Joint Modelling Of Time-to-event And Longitudinal Data

By

YI-KUAN TSENG

B.E. (National Taipei Teachers College, 1992)
M.S (National Sun Yat-sen University, 1994)
M.S (University of California, Davis, 2002)

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in

Statistics

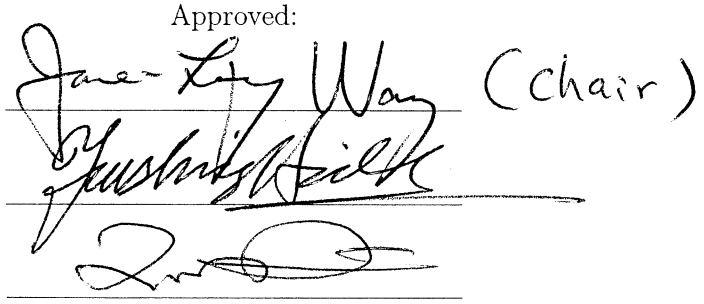
in the

OFFICE OF GRADUATE STUDIES

of

UNIVERSITY OF CALIFORNIA
DAVIS

Approved:



Committee in Charge

2005

Abstract

Jointly modelling the survival time and its longitudinal covariates has recently generated considerable interest. The longitudinal processes are typically assumed to follow a mixed effects model which is linked to the survival time through the Cox's proportional hazards model. Among the various estimating approaches, the maximum likelihood approach based on the joint likelihood of the observed survival and longitudinal data is perhaps the most satisfying approach. We investigate several intriguing issues including robustness of the MLEs against departure from the normal random effects assumption, and difficulties with profile likelihood approach to provide reliable estimates for the standard errors of the MLEs. We propose to use bootstrap procedures to estimate these standard errors and illustrate its effectiveness. Furthermore, the accelerated failure time model is proposed as an alternative to the proportional hazards model when the proportionality assumption fails to capture the relationship between the survival time and longitudinal corvariates. We propose a joint likelihood approach to model the accelerated failure time and its longitudinal covariates simultaneously, where the random effects of the longitudinal processes 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.

Key words: EM algorithm; Joint modelling; Missing data; Nonparametric Maximum likelihood; Posterior density; Profile likelihood; Monte Carlo integration; Survival data.

Contents

Introduction

In biomedical research, it has become increasingly common to observe the event time of interest, called survival time, along with longitudinal covariates. The relationship between these variables is of interest, with the primary focus in survival analysis being modelling the effect of key longitudinal covariates on the survival times of patients. Due to several complications (See Patwitan and Self (1993), Tsiatis, DeGruttola and Wulfson (1995), Wulfsohn and Tsiatis (1997)), traditional approaches, including the partial likelihood approach for Cox Proportional hazards model, encounter difficulties when time-dependent (longitudinal) coavriates are used to model survival times. Joint modelling the survival and longitudinal data emerges as an effective way to overcome these difficulties. It is also a more efficient way to utilize the information available on both types of data as illustrated in the simulation studies in Henderson, Diggle and Dobson (2000)) and Tsiatis and Davidian (2001). So far, attention has been concentrated on Cox model for the survival data and simple parametric random effects model, such as linear mixed effects models, for the longitudinal data. No parametric assumptions were imposed on the baseline hazard rates following the semiparametric spirit of Cox model. EM algorithms were often employed to impute the missing random effects of the longitudinal covariates in the joint likelihood of the survival and longitudinal data. These last two aspects, namely the nonparametric feature of the baseline hazards and the missing data stucture, render complications in statistical inference procedures. No formal asymptotic results, other than some heuristic arguments, have been established to validate the inference procedures. The purpose of the thesis is to explore several challenging issues arise from joint model of time-to-event data and its longitudinal covariates. Three issues are elaborated. The first one pertains to the difficulties to obtain standard error formulae for existing parametric estimates in the literature. We point out a gap in the literature where standard errors were derived using profile likelihood approach. The second issue is the intriguing robustness of likelihood approach against

the normality assumption of the random effect distribution as observed by various authors in numerical studies. This robust feature makes the likelihood approach very attractive and we provide some insights to this phenomenon. The third issue pertains to situations when the proportional hazards assumption fail. So far in the literature the joint modelling approach has exclusively been developed under the proportional hazards assumption, which often does not fit the data in practice. We develop a new joint modelling methodology under the proposed accelerated failure time (AFT) assumption.

