Joint Modelling of Accelerated Failure Time and Longitudinal Data

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

Summary

The accelerated failure time (AFT) model is an attractive alternative to the Cox model when the proportionality assumption fails to capture the relation 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 AFT assumption when covariates are assumed to follow a linear mixed effects model with measurement errors. The procedure is based on maximizing the joint likelihood function where random effects are treated as missing data. A Monte Carlo EM algorithm is employed 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 relation to longevity demonstrate the effectiveness of the new procedure.

Some key words: EM algorithm; measurement errors; missing data; Monte Carlo integration; random effects, 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. A growing interest in the health community is to model both processes simultaneously to explore their relationship and to borrow strength from each component in the model building process. Such a joint modelling approach has emerged as an effective way to utilize the information available on both processes, and has become feasible due to rapidly improving computing environments.

In the joint modelling approach, the longitudinal covariates are usually assumed to be of parametric form with random effects, such as a linear mixed effects model. Moreover, the longitudinal covariate may not be directly observed due to intermittent sampling schedule and/or measurement errors. Let $X(t)$ denote such a longitudinal covariate with additive measurement error, $e(t)$. So 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, as the case of multiple longitudinal covariates and additional time independent covariates can easily be adapted.

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 $\bar{X}(t) = \{X(s) : 0 \leq s < t\}$ 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 are 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 time varying CD4 count and the event time is an

important goal of these experiments and has been studied extensively in the literature, for instance in Pawitan and Self (1993), Tsiatis, DeGruttola and Wulfsohn (1995), Wulfsohn and Tsiatis (1997), and Wang and Taylor (2001).

If there were no measurement errors 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, and it is thus necessary to find alternative approaches. Intuitively, one could overcome both difficulties by imputing the unobserved covariate process, $X(t)$, in the partial likelihood. Such an approach is called "two-stage procedure" in the joint modelling literature, and has been studied in Tsiatis et al. (1995) and Dafini and Tsiatis (1998) among others. This approach encounters bias when the observation of the longitudinal process was interrupted by the event time, that is, when death strikes. In such situations, only measurements before death are available, which results in informative missing longitudinal data. Bias will occur in both the longitudinal and survival components, if unmodified linear mixed effects model procedures were employed to fit the longitudinal component. Various remedies were proposed and the most satisfactory approach is perhaps the joint likelihood approach in Wulfsohn and Tsiatis (1997), who constructed a joint likelihood of (1) and (2) under certain assumptions including 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 and Davidian (2001) through a conditional score approach, and relaxed to a flexible parametric class of smooth density functions in Song, Davidian and Tsiatis (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 and Taylor (2001) consider a similar model as Henderson et al. (2000) and applied a Bayesian framework as well as MCMC methods to fit the joint model. For additional information about joint modelling, see the insightful reviews in Tsiatis and Davidian (2004) and Yu et al. (2004).

So far the literature on joint modelling of survival and longitudinal data only focused on the Cox proportional hazards model to characterize the relation between the longitudinal

covariates and the survival information. There are, however, many situations (such as the fecundity data in section 5) where the proportionality assumption in (2) fails. For such situations an accelerated failure time (AFT) model is a viable alternative. The AFT 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 a time independent covariate and e 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 and Oakes (1984, chapter 5, pages 64-65) propose the following extension of the AFT model:

$$U \sim S_0, \qquad where \quad U = \psi\{X(T); \beta\} = \int_0^T exp\{\beta X(s)\} ds. \tag{4}$$

With this transformation, the survival function for an individual with covariate history $\bar{X}(t)$, is $S\{t|\bar{X}(t)\} = S_0\{\psi(X(t;\beta)\}$. This means that individuals age on an accelerated schedule , $\psi\{X(t);\beta\}$, under a baseline survival function $S_0(\cdot)$. Such a model is biologically meaningful and allows the influence of the entire covariate history on 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

$$\lambda\{t|\bar{X}(t)\} = \lambda_0[\int_0^t exp\{\beta X(s)\} ds] exp\{\beta X(t)\} = \lambda_0[\psi\{X(t);\beta\}]\psi'\{X(t);\beta\}, \tag{5}$$

where $\lambda_0(\cdot)$ is the hazard function for S_0 and ψ' is the first derivative of ψ. Here, U serves the role of a baseline failure time variable and we thus refer to $\lambda_0(\cdot)$ as the baseline hazard function, which is usually left unspecified. Thus, (5) corresponds to a semi-parametric model, which has been studied first by Robins and Tsiatis (1992) using a certain class of rank estimating equations for β. These rank estimates were shown to be consistent and asymptotically normal by Lin and Ying (1995). Recently, Hsieh (2003) proposed an over-identified estimating equation approach to achieve semiparametric efficiency and to

extend (5) to a heteroscedastic version. All this aforementioned work assumes, however, that the entire covariate process, $X(t)$, can be observed without measurement errors.

For the rest of the paper, we consider the joint AFT model as specified by (1) and (5) (or equivalently, (1) and (4)), subject to the further complication that the observation of the longitudinal coavriate 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 $\lambda_0(\cdot)$ in the survival components (4) (or (5)); as well as effective estimators for the model components of the longitudinal process. This is accomplished via the likelihood approach, so one could consider our proposal the counterpart of the approach in Wulfsohn and Tsiatis (1997) for the proportional hazards mode.

As with the traditional time-independent AFT model, the AFT structure in the joint modelling setting is much harder to handle than the proportional hazards model. We assume that the baseline hazard function is a step function in section 2 when specifying the joint likelihood of T and $X(t)$. This is different from the approaches in Tsiatis and Wulfsohn (1997), where the baseline hazard function is assumed to be discrete. The step function structure is prompted by the continuous nature of the AFT model in (5), and it allows us to implement the EM algorithm in section 3. The simulation studies in section 4 show that the proposed estimating procedures perform reasonably well.

