, the average estimated function, solid, and bands made up of the upper, dot-dashed, and lower, dotted, limits of the 95% confidence intervals.

5. Application to medfly fecundity data

5·1. *Introduction*

We apply our procedures to the egg-laying data in Carey et al. (1998), which motivated our model. The original dataset consists of 1000 female Mediterranean fruit flies, known as medflies, for which the numbers of eggs produced daily until death were recorded without missing values. The goal there was to explore the relationship of the pattern of these fecundity curves, $X(t)$, to longevity, as measured by the associated lifetimes of

the medflies. Such information is important because reproduction is considered by evolutionary biologists to be the single most important life history trait besides lifetime itself. This dataset is unusual and is selected for illustration for several reasons.

First, the proportional hazards assumption fails for the most fertile medflies; we use data from the 251 flies that produced more than 1150 eggs in their lifetime. The proportional hazards assumption was rejected by the test based on Schonfeld residuals in S-Plus, as described later. This is not surprising because of the complexity of the reproductive dynamics and its association with lifetime. On the other hand, an accelerated failure time model, as defined in (5), provides a biologically more sensitive model as it reflects covariate risks on an accelerated time scale and involves the cumulative reproductive effects and not just daily effects.

Secondly, this dataset contains the complete event history, the reproductive history in this case, for all experimental subjects, which is rare for data collected in medical longitudinal studies. The complete data setting allows us to discard most of the original data artificially and to apply our procedure to both the complete and incomplete datasets. This allows us to check the stability of the joint accelerated failure time procedure.

5·2. *Fitting the model to the complete medfly data*

A key to the proposed procedure is a suitable parametric longitudinal model. The fecundity profiles of four typical flies are shown in Fig. 2, and suggest the adoption of a Gamma-like parametric model, with individual 'random' shape and scale parameters