In chapter 1, the joint model consists of two components, the Cox's proportional hazards model is used to describe the survival information and the mixed effect model is used to fit the time dependent covariates. we focus on the discussion concerning the difficulties in statistical inference and the robustness phenomenon to the distribution assumption of random effects for the longitudinal covariates. The difficulty in deriving efficiency is caused by several complications. The number of parameters used for the unspecified baseline hazard function is the same order as the sample size. Since the number of parameters increases as the sample size increasing, the classic parametric maximum theory is not valid here. Moreover, the mixture structure of the likelihood function of joint model results in no explicit profile and hence the profile likelihood approach is not promising. Consequently, we suggest a bootstrap technique to overcome the difficulty. As to the robustness phenomenon of likelihood approach, we provide a explanation via posterior distribution and Laplace approximation that if the longitudinal covariates is not too sparse or the measurement error is not too large the robustness to the distribution of random effects holds. We illustrate this through a simulation study and show how the robustness feature hinges upon these properties.

In chapter 2, when the proportionality assumption fails, we propose an accelerated failure time model as an alternative to Cox's model. The likelihood function under AFT assumption requires continuous lifetime distribution, hence the baseline hazard functions cannot be treated the same way as the Cox's model, where the maximum

likelihood estimator for the baseline hazard function is a discrete function. We used instead a step function to approximate the baseline hazard function in the likelihood. Another complication for the baseline hazards estimation is that Monte Carlo EM algorithm is employed to derive point estimates and a bootstrap technique, as the suggestion in chapter 1, is used to obtain standard error estimates.

Chapter 1

Cox Models For The Event Time

1.1 Introduction

In medical follow up or biological longitudinal studies, the trajectory of a biomarker process and other key covariates are often observed along with the time to event of interest, called survival or failure time. Modelling the survival time based on time-independent or longitudinal covariates is of keen interest, and so is understanding of the key biomarker/covariate processes. Both can be accomplished through an approach now termed "Joint Modelling". The most widely explored joint modelling approach consists of a parametric mixed effects models for the longitudinal covariate process, which is linked to the survival time through the Cox's proportional hazards regression model.

If the entire history of the longitudinal covariate is available, partial likelihood can be employed readily to estimate the regression coefficients of the covariates and there would be no complication in modelling the survival times. However, modelling of the longitudinal process separately encounters difficulties for internal longitudinal covariates, as defined in Kalbfleisch and Prentice (2002, section 6.3.2), due to the abrupt ending of the data collection process when failure (or death) strikes an individual. Thus, the observation of the longitudinal process is truncated by lifetime, causing informative missing longitudinal data.

The aforementioned ideal situation to model the survival process separately is often not feasible as the longitudinal covariate processes are typically measured intermittently at different times for the experimental subjects, and often with measurement errors. Any naive effort to impute the missing or mis-measured internal longitudinal covariates without taking into account the survival information will induce bias for the regression

coefficient estimates, if unmodified partial likelihood are used.

Several remedies have been proposed in the literature, including a parametric likelihood approach in Patwin and Self (1993), the two-stage methods in Tsiatis, Degruttola, and Wulfsohn (1995), the semiparametric joint likelihood approaches in Wulfsohn and Tsiatis (1997) and Henderson, Diggle, and Dobson (2000), a conditional score method in Tsiatis and Davidian (2001), and a semiparametric likelihood approach in Song, Davidan and Tsiatis (2002) among many others including several Bayesian approaches. We refer the readers to two insightful surveys in Tsiatis and Davidian (2004) and Yu, et al. (2004). Simulations in Henderson et al. (2000) demonstrate the advantage to combine information available from both the longitudinal and survival processes versus marginal approaches to model them separately in stages. Numerical studies in Tsiatis and Davidian (2004) also suggests that the "joint maximum likelihood" (ML) approach in Wulfsohn and Tsiatis (1997), hereafter abbreviate as **WT**, is among the most satisfactory ones to combine information. We focus on this approach in what follows.