Standard errors for the estimator for β turn out to be a difficult issue due to the missing information on random effects in the EM step. We propose a bootstrap method to estimate the standard error of $\hat{\beta}$ and illustrate it through a data set in Section 5, where a case study for this fecundity data from Carey et al. (1998) is discussed. An intriguing parametric model is proposed to model the longitudinal covariate which consists of the daily egg-laying history of each of 251 female Mediterranean fruit flies (medflies). This data is unique in that the entire history of the longitudinal process is available and there is no censoring involved. We can thus artificially select part of the longitudinal data to examine the performance of our procedure. This data also motivate the joint AFT and longitudinal models proposed in this paper.

2. Joint AFT and longitudinal 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 $T_i \leq C_i$, and 0 elsewhere. Without lose 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 $X_i(\cdot)$ are scheduled to be measured (with error) at times t_{ij}, but no measurements are available after the event time. Thus, the measurement schedule of subject i is $\mathbf{t}_i = (t_{ij},\, t_{ij} \leq V_i)$ and there are m_i repeated measurements for subject i, so that $j = 1, ..., m_i$. The measurements for subject i are $\mathbf{W}_i = (W_{ij})$ with measurement error $\mathbf{e}_i = (e_{ij})$, $j = 1, ..., m_i$, where $W_{ij} = X_i(t_{ij}) + e_{ij}$. Therefore, the observed data for each individual is $(V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i)$, with all variables independent across i.

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

$$X_i(t) = \mathbf{b}_i^T \boldsymbol{\rho}(t), \tag{6}$$

where $\boldsymbol{\rho}(t) = \{\rho_1(t), \cdots, \rho_p(t)\}^T$ and $\boldsymbol{\rho}(t)$ are known functions; $\mathbf{b}_i^T = (b_{1i}, \cdots, b_{pi})$ are p-dimensional multivariate normal distributions, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, independent of the measurement errors $\mathbf{e}_i$. The measurement errors, $\mathbf{e}_i$, 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 $\mathbf{b}_i$, which are not observed and 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 $\{\rho_1(t), \cdots, \rho_p(t)\} = (1, \cdots, t^{p-1})$ can be used to include more complicated growth curves model at high computational cost, as the EM steps involve evaluation of p-dimensional integrals. Because of this, only a few random effects are employed in practice

and different base 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 section 5, where we show that $\{\rho_1(t), \rho_2(t)\} = (logt, t-1)$ is a good choice. This data 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 AFT assumption and the parametric longitudinal model (6), the hazard function in (5) now takes the form:

$$\lambda\{t|\bar{X}(t)\} = \lambda(t|\beta, \mathbf{b}_i) = \lambda_0\{\psi(t;\beta,\mathbf{b}_i)\}\psi^{'}(t;\beta,\mathbf{b}_i), \tag{7}$$

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

$$\psi(t;\beta,\mathbf{b}_i) = \int_0^t exp\{\beta X(s)\}ds = \int_0^t exp\{\beta \mathbf{b}_i^T \boldsymbol{\rho}(s)\}ds,$$

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

$$\psi^{'}(t;\beta,\mathbf{b}_i) = exp\{\beta X(t)\} = exp\{\beta \mathbf{b}_i^T \boldsymbol{\rho}(t)\}.$$

To construct the likelihood function, we assume noninformative censoring and measurement schedule t_{ij}, which is also independent of future covariate history and random effects $\mathbf{b}_i$. With such assumptions, both probability mechanisms of censoring and measurement schedule can be factorized out of the likelihood function, and the joint observed likelihood for model (1) and (7) can be expressed as:

$$\begin{aligned} L(\theta) &= L(\beta, \boldsymbol{\mu}, \Sigma, \sigma_e^2, \lambda_0) \\ &= \Pi_{i=1}^n [\int \{\Pi_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \mathbf{t}_i, \sigma_e^2)\} f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) f(\mathbf{b}_i|\Sigma, \boldsymbol{\mu}) d\mathbf{b}_i], \end{aligned} \tag{8}$$

where $f(W_{ij}|\mathbf{b}_i, \mathbf{t}_i, \sigma_e^2)$and $f(\mathbf{b}_i|\Sigma, \boldsymbol{\mu})$ are the density of $N\{\mathbf{b}_i^T \rho(t), \sigma_e^2\}$ and $N(\boldsymbol{\mu}, \Sigma)$ respectively. The function, $f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta)$, from the survival component of the model is given as

$$f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) = [\lambda_0\{\psi(V_i;\beta,\mathbf{b}_i)\}\psi^{'}(V_i;\beta,\mathbf{b}_i)]^{\Delta_i} exp\{-\int_0^{\psi(V_i;\beta,\mathbf{b}_i)} \lambda_0(t)dt\}. \tag{9}$$