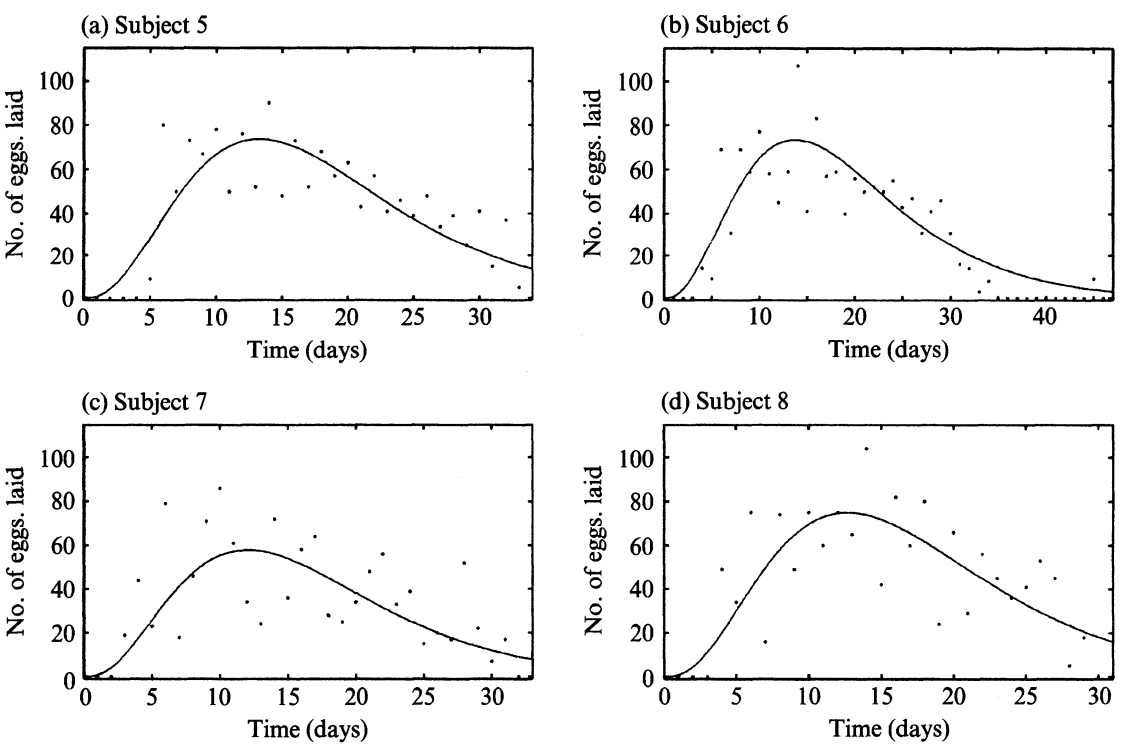

Fig. 2: Medfly data. Individual profiles are fitted by the gamma function. Number of eggs laid by (a) subject 5 is fitted by $t^{2\cdot710}e^{-0\cdot204t}$, (b) subject 6 by $t^{2\cdot652}e^{-0\cdot193t}$, (c) subject 7 by $t^{2\cdot725}e^{-0\cdot226t}$, and (d) subject 8 by $t^{2\cdot803}e^{-0\cdot221t}$. The shape and scale parameters are obtained by least squares.

for the ith fly:

$$W_i(t) = X_i^*(t) + e_i(t), \quad X_i^*(t) = t^{b_{1i}} \exp(b_{2i} t).$$

Here $W_i(t)$ is daily egg-laying, which is subject to random daily fluctuations. The actual underlying fecundity process, $X_i^*(t)$, is not observed, and (b_{1i}, b_{2i}) are the random effects of the ith fly. However, this choice of parametric model for $X_i^*(t)$ yields a nonlinear random effects model and hence it is very complicated to derive a joint likelihood function and conditional expectation in every iteration of the EM algorithm. To overcome this computational difficulty, we apply a logarithmic transformation to both $W_i(t)+1$ and $X_i(t)+1$. The unit constant is added to avoid ill-defined logarithmic function values, since the daily egg-laying of any individual could be zero. Consequently, the final longitudinal model for the ith individual becomes

$$\log(W_{ij} + 1) = X_{ij} + e_{ij}, \tag{20}$$

$$X_{ij} = b_{1i} \log(t_{ij}) + b_{2i}(t_{ij} - 1), \tag{21}$$

where $e_{ij} \sim N(0, \sigma_e^2)$ and $b_i = (b_{1i}, b_{2i})' \sim N(\mu_{2\times 1}, \Sigma_{2\times 2})$, for $i = 1, \ldots, 251$, $j = 1, \ldots, m_i$ and $22 \leqslant m_i \leqslant 99$. Note here that $m_i = T_i$ for the complete medfly data. After taking log transformation of daily egg-laying of those medflies, we test, in S-Plus, the Cox proportional hazards assumption again using the scaled Schonfeld residuals in Grambsch & Therneau (1994) and Therneau & Grambsch (2000, pp. 130–5). The proportional hazards model was rejected at p-value = 0·003. An accelerated failure time survival model is thus proposed, based on its aforementioned biologically appealing feature. The results of the joint accelerated failure time procedure developed in § 3 are summarised in Table 2(a), where the standard error estimate for each parameter is derived from 100 bootstrap samples as described in § 3·5.

Table 2: *Medfly data. Parameter estimation based on* (a) *complete data and* (b) *incomplete data, in each case together with results from* 100 *bootstrap samples under the joint accelerated failure time model*

(*a*) *Complete data*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Fitted value	−0·4340	2·1227	−0·1442	0·3701	−0·0482	0·0068	0·8944
Boostrap mean	−0·4313	2·1112	−0·1429	0·3651	−0·0483	0·0066	0·8958
Boostrap SD	0·0115	0·0375	0·0051	0·0353	0·0002	0·0005	0·0223

(*b*) *Incomplete data*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Fitted value	−0·3890	2·2011	−0·1665	0·2833	−0·0382	0·0051	0·9775
Boostrap mean	−0·3526	2·1986	−0·1575	0·2862	−0·0398	0·0057	0·9712
Boostrap SD	0·0323	0·0461	0·0074	0·0351	0·0046	0·0006	0·0570

SD, standard deviation of 100 bootstrap estimates.