Briefly described, the approach in **WT** is semiparametric in that no parametric assumptions were imposed on the baseline hazard function. This is in line with the semiparametric spirit of the partial likelihood approach of Cox (1972), and is accomplished by using the nonparametric maximum likelihood estimate (see Kiefer and Wolfowitz(1956)) of the baseline hazard function under the traditional Cox model, which turns out to be point mass function with positive masses assigned only at uncensored survival times. Thus, a large number of parameters, same as the number of uncensored survival times, are needed to model the baseline hazards. The joint likelihood of the observed longitudinal and survival data for a subject was then specified by conditioning on the unobserved random effects of the longitudinal process of this subject. These random effects were assumed to be normally distributed and treated as missing data and estimated via EM-algorithm iteratively together with other maximum likelihood estimates for the other parameters, including those from the baseline hazard functions.

This approach successfully remove the bias discussed above and is computationally feasible for moderate number of random effects.

An attractive feature of the joint ML approach in **WT** is that although it utilizes the normality assumption of the random effects, simulation results conducted in Tsiatis and Davidian (2004) reveal that it is fairly robust against departure of this assumption. This "robustness" phenomenon is persistent for various types of random effects distributions, and is perhaps the most compelling evidence towards the superiority of this approach. They called for further investigation on this intriguing robustness feature, and we provide an explanation in section 4. The same simulation results in that paper also led them to heuristically suggest that this joint ML approach should be efficient for statistical inference in the joint modelling setting.

It turns out that the efficiency issue along with the asymptotic distributions of the parametric estimates in the joint modelling framework pose theoretical challenges. Even the variance of the parametric estimators, including the regression estimators in the Cox model, are difficult to obtain due to several technical complications. The first is the large number of parameters used in **WT** to model the baseline hazard function, which is of the same order as the number of subjects. Thus, classical parametric maximum likelihood theory does not apply here. Secondly, EM-algorithm were employed to estimate the missing random effects resulting in information loss from the complete likelihood. The actual amount of information loss due to imputing the missing random effects is difficult to estimate, and this further complicates the asymptotic studies . Thirdly, all estimates were obtained via an implicit profile approach during the iterative estimating process, and there is no semiparametric profile theory (See Murphy and Van der Vaart (2000)) that is readily available in this setting.

Regarding the second difficulty, a heuristic argument was made in the literature to estimate the observed Fisher information by using the slope (or derivative) of the implicit profile score (defined in (1.9)) in the last step of the EM-algorithm. This

approach relies on two assumptions: that statistical inferences based on the maximum likelihood approach are applicable in the setting of joint modelling, and that the EM-algorithm produces profile likelihood estimates with no or small amount of information loss. We show in section 3 that there is indeed information loss in the EM-algorithm, and the loss might be significant as illustrated in the numerical study in Section 5 and the data analysis in Section 6. This is consistent with the general missing information principle in the EM-algorithm. As a result, the inverse matrix based on the aforementioned heuristic profiled slope in EM-algorithm underestimates the actual sampling errors. Moreover, the implicit profile likelihood is not the real profile likelihood. Therefore the net Fisher information for any parameter needs to be calculated by applying further projection procedures as in the standard semiparametric approaches, and this remains an open and challenging problem. We propose in this paper to use bootstrap procedure to estimate the standard errors.

The rest of the paper is organized as follows: Section 2 provides the background of joint model structure and joint likelihood. In section 3, we show that no explicit profile likelihood function is available, and EM-algorithm leads to implicit profile estimates. However, the standard deviations, hereafter referred to as the standard error, of these estimator turn out illusive due to the implicit profile feature. Important discrepancy between the valid statistical precision and the heuristic profiled slope in EM-algorithm is theoretically demonstrated. Such discrepancy is also confirmed in simulation study in Section 5 as well as in real data analysis via bootstrapping method in Section 6. In section 4, we show when to expect the parametric estimators in **WT** to be robust against departure from the normal prior distribution assumption on the random effect structure. Section 7 summarizes the findings and future directions.