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 counter part 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 and Tsiatis (1997) to take the form of its nonparametric MLE, 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 subject size n. While this growing parameter size creates theoretical difficulties, it has no computational complications. However, the baseline function under the AFT model now becomes a computational challenge, as the AFT model in (5) or (9) excludes discrete survival times and hence the point mass approach for baseline hazards. Moreover, direct maximum likelihood estimate for baseline hazard function fails for (9), as it involves a set of transformed variables (or baseline failure time), $U_i = \psi(V_i; \beta, \mathbf{b}_i)$, which are not observed and further involve both the random effects and the unknown parameter 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 MLE under the time-independent AFT model has not been resolved. To circumvent this problem, we assume that $\lambda_0(\cdot)$ is constant between two consecutive estimated baseline failure times, i.e. $\lambda_0(\cdot)$ is a step function. This allows the feasibility of the EM algorithm described in the next section to impute the unobserved random effects $\mathbf{b}_i$'s in (8) and (9). Such a step function assumption on the baseline hazard function resembles the sieves method approach to MLE as proposed in Grenander (1981), and thus provides hope that the resulting procedures in this paper will be quite efficient. The simulation study and data application in sections 4 and 5 later demonstrate the satisfactory performance of the proposed procedure and its computational algorithm. The theoretical properties of the new procedure is a complex problem and is currently under investigation. In fact, even the simpler procedure in Wulfsohn and Tsiatis (1997) poses theoretical challenges and remains unsolved due to the high dimensional nature of the problem.

3. EM Algorithm

The joint likelihood in (8) will be maximized via the EM algorithm. For this, we need to construct the complete data likelihood. The complete data for each subject is $(V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i, \mathbf{b}_i)$ and the complete data likelihood is

$$L^*(\theta) = \Pi_{i=1}^{n}[\{\Pi_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \mathbf{t}_i, \sigma_e^2)\} f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) f(\mathbf{b}_i|\Sigma, \boldsymbol{\mu})]. \tag{10}$$

We will then compute the expected log likelihood of the complete data, conditioning on observed data and current parameter estimates in the E-step, and maximize the conditional expected log likelihood to update estimates of current parameters in the M-step. This is repeated until the parameter estimates converge. The detailed procedure is described in the next two subsections.

3·1. M-step

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

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

$$\hat{\Sigma} = \sum_{i=1}^{n} E_i(\mathbf{b}_i - \hat{\mu})(\mathbf{b}_i - \hat{\mu})^T/n, \tag{12}$$

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i} E_i\{W_{ij} - \mathbf{b}_i^T\ \boldsymbol{\rho}(t_{ij})\}^2 / \sum_{i=1}^{n} m_i. \tag{13}$$

To estimate the baseline hazard function, we need to parameterize 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 parametrization 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: $u_k = \int_0^{T_k} exp\{\beta\ \mathbf{b}_k^T \boldsymbol{\rho}(s)\} ds, k = 1, ..., d$. They can then be estimated by plugging in the current estimate of β and the current empirical Bayes estimate of $\mathbf{b}_k$. Let $\hat{u}_k$ denote these estimates in ascending order. We have $0 = \hat{u}_{(0)} \leq \hat{u}_{(1)} \leq \cdots \leq \hat{u}_{(d)}$, and a natural parametrization 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 :

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

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

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

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

$$\begin{aligned} & \frac{\partial}{\partial C_k} E_i\{logL^*(\theta)\} \\ = & \frac{\partial}{\partial C_k} \sum_{i=1}^{n} E_i[\Delta_i log\ \lambda_0(u_i) - \Lambda_0\{\psi(V_i;\beta,\mathbf{b}_i)\}] \\ = & 0. \end{aligned} \tag{16}$$

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

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

Therefore, the maximum likelihood estimate for C_k is

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

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 elusive as under the assumption that $\lambda_0(\cdot)$ is piecewise constant, $E_i\{logL^*(\theta)\}$ is equal to

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

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

3·2. E-step

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

$$\begin{aligned} & f(\mathbf{b}_i | V_i, \Delta_i, \boldsymbol{W_i}, \mathbf{t}_i, \hat{\theta}) \\ = & \ \frac{f(\mathbf{b}_i, V_i, \Delta_i | \boldsymbol{W_i}, \mathbf{t}_i, \widehat{\theta})}{f(V_i, \Delta_i | \boldsymbol{W_i}, \mathbf{t}_i, \hat{\theta})} \\ = & \ \frac{f(V_i, \Delta_i | \mathbf{b}_i, \mathbf{t}_i, \hat{\theta}) \cdot f(\mathbf{b}_i | \boldsymbol{W_i}, \mathbf{t}_i, \hat{\theta})}{\int_{-\infty}^{\infty} f(V_i, \Delta_i | \mathbf{b}_i, \mathbf{t}_i, \hat{\theta}) \cdot f(\mathbf{b}_i | \boldsymbol{W_i}, \mathbf{t}_i, \hat{\theta}) d\mathbf{b}_i}, \end{aligned}$$

where $f(V_i, \Delta_i | \mathbf{b}_i, \mathbf{t}_i, \hat{\theta})$ is the same as in (10) with parameters replaced by current estimates and $f(\mathbf{b}_i | \boldsymbol{W_i}, \mathbf{t}_i, \hat{\theta})$ is a density of conditional multivariate normal distribution, whose exact form can be derived. More specifically, let $\boldsymbol{\rho}^*= \{\boldsymbol{\rho}^T(t_{i1})\boldsymbol{\mu}, \cdots, \boldsymbol{\rho}^T(t_{im_i})\boldsymbol{\mu}\}^T$ and $A = \{\boldsymbol{\rho}(t_{i1}), \cdots, \boldsymbol{\rho}(t_{im_i})\}^T$. Given $\mathbf{t}_i$, we have

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

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

$$\mathbf{b}_i|\boldsymbol{W_i}, \boldsymbol{t_i}, \widehat{\theta} \sim N\{\boldsymbol{\mu}+\Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{W_i} - A\boldsymbol{\mu}),\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\}. \tag{19}$$