The mean of the 100 bootstrap estimates, as reported in the second row, is close to the estimate based on the data, reported in the first row. This provides positive evidence of the reliability of the bootstrap procedure under the joint modelling framework. Based on the bootstrap standard deviations, all the parameters are highly significant, and the

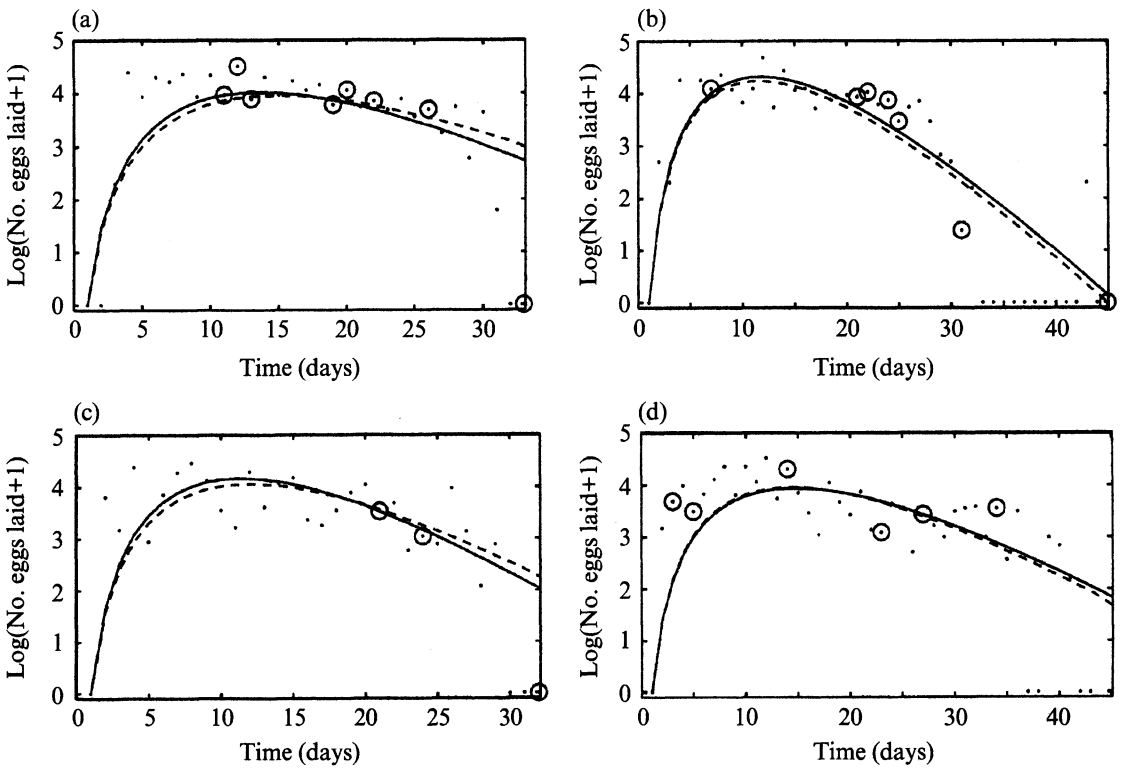

Fig. 3. Fitted fecundity curves, that is log (No. of eggs laid +1) per day, for four medflies based on complete (dots) and incomplete (circled dots) data. The fitted curves based on complete data are shown by dashed lines, and those for incomplete data by solid lines.

negative estimated regression coefficient, −0·4340, suggests that, for highly fertile flies, reproduction activity is positively associated with longevity. In other words, the commonly observed 'cost of reproduction' (Partridge & Harvey, 1985) does not hold for the most fertile flies; in fact, fertility seems to be an indicator of genetic fitness for those flies.