1.2 Joint Likelihood Function

Without loss of generality, we assume a single time dependent covariate for the survival time, as the case of multiple longitudinal covariates and additional time-independent covariates can be handled straightforward. Let $X_i(t)$ be the longitudinal covariate and L_i the survival time for the i-th individual, with $i = 1, .., \ldots, n$. The survival time L_i is subject to usual independent random censoring by C_i, and for the i-th individual we observe (V_i, Δ_i), where $V_i = min(L_i, C_i)$ and $\Delta_i = 1(L_i \leq C_i)$.

In reality, the longitudinal process are scheduled to be measured at discrete time points and the schedule for each individual can be different. If $X_i(t)$ is an internal covariate (Kalbfleisch & Prentice 2002, page 198), it is reasonable to assume that the original schedule times $\{t_{i1}, t_{i2}, ...\}$ for subject i will be terminated at the endpoint when the event time, V_i, occurs. Hence, the schedule that is actually observed is $\mathbf{t}_i, = \{t_{i1}, \ldots, t_{im_i}\}$ with $t_{im_i} \leq V_i < t_{im_i+1}$, and no longitudinal measurements are available after time V_i leading to informative missing longitudinal data. Consequently, marginal inference of the longitudinal process, such as a two-stage procedure, without incorporating the survival information will lead to bias. This calls for the need to jointly modelling the longitudinal and survival processes.

Moreover, the longitudinal covariate, $X_i(t)$ which affects the survival time is often subject to measurement errors or random fluctuation. Instead, another process, $W_i(t)$, is observed at the scheduled time, $\mathbf{t}_i$, with the following measurement error model:

$$W_{ij} = X_{ij} + e_{ij}, \quad j = 1, \ldots, m_i, \tag{1.1}$$

where $W_{ij} = W_i(t_{ij})$, $X_{ij} = X_i(t_{ij})$, and e_{ij} is measurement error independent of X_{ij}. It is customary in the literature to assume that e_{ij} follows a normal distribution, $N(0, \sigma_e^2)$, but this is not necessary for our purpose. For computational simplicity, the structural assumption of $X_i(t)$ has often been a linear random effects model, such as $X_i(t) = b_{1i} + b_{2i}t$. However, such a parametric model restriction is also not necessary for the discussion

in this paper, and X_i can be either a parametric, nonparametric or semiparametric process with k-dimensional random effects $\mathbf{b}_i$. Henceforth, we use $X_i(t;\mathbf{b}_i) = X_i(t)$ to denote the unknown structure built in the covariate process, and assume a parametric prior distribution g_α for the random effect $\mathbf{b}_i$. We adopt the boldface symbols $\mathbf{X}_i = (X_{i1}, \cdots, X_{im_i})$ and $\mathbf{W}_i = (W_{i1}, \cdots, W_{im_i})$ to denote the vectors associated with the longitudinal processes.

To model the survival part, we need to introduce the covariate history process $\bar{X}_i(t;\mathbf{b}_i) = \{X_i(s;\mathbf{b}_i)|0 \le s \le t\}$ and this is related to the survival time L_i through the time-dependent Cox proportional hazards model with the hazard function for the ith individual specified by:

$$\lambda_i(t|\bar{X}_i(t;\mathbf{b}_i)) = \lambda_0(t)e^{\beta X_i(t;\mathbf{b}_i)}. \tag{1.2}$$

The "joint Models" of the longitudinal and survival parts is specified by (1.1) and (1.2), with (1.2) linking the two parts. The two model components, W_i and L_i are independent of each other once the random effect b_i is specified. The observed data for the ith individual is $(V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i)$, with all variables independent across i. The parameter that specifies the model is $\theta = (\beta, \lambda_0, \alpha, \sigma_e^2)$, where λ_0 is really a nonprametric parameter.