The empirical Bayes estimate or BLUP for $\mathbf{b}_i$ is thus the estimated mean of (19). Moreover, Monte Carlo integration is used to derive all $E_i(\cdot)$, similar to Henderson et al.(2000), by generating a number, M, of multivariate normal sequences for $\mathbf{b}_i|W_i, \boldsymbol{t_i}, \widehat{\theta}$, denoted by $\boldsymbol{N_i} = (N_{i1}, \cdots N_{iM})^T$. Then for any function, $h(\cdot)$ of $\mathbf{b}_i$, we have

$$\begin{aligned} E_i\{h(\mathbf{b}_i)\} &= \frac{\int_{-\infty}^{\infty} h(\mathbf{b}_i) f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \widehat{\theta}) \cdot f(\mathbf{b}_i|\boldsymbol{W_i}, \mathbf{t}_i, \widehat{\theta}) d\mathbf{b}_i}{\int_{-\infty}^{\infty} f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \widehat{\theta}) \cdot f(\mathbf{b}_i|\boldsymbol{W_i}, \mathbf{t}_i, \widehat{\theta}) d\mathbf{b}_i} \\ &\approx \frac{\sum_{j=1}^{M} h(N_{ij}) f(V_i, \Delta_i|N_{ij}, \mathbf{t}_i, \widehat{\theta})}{\sum_{j=1}^{M} f(V_i, \Delta_i|N_{ij}, \mathbf{t}_i, \widehat{\theta})}, \text{when } M \text{ is large.} \end{aligned}$$

The accuracy of the Monte Carlo integration increases as M increases, at the cost of computational time. In order to have higher accuracy and less computing time, we may follow the suggestion for Monte Carlo EM in Wei and Tanner (1990). That is, to use small values of M in the initial iterations of the algorithm, and increase the values of M as the algorithm moves closer to convergence. This strategy is found effective in the simulation studies.

3·3. Summary and remarks

The EM-algorithm can be summarized as follows:

Obtain reasonable initial values for all parameters $\hat{\theta}^{(0)}$, and at the k^{th} step:

1. Estimate $\mathbf{b}_i$ by the empirical Bayesian estimate as specified in (19), and then estimate the ordered baseline failure times $\{\widehat{u}_{(1)}, ...\widehat{u}_{(d)}\}$.

2. Compute (11), (12), (13) and (17) to get $\widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\Sigma}^{(k)}, \widehat{\sigma}_e^{2(k)}, \widehat{\lambda}_0^{(k)}$, where E_i in those formulae are performed according to the E-step in section 3.2.

3. Find the maximizer $\widehat{\beta}^{(k)}$of the conditional expected log likelihood from all vicinal grid points of current $\widehat{\beta}^{(k-1)}$.

Repeat steps 1-3 until all parameters converge.

Computation Remarks:

1. The monotonicity property of the EM algorithm is lost due to the Monte Carlo integrals in the EM algorithm. However, following a suggestion of Chan and Ledolter (1995), under suitable regularity conditions, the EM algorithm will approach the maximizer of the likelihood with high probability, and this probability increases as the Monte Carlo sample size increases.

2. Due to potential multiple modes of the likelihood function, it is necessary to choose various initial values to make sure the global maximum likelihood estimates are obtained. A reasonable initial value is needed to speed up convergence. A simple two-stage procedure can be employed for the initial value, with the procedure in Hsieh (2003) providing the initial estimate for β at the second stage. Alternatively, one could also apply the last-value-carry-forward technique to implement Hsieh's procedure at the second stage. We remind the reader that any two-stage approach is likely to induce biased due to truncation at lifetime, but could be used to gain initial estimates.

3. Even with all precautionary measures taken as above, the EM-algorithm may still take a long time to converge, especially if a large number of basis functions is used in (6). It is thus very important to find good but few basis functions. We illustrate in the case study in section 5 how to accomplish this.

3·4. Bootstrap estimate of the standard errors

When estimating the standard error of $\widehat{\beta}$, we encounter two difficulties. The first is that implementation of the EM algorithm involves missing information, and as noted in Orchard and Woodbury (1972) the exact information matrix of parameters of interest can not be obtained directly in the EM algorithm. This is the so called "missing information principle". Various remedies have been proposed in Louis (1982) and McLachlan and Krishnan (1997, chapter 4) by approximating the observed Fisher information matrix. It is noted that these approximations are asymptotically valid for a finite dimensional parameter space. Since we consider the baseline hazard to be unspecified, the asymptotic validity of such approximations is dubious for infinite dimensional parameter space.
The second difficulty is that a promising way to derive the information matrix is provided by profile likelihood. However, the mixture structure of the joint AFT model results in no explicit profile likelihood. Hence we need to project onto all other parameters, including the infinite dimensional parameter, λ_0, to derive estimated standard errors for $\hat{\beta}$. This projection, which involves the infinite dimensional parameter λ_0, is very difficult to derive.