Figure 3 provides the empirical Bayes estimates of the four individual $X(t)$, with b_i estimated from the mean of the bivariate normal distribution in (19). The four fitted curves, denoted by dashed lines, capture the egg-laying trajectories quite well. Figure 4 shows the cross-sectional sample mean of the log daily egg-laying and the mean of the 251 fitted curves. The fitted mean curve, denoted by dashes, is very close to the sample means up until day 60, at which time only 10% of the medflies are still alive. The variation becomes larger afterwards, as expected. We have thus demonstrated the feasibility of the joint models (20) and (21) for female medfly fecundity and survival data.

5·3. *Fitting incomplete medfly data*

We now test our procedure in the presence of censoring and irregular sampling plans. We randomly select 1 to 7 days as the corresponding schedule times for each individual and then add the day of death as the last schedule time. Therefore, a minimum of two and a maximum of eight repeated measurements of egg production are recorded for each medfly, and all other reproduction information is discarded. This resulted in artificially induced irregular sampling plans on the longitudinal data. The data are further censored

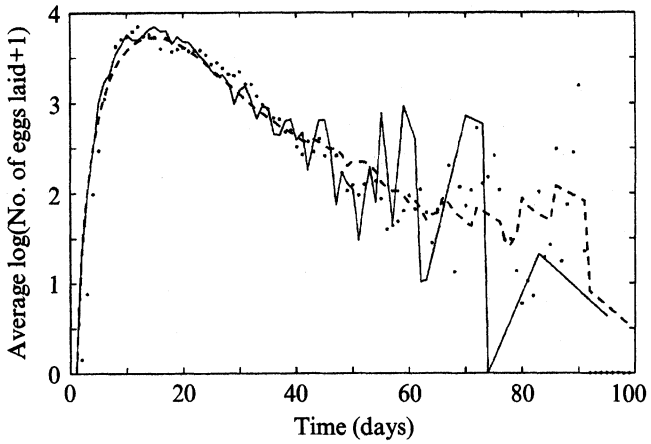

Fig. 4: Medfly data. Fitted cross-sectional mean curves for complete, dashed line, and incomplete, solid line, data. The dots represents the daily observed cross-sectional mean eggs of those medflies that are still alive.

by an exponential distribution with mean 500, which resulted in censoring of lifetimes for 20% of the medflies and many fewer longitudinal measurements for the censored subjects. The joint accelerated failure time procedure is then applied to this incomplete dataset, and the results are presented in Table 2(b). Here again, the bootstrap procedures seem to be effective, all parameters are highly significant, and the point estimates based on the incomplete data are close to those based on the complete data.

The individual fitted curves for the four subjects based on the incomplete data are also shown as solid lines in Fig. 3 and are essentially the same as the fitted curves based on the complete data. The mean of the 251 fitted curves, also based on incomplete data, is shown as a solid line in Fig. 4. While the two fitted mean curves are close to each other until day 50, the impact of the sparsity of the longitudinal data is clear in the high variability of the mean fitted curve based on incomplete data.

6. Discussion

We have demonstrated that our procedure can be insensitive to the normality assumption, but this must not be mistaken for global robustness of the procedure. Like all parametric approaches, joint likelihood is sensitive to model assumptions for the longitudinal covariates, that is the choice of the basis functions, ρ_k. A misspecified functional form of the longitudinal covariates could induce large bias. For example, if instead of (20) and (21) we fit the longitudinal covariates for the medfly data by a simple linear mixed model given by (6) with $\rho(t) = (1, t)'$ and $b_i = (b_{1i}, b_{2i})'$, the estimate of β becomes $-0{\cdot}021$ with standard deviation $0{\cdot}14$, which results in nonsignificance of the fecundity curve for the medfly data.