Since the covariate history is determined by the random effects $\mathbf{b}_i$, the p.d.f. corresponding to (1.2) and subject i can be expressed as:

$$\begin{aligned} &f(V_i, \Delta_i|\mathbf{b}_i, \beta, \lambda_0) = \\ &[\lambda_0(V_i)\exp\{\beta X_i(V_i;\mathbf{b}_i)\}]^{\Delta_i}\exp[-\int_0^{V_i}\lambda_0(t)\exp\{\beta X_i(t;\mathbf{b}_i)\}dt]. \end{aligned} \tag{1.3}$$

Let $f(W_{ij}|\mathbf{b}_i, \sigma_e^2)$ denote the p.d.f. of the longitudinal measurement W_{ij} given the random effects $\mathbf{b}_i$ for subject i, and g_α denote the p.d.f. of the random effects $\mathbf{b}_i$. The joint likelihood based on the observed data $(V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i)$ $i = 1, \ldots\ldots, n$, as derived in

WT, is

$$L(\theta) = \prod_{i=1}^{n}[\int_{R^k}\{\prod_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \sigma_e^2)\}g_\alpha(\mathbf{b}_i)f(V_i, \Delta_i|\mathbf{b}_i, \beta, \lambda_0)d\mathbf{b}_i]. \tag{1.4}$$

Strictly speaking, the time schedule, t_{ij}, of the longitudinal measurements may carry information of the parameters, so some assumptions on the "uninformative" nature of the censoring and time schedule, t_{ij}, of longitudinal measurements are required for (1.4). A detailed account can be found in section 3 of Tsiatis and Davidian (2004), and we assume that these assumptions are satisfied so that (1.4) is a legitimate likelihood. Direct maximization of the joint likelihood in (1.4) is impossible in this semiparametric setting due to the nonparametric component, λ_0, which leads to an unbounded likelihood function. However, due to the fact that the term $\prod_{i=1}^{n}\{\lambda_0(V_i)\}^{\Delta_i}$ can be factored out of the integration sign of $L_(\theta)$, the nonparametric maximum likelihood estimate (MLE) of $\lambda_0(t)$, as defined in Kiefer and Wolfowitz (1956), has discrete masses at each uncensored event time V_i. This was reported in Johansen (1983) for the Cox model and continues to hold for the joint models setting with the likelihood function in (1.5). This implies that the joint likelihood approach in **WT** to parameterize the baseline hazard function by point mass function with positive mass at uncensored V_i indeed yields nonparametric MLE for the baseline hazard function and other parameters in θ as reported in Song, Davidian and Tsiatis (2002).

Hereafter, the baseline function, λ_0, in (1.4) is replaced by its nonparametric MLE, a point mass function with positive mass only at the uncensored V_i. Consequently, the semiparametric problem in (1.4) has been converted to a parametric problem with the parameter representing λ_0 being of dimension in the order of n. Such a parametrization comes at no cost to the baseline hazards assumption, which still enjoy the flexibility of the semiparametric approach as in Cox model. Simulation results reported in several papers mentioned in section 1 suggest very satisfactory performance of the nonparametric MLE approach in **WT**, and led to the speculation that it might be semiparametric efficient. However, the fact that the dimension of the parameters in (1.4) being of the

same order as the sample size poses theoretical challenges as no standard asymptotic theory apply to this setting. In fact, there are no theoretical results so far that apply to any of the estimators in θ in the joint modelling setting. Moreover, no precision estimates is available to evaluate the standard deviation of these point estimator either. We focus in the next section on the last issue.

1.3 Relation of EM-algorithm and Fisher Information.

Expectation-maximization (EM)-algorithm was employed in **WT** to maximize the joint likelihood (1.4), which involves the unobserved random effects $\mathbf{b}_i$. This is a common approach in random effects model to treat the unobserved random effects as missing data within the EM-algorithm. In the special case considered in **WT**, where X_i follows a linear mixed effects model with multivariate normal random effects and normal measurement errors, close form expressions are available in the M-step for all the parameters in θ except for β, where Newton-Raphson algorithm was needed to perform the maximization operation. See formulae (3.1)-(3.4) and the discussion afterwards on page 333 of **WT** for details. In the more general setting considered in this paper, close form solutions to some or all of the parameters in θ may not exist and Newton-Raphson algorithm needs to be employed to those parameters. While this adds computational cost, they do not trigger further theoretical complications in statistical inference. The theoretical challenge in the joint modelling setting, including the one in **WT**, attributes to two sources. The first is the aforementioned high dimensional nature of the baseline hazards parameter, which is of the same order as the sample size. This renders difficulties in asymptotic theory for statistical inference. The usual parametric asymptotic arguments for MLE do not apply here. Profile likelihood approach would be an alternative but it encounters difficulties as well. This is the second source of complication

which will be elaborated below.