Due to the above difficulties, we suggest using a bootstrap technique for missing data by Efron (1994) to derive the standard error estimates. The following is an outline of the procedure:

1. Generating bootstrap sample w^* from original observed data w.

2. The EM algorithm is applied to the bootstrap sample w^* to derive the MLE $\widehat{\theta}^*$

3. Repeat step 1 and 2 B times.

4. Compute $Cov(\widehat{\theta}^*) = 1/(B-1)\sum_{b=1}^{B}(\widehat{\theta}_b^* - \overline{\theta}_b)(\widehat{\theta}_b^* - \overline{\theta}_b)^T$, where $\overline{\theta}_b = \sum_{b=1}^{B}\widehat{\theta}_b^*/B$

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

4. Simulation studies

We study the performance of the EM-procedures in section 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 $\lambda_0 \equiv 0.01$, 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 $\boldsymbol{\mu} = (1, 0.5)^T$, and measurement errors with $\sigma_e^2 = 0.25$ in (1). The preliminary scheduled measure times for each subject are $(0, 1, \cdots, 7)$, but no measurement are available after death or censoring time. Three different settings are considered for the variance components, Σ and censoring schemes: (i) $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.01, -0.001, 0.001)$, and no censoring on scheduled measure times; (ii) with the same values of σ_{ij} as (i), but the lifetime is subject to censoring by exponential distribution with mean 25. This resulted in about 20% censoring among all subjects. (iii) same setting as (ii) except $\sigma_{22} = 0.3$. Because of the larger 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 like in the cure model setting. We choose to discard the negative values and the resulting $\mathbf{b}_i$ 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 exam the impact of censoring and violations of the Gaussian random effects model on the performance of the proposed joint AFT procedure. In the first and second setting the random effects are normally distributed as assumed, but in the third setting the random effects depart from the normality assumption.

Table 1. Simulation (i) with no censoring and 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

Table 2. Simulation (ii) with 20% censoring and 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

Table 3. Simulation (iii) with 20% censoring and random effects that are truncated bivariate normal distribution.

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
parameter values	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

Table 1 and Table 2 show the simulation results of the first and second setting respectively. The proposed joint AFT procedure provides approximately unbiased estimates in both settings, and censoring mostly affects the variances of the estimators but not the biases. In Table 3, simulation of the third setting, the original parameter values reported in the second row are no longer the actual model parameters due to the truncation of the normal random effects. The actual targets were estimated empirically and reported in the third row. This should be the actual base of comparison for the mean estimates reported in the fourth row. As can be seen from Table 3, the proposed joint AFT procedure also resulted in good estimates for all parameters. Although the estimates for μ_2, σ_{12} and σ_{22} now have much larger standard deviations than their counterparts in Tables 1 and 2, this is probably due to the increase in the target variance components rather than the stability of the procedures. Comparing Table 3 to Table 2, violation of the normality assumption on random effects has little impact on the biases of the procedures, 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 bigger variance components on the random effects may offer larger information on $\hat{\beta}$ and hence a smaller standard error for $\hat{\beta}$.

To summarize, the simulation results reported in Table 1 , 2 and 3 reveal that the estimates for β are approximately unbiased, and so are all the other parameter estimates. This is true even when the random effects are not normally distributed (cf. Table 3), suggesting the robustness of the joint likelihood approach. This robustness property was also observed in Song et al. (2002) and Tsiatis and Davidian (2004) for the joint Cox model setting when the true random effects have bimodal or skew distributions. This is probably due to the fact that when there are sufficient repeated measurements on the longitudinal data, the posterior density of $\mathbf{b}_i$ given the $\mathbf{W}_i, \boldsymbol{\mu}, \Sigma$, has a mode near the true parameters regardless of the random effects distribution. Thus, one could comfortably apply the AFT procedure in this paper by assuming normal random effects, whenever there are enough measurements on the longitudinal data. Caution, however, must be taken when the data are sparse, as departure from the normal random effects assumption may have due effects on the estimating procedures.

In addition to those estimates in Table 1-3, we plot the average estimated cumulative baseline hazard function vs. the true one for each simulation setting. All curves ended at the 95% percentile of the true survival distribution. Figure 1 (a), (b), and (c) show that these average estimated cumulative baseline hazard functions derived from our proposed approach are all close to the true ones. Pointwise 95% confidence bands based on the Monte carlo simulations are also reported in Figure 1, and all of them include the true cumulative baseline hazard functions. Therefore, the proposed EM-algorithm also provides approximately unbiased estimates for the cumulative baseline hazard function.

5. Application to Medfly fecundity data

We apply our procedures to the egg-laying data in Carey, et al. (1998), which motivated our joint AFT model. The original data set consists of 1000 female Mediterranean fruit flies (medflies), for which number of eggs produced daily until death were recorded without missing and censoring. The goal there was to explore the relation of the pattern

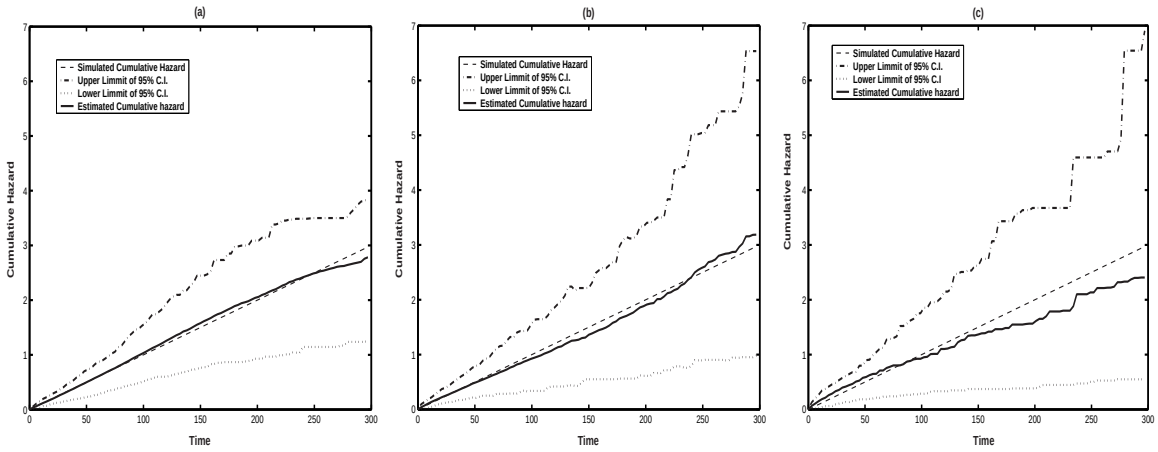

Figure 1: *Estimated cumulative baseline hazard function. Figure (a)-(c) are plotted under the settings of simulations (i)-(iii) respectively.*

of these fecundity curves, $X(t)$, to longevity, as measured by the associated lifetime of the medflies. Such information is important because reproduction is considered by evolutionary biologists as the single most important life history trait besides lifetime itself. This data set is unusual and selected for illustration for several reasons.