It is straightforward to extend our procedure to accommodate multivariate time-dependent covariates and/or baseline covariates. Instead of (4) we have

$$U = \psi\{X(T), Z; \beta, \eta\} = \int_0^T \exp\{\beta' X(s) + \eta' Z\} ds,$$

where X is a q-dimensional longitudinal process, β is a q-dimensional vector and η is the regression coefficient vector corresponding to baseline covariates Z. A slight adjustment is required in Step 3 of the summary of the EM algorithm, to indicate finding the maximisers of β and η simultaneously. This can be achieved by using a simplex algorithm (Nelder & Mead, 1965) or simulated annealing (Kirkpatrick et al., 1983).

Important issues remain to be resolved. For instance, the asymptotic theory of the estimators is not yet available. In fact, this is true even for the simpler case of a proportional hazards model. Both problems are currently under investigation. Until reliable estimates for the standard deviations of the estimators are derived, we recommend the use of the bootstrap estimates of standard deviations, which seem to work well in the data illustration.

ACKNOWLEDGEMENT

The authors thank the reviewers for insightful comments. The research of Jane-Ling Wang was supported in part by the U.S. National Science Foundation and National Institutes of Health.

REFERENCES

CAREY, J. R., LIEDO, P., MÜLLER, H. G., WANG, J. L. & 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. *J. Gerontol.: Biol. Sci.* **53**, 245–51.

CHAN, K. S. & LEDOLTER, J. (1995). Monte Carlo estimation for time series models involving counts. *J. Am. Statist. Assoc.* **90**, 242–52.

COX, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

DAFINI, U. G. & TSIATIS, A. A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics* **54**, 1445–62.

EFRON, B. (1994). Missing data, imputation and bootstrap (with Discussion). *J. Am. Statist. Assoc.* **89**, 463–79.

GRAMBSCH, P. M. & THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–26.

GRENANDER, U. (1981). *Abstract Inference.* New York: Wiley.

HENDERSON, R., DIGGLE, P. & DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **4**, 465–80.

KIRKPATRICK, S., GELATT, C. D. JR. & VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–80.

LIN, D. Y. & YING, Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *J. Statist. Plan. Infer.* **44**, 47–63.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226–33.

MCLACHLAN, G. J. & KRISHNAN, T. (1997). *The EM Algorithm and Extensions.* New York: Wiley.

NELDER, J. A. & MEAD, R. (1965). A simplex method for function minimization. *Comp. J.* **7**, 308–13.

PARTRIDGE, L. & HARVEY, P. H. (1985). Costs of reproduction. *Nature* **316**, 20–1.

PAWITAN, Y. & SELF, S. (1993). Modeling disease marker process in AIDS. *J. Am. Statist. Assoc.* **88**, 719–26.

ROBINS, J. & TSIATIS, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time dependent covariates. *Biometrika* **79**, 311–9.

SONG, X., DAVIDIAN, M. & TSIATIS, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58**, 742–53.

THERNEAU, T. M. & GRAMBSCH, P. M. (2000). *Modeling Survival Data.* New York: Springer.

TSIATIS, A. A. & DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447–58.

TSIATIS, A. A. & DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statist. Sinica* **14**, 809–34.

TSIATIS, A. A., DEGRUTTOLA, V. & WULFSOHN, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J. Am. Statist. Assoc.* **90**, 27–37.

WANG, Y. & TAYLOR, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J. Am. Statist. Assoc.* **96**, 895–905.

WEI, G. C. G. & TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and poor man's data augmentation algorithm. *J. Am. Statist. Assoc.* **85**, 699–704.

WULFSOHN, M. S. & TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–9.

YU, M., LAW, N. J., TAYLOR, J. M. G. & SANDLER, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statist. Sinica* **14**, 835–62.

[*Received May* 2004. *Revised March* 2005]