1.3.1 Implicit Profile Estimates

Under the noninformative assumption on censoring and measurement schedule as in Tsiatis and Davidian (2004) and following the EM principal, we can decompose the likelihood (1.4) into two parts:

$$L(\theta) = L_1(\alpha, \sigma_e^2) L_2(\theta), \tag{1.5}$$

where

$$L_1(\alpha, \sigma_e^2) = \prod_{i=1}^{n} [\int \{\prod_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \sigma_e^2)\} g_\alpha(\mathbf{b}_i) d\mathbf{b}_i], \tag{1.6}$$

$$L_2(\theta) = \prod_{i=1}^{n} L_{2i}(\theta) = \prod_{i=1}^{n} [\int g(\mathbf{b}_i|\mathbf{W}_i, \mathbf{t}_i; \alpha, \sigma_e) f(X_i, \Delta_i|\mathbf{b}_i, \beta, \lambda_0) d\mathbf{b}_i]. \tag{1.7}$$

The first part, which is the marginal distribution of $(\mathbf{W}_1, \cdots, \mathbf{W}_n)$, involves the longitudinal components only. The survival component was reflected only in the second part, $L_2(\theta)$, which involves all parameters. Knowing that λ_0 is the hardest one to estimate due to its high dimensionality and the arguments in section 2 leading to the form of the nonparmetric MLE of λ_0 suggest that we should first calculate the NPMLE, denoted as $\hat{\lambda}_0(\alpha, \sigma_e, \beta)$, given the parameter $(\alpha, \sigma_e, \beta)$. Next, this NPMLE is substituted into $L_2(\theta)$ to produce a profile likelihood $L(\alpha, \sigma_e, \beta, \hat{\lambda}_0(\alpha, \sigma_e, \beta))$. Maximizing this likelihood would then yield estimates of $(\alpha, \sigma_e, \beta)$.

However, for the profile approach to work here as in the classic parametric setting, the profiled likelihood $L(\alpha, \sigma_e, \beta, \hat{\lambda}_0(\alpha, \sigma_e, \beta))$ with the NPMLE $\hat{\lambda}_0(\alpha, \sigma_e, \beta)$ in place should not involve λ_0. Unfortunately, this is not the case here because the NPMLE $\hat{\lambda}_0(\alpha, \sigma_e, \beta)$ can not be solved explicitly here under the random effect structure. This is why EM-algorithm was employed to update the profile likelihood estimate for $\lambda_0(t)$ as

$$\hat{\lambda}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i 1_{(V_i=t)}}{\sum_{j=1}^{n} E_j[\exp\{\beta X_j(t; \mathbf{b}_j)\}] 1_{(V_i \geq t)}} \tag{1.8}$$

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

$$h(\mathbf{b}_i|V_i,\Delta_i,\mathbf{W}_i,\mathbf{t}_i;\theta)=\frac{g(\mathbf{b}_i|\mathbf{W}_i,\mathbf{t}_i;\alpha,\sigma_e^2)f(V_i,\Delta_i|\mathbf{b}_i,\beta,\lambda_0)}{\int g(\mathbf{b}_i|\mathbf{W}_i,\mathbf{t}_i;\alpha,\sigma_e^2)f(V_i,\Delta_i|\mathbf{b}_i,\beta,\lambda_0)d\mathbf{b}_i}. \tag{1.9}$$

We can now see that (1.8) yields an implicit profile estimate since E_j involves $\lambda_0(t)$. The implication of this complexity is that although a point estimation of θ can be derived via the EM-algorithm as in **WT**, the usual profile approach to calculate the Fisher information cannot be employed here. The asymptotic covariance matrix for $(\alpha,\sigma_e^2,\beta)$ could not easily be evaluated through derivatives of the implicit profile likelihood as it now involves the chain rule. We will resume this information issue after we discuss the remaining profile estimates.