First, the proportional hazards assumption fails for medflies that are most fertile, those in the highest quartile of lifetime reproduction (measured by total number of eggs produced in a lifetime). We use data of the 251 flies that produced more than 1150 eggs in their lifetime. This choice is motivated by issues in the study of longevity in aging research, as these flies are most successful in terms of reproduction. The proportional hazards assumption was rejected by the test based on Schonfeld residuals in S-Plus as described later. This is not surprising due to the complexity of the reproductive dynamics and its association to lifetime. A simple proportional hazards assumption fails to capture their relation. An AFT model, as defined in (5), on the other hand provides a biologically more sensible model as it reflects covariate risks on an accelerated time scale and involves the cumulative reproductive effects and not just daily effects.

Secondly, this data set contains the complete event history (reproductive history in this case) for all experimental subjects, which is rare for data collected in medical longitudinal studies. The complete data setting allow us to artificially discard most of the original data

and fit our procedure on both the complete and incomplete data sets by the joint AFT procedure. With this contrast, we could check the stability of the joint AFT procedure.

5·1. Fitting complete medfly data

A key to the proposed procedure is a suitable parametric longitudinal model. Towards this goal, we examine the individual fecundity curves and its cross-sectional mean curve (taken as the daily sample means). The original 251 fecundity curves are very noisy and hence it is difficult to examine the overall shape of the fecundity curves, if we plot all the 251 curves in one figure. However, they all express a strong mode between day 10 to 20 and then tamper off to zero towards the end of lifetime. A sample of four flies are selected and their fecundity profiles are shown in Figure 2. The mode represents peak reproduction, which is expected, so we tried to fit these fecundity curves with unimodal smooth functions that have zero as asymptote. Using a least squares method, the Gamma functions seem to provide good approximations as illustrated in Figure 2.

Those individual fitted curves are gamma functions with different scale and shape parameters. Therefore, a gamma function with random shape and random scale parameters seems appropriate as an initial longitudinal model:

$$W(t) = X^*(t) + e(t), \quad X^*(t) = t^{b_1} exp(b_2 t).$$

Here $W(t)$ is daily egg-laying, which are subject to random daily fluctuations. The actual underlying fecundity process, $X^*(t)$, is not observed, and (b_1, b_2) are the random effects. However, this choice of parametric model for $X^*(t)$ yields a nonlinear random effects model and hence it is very complicated to derive joint likelihood function and conditional expectation in every iteration of the EM algorithm. To overcome this computational difficulty, we apply logarithmic transformation to both $W(t) + 1$ and $X(t) + 1$. The constant one is added to avoid ill-defined logarithmic function values, since daily egg-laying of each individual could be zero. Consequentially, the final longitudinal model for

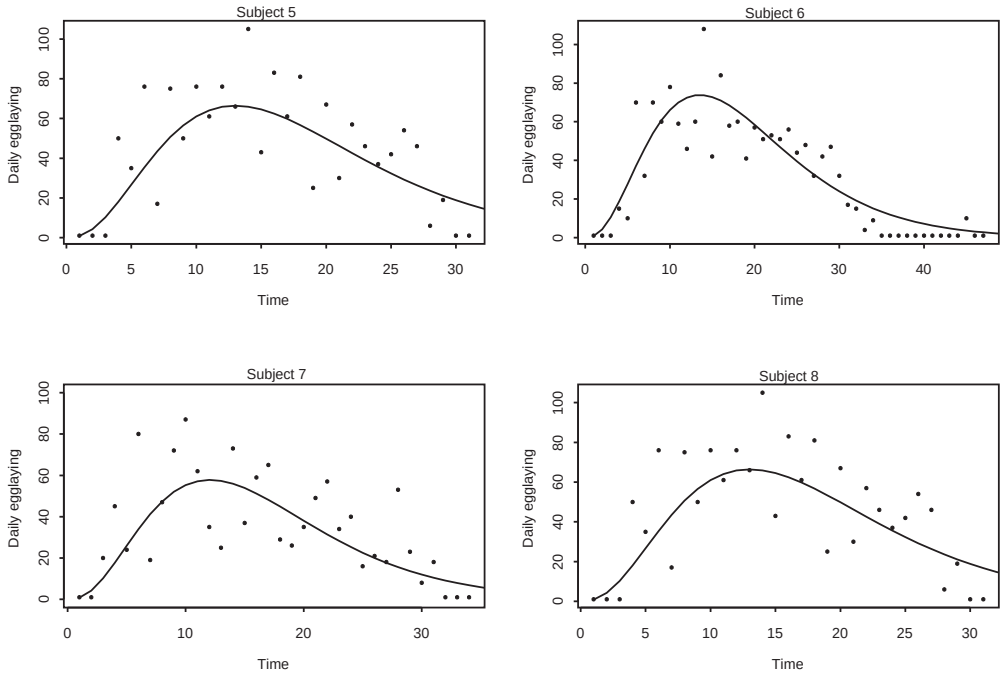

Figure 2: *Individual profiles are fitted by the gamma function. Daily egg-laying of subject 5 is fitted by* $t^{2.710}e^{-0.204}$*, subject 6 by* $t^{2.652}e^{-0.193}$*, subject7 by* $t^{2.725}e^{-0.226}$*, and subject 8 by* $t^{2.803}e^{-0.221}$*.*