Likewise, the maximum "profile" likelihood estimates of α is also in implicit forms:

$$0=\sum_{i=1}^{n}E_i\{\frac{\partial}{\partial\alpha}\ log\ \ g_\alpha(\mathbf{b}_i)\},$$

since E_i involves α itself. Same argument applies to the estimate of σ_e^2, but the estimation of β is more complicated. To see this, consider $L_2(\theta)$ in (1.7), the relevant component in the likelihood for β. The score equation for β is thus

$$S_\beta=\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log L_{2i}(\theta)=\sum_{i=1}^{n}E_i\{S^c(\beta;\lambda_0,\mathbf{b}_i)\}, \tag{1.10}$$

where

$$\begin{aligned} S^c(\beta:,\lambda_0,\mathbf{b}_i) &= \sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(V_i,\Delta_i|\mathbf{b}_i,\beta,\lambda_0), \qquad (1.11)\\ &= \sum_{i=1}^{n}\Delta_i X_i(V_i;\mathbf{b}_i)-\int_0^{V_i}X_i(t;\mathbf{b}_i)\exp\{\beta X_i(t;\mathbf{b}_i)\}\lambda_0(t)dt. \end{aligned}$$

Again, the maximum profile score of β is in implicit form by substituting $\lambda_0(t)$ by $\hat{\lambda}_0(t)$

in equations (1.8), and denoted as

$$\begin{aligned} S_\beta^{IP} &= \sum_{i=1}^{n} E_i\{S^c(\beta; \hat{\lambda}_0(t), \mathbf{b}_i)\} \\ &= \sum_{i=1}^{n} \Delta_i[E_i\{X_i(V_i; \mathbf{b}_i)\} - \frac{\sum_{j=1}^{n} E_j[X_j(V_i; \mathbf{b}_i) \exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}}{\sum_{j=1}^{n} E_j[\exp\{\beta X_j(V_i)\}]1_{(V_j \geq V_i)}}. \end{aligned} \tag{1.12}$$

The EM-algorithm then iterative between the E-step, to evaluate the conditional expectations E_i with parameters values obtained from the previous iteration, $\hat{\theta}^{(k-1)}$; and the M-step, to update the estimate values via the above score equation. Since (1.12) has no close form solution, Newton-Raphson method is performed as:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + I^{-1}_{\hat{\beta}_{k-1}} S^{IP}_{\hat{\beta}_{k-1}}, \tag{1.13}$$

where $S^{IP}_{\hat{\beta}_{k-1}}$ is the value of the incomplete profile score given in (12) with $\beta = \hat{\beta}_{k-1}$, and the slope $I^{-1}_{\hat{\beta}_k}$ is obtained through the following working formula

$$\begin{aligned} I_\beta^W &= \sum_{i=1}^{n} \{ \frac{\sum_{j=1}^{n} E_j[X_j(V_i; \mathbf{b}_i)^2 \exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}}{\sum_{j=1}^{n} E_j[\exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}} \\ &- (\frac{\sum_{j=1}^{n} E_j[X_j(V_i; \mathbf{b}_i) \exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}}{\sum_{j=1}^{n} E_j[\exp\{\beta X_j(V_i; \mathbf{b}_i)\}]1_{(V_j \geq V_i)}})^2 \}. \end{aligned} \tag{1.14}$$

The above iterative procedure is implemented until a convergence criterion is met. We next examine what has been achieved in the EM-algorithm and what has not.

1.3.2 Fisher Information

The iterative plan of EM-algorithm should achieve the consistency of parameters $(\alpha, \beta, \sigma_e^2)$ and the cumulative baseline hazard function $\lambda_0(t)$. The working formula I_β^W in (1.14) at the last step of the EM-algorithm has been suggested in the literature to provide the precision estimate for the standard deviation (standard error) of the β-estimator. However, this working formula is derived by taking the partial derivative of