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 $e_{ij} \sim N(0,\sigma_e^2)$; $\mathbf{b}_i = (b_{1i}, b_{2i})^T \sim N(\boldsymbol{\mu}_{2\times1}, \Sigma_{2\times2}), i = 1, ..., 251, j = 1, ..., m_i$ and $22 \leq m_i \leq 99$. Note here that $m_i = T_i$ for the complete medfly data. After taking log transformation on daily egg-laying of those medflies, we test, in S-plus, the Cox proportional hazard assumption again using the scaled Schonfeld residuals in Grambsch and Therneau(1994, 2000). The proportional hazards model was rejected at P-value=0.003. An AFT survival model is thus proposed as the alternative based on its aforementioned biological appealing feature. The results of the joint AFT procedure in Section 3 are summarized in Table 4, where the standard error estimate for each parameter is derived by 100 bootstrap samples as described in Section 3·5.

Table4. The parameter estimates derived from the original complete data and 100 bootstrap samples under the joint AFT model.

	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
bootstrap mean	-0.4313	2.1112	-0.1429	0.3651	-0.0483	0.0066	0.8958
bootstrap SD	0.0115	0.0375	0.0051	0.0353	0.0002	0.0005	0.0223

The mean of the 100 bootstrap estimates, as reported in the third row, is close to the estimate based on the data (reported in the second row). This provides positive evidence towards the reliability of the bootstrap procedure under the joint modelling framework. Based on the bootstrap SD reported in the last row, all the parameters are highly significant, and the 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 and Harvey (1985)) does not hold for the most fertile flies. In fact, fertility seems to be an indicator for genetic fitness for those flies.

Fig. 3 provides the empirical Bayes's estimate (or BLUP) of the four individual $X(t)$, with $\mathbf{b}_i$ estimated from the mean of the bivariate normal distribution in (19). The four fitted curves (dashed lines) capture the egg-laying trajectories quite well. Fig. 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 (dashed lines) is very close to the sample means up to day 60, where 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.

5·2. Fitting incomplete medfly data

So far, we have applied our procedure to the complete data set, which involves no censoring and contains the complete covariate history. Since most longitudinal studies in

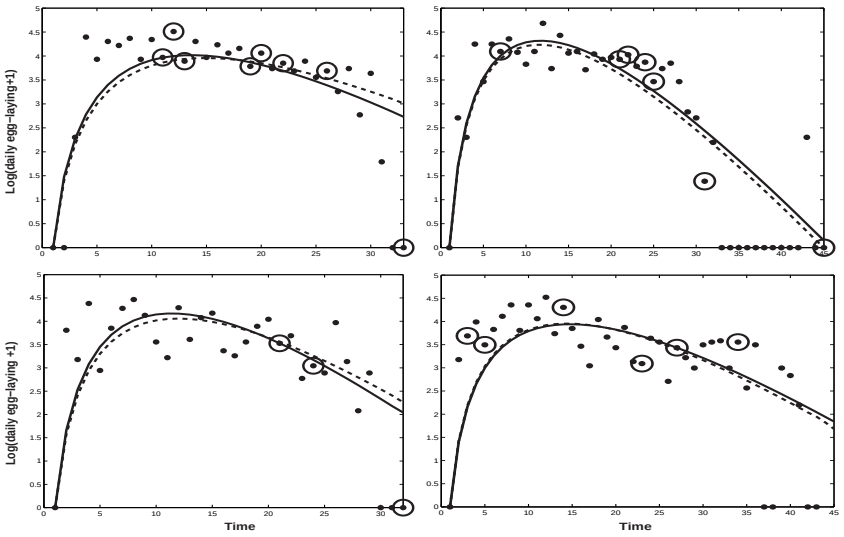

Figure 3: *Fitted fecundity curves for four medflies based on complete (dots) and incomplete (circled dots) data. The dashed lines are the fitted curves based on complete data, and the solid lines for incomplete data.*

clinical trials or medical follow up studies result in incomplete data either through censoring or irregular sampling plan, we want to check the performance of our procedure under these common sampling schemes. We thus 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 2 and a maximum of 8 repeated measurements on the number of egg production are recorded for each mefly, and all other reproduction information is discarded. This resulted in artificially induced irregular sampling plans on the longitudinal data. The sub data set is further censored by an exponential distribution with mean 500, which resulted in censoring of lifetimes for 20 % of the medflies and much fewer longitudinal measurements for the censored subjects. The joint AFT procedure is then applied to this incomplete data set, and the results are presented in Table 5.

Table 5. The parameter estimates derived from incomplete data and 100 bootstrap samples under the joint AFT model.

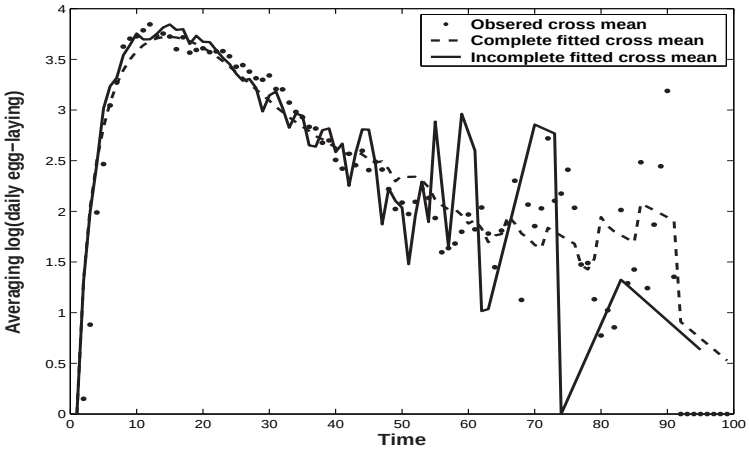

Figure 4: *Fitted Cross-sectional mean curves for complete and incomplete data. The dots represents the daily mean eggs of those that are still alive.*

	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
bootstrap mean	-0.3526	2.1986	-0.1575	0.2862	-0.0398	0.0057	0.9712
bootstrap SD	0.0323	0.0461	0.0074	0.0351	0.0046	0.0006	0.0570

Here again, the bootstrap procedures seems to be effective, all parameters are highly significant, and the estimates based on the incomplete data are close to those based on the complete data. The standard deviations in Table 5 are all much larger than those in Table 4 because a large proportion of information is lost due to the unavailability of the measurements.

The fitted individual curves for the four subjects based on the incomplete data are also shown in Figure 3 (solid lines), and they are essentially the same as the fitted curve based on the complete data (dashed lines). The mean of the 251 fitted curves, also based on incomplete data, is shown in Figure 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 prominently expressed through the high variability of the mean fitted curve based on incomplete data. Overall, we can comfortably claim that the joint AFT and longitudinal procedure proposed in this paper handle incomplete data very well even when the majority of the longitudinal covariates are not available.

6. Discussion and Conclusion

We have demonstrated the applicability of the proposed joint likelihood approach, and that it is insensitive to the normality assumption, if rich information is available on the longitudinal data, meaning that reasonably many repeated measurements are available on the subjects. However, this must not be mistaken for a 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 base functions, ρ_k. 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 which is (6) with $\boldsymbol{\rho}(t) = (1, t)^T$ and $\mathbf{b}_i = (b_{1i}, b_{2i})^T$, the estimate of β becomes -0.021 with standard deviation 0.14, which results in insignificance of the fecundity curve for the medfly data.

Practically, a data set may contain multivariate time dependent covariates and/or baseline covariates, such as treatment status , sex, etc. In these situations, the extension of the proposed joint AFT procedure is straightforward and we may consider the transformation (4) as:

$$U = \psi\{X(T), Z; \beta, \eta\} = \int_0^T exp\{\beta^T X(s) + \eta^T Z\} ds$$

where X is a q-multivariate longitudinal process and β is a q-dimensional vector, η is the regression coefficient vector corresponding to baseline covariates Z. A slight adjustment of the EM algorithm is required in step 4 of the summary of EM algorithm by finding the maximizer of β and η simultaneously. This can be achieved by using a simplex algorithm in Nelder and Mead (1965) or method of simulated annealing in Kirkpatrick et al(1983).

We have proposed a viable joint modelling approach for accelerated failure time data and longitudinal and time-independent covariates. To our knowledge, this is the first attempt of such a joint modelling approach. There are obviously many remaining issues to be resolved. For instance, the asymptotic theory of the estimates is not yet available. In fact, this is also not available even for the simpler case of a proportional hazards model.

Both are challenging technical problems currently under investigation. Until reliable estimates for the standard deviations of the estimators are derived, we recommend to use the bootstrap SD estimates as they 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 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. *Journal of Gerontology–Biological Sciences* **53**, 245-251.

Chan, K.S., & Ledolter, J. (1995). Monte Carlo estimation for time series models involving counts. *Journal of American Statistical Association* **90**, 242-252.

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

Cox, D.R. & Oakes, D. (1984). *Analysis of survival data.* Chapman & Hall, London.

Dafini, U.G. & Tsiatis, A.A. (1998). *Evaluating surrogate markers of clinical outcome measured with error. Biometrics* **54**, 1445-1462.

Efron, B. (1994). Missing data, imputation and bootstrap(with discussion).*Journal of American Statistical Association* **89**, 463-479.

Grambsch P.M. & Therneau T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrics* **81**, 515-526.

Grenander, U. (1981). *Abstract inference*. Wiley, New York.

Henderson, R.,Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **4** , 465-480.

Hsieh, F. (2003). Lifetime regression model with time-dependent covariate. I: semiparametric efficient inference on identical time scale model. Manuscript

Kirkpatrick, S., Gelatt, C.D. Jr. & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Lin, D.Y. & Ying, Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal of statistical planning and inference* **44**, 47-63.

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

McLachlan, G. J. & Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308-313.

Orchard, T. & Woodbury, M. A. (1972). A missing information principle: the theory and applications. In *Processing of the 6th Berkeley Symposium on Mathematical Statistics and Probability Vol. 1* Berkeley, California: University of California Press, 697-715.

Partridge, L. & Harvey, P.H. (1985). Costs of reproduction. *Nature* **316**, 20-21.

Robins, J. & Tsiatis, A.A. (1992). Semiparametric estimation of an accelerated failure time model with time dependent covariates. *Biometrika* **79**, 311-319.

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

Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data* . Spinger

Tsiatis, A.A. & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447-458.

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

Tsiatis, A.A., Degruttola, V. and Wulfsohn, M.S. (1995). Modelling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American statistical association* **90**, 27-37.

Wang, Y. & Taylor, J.M.G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of American Statistical Association* **96**, 895-905.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and poor man's data augmentation algorithm. *Journal of American Statistical Association* **85**, 699-704.

Wulfsohn, M. S. & 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. & Sandler H.M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, **14**, 835-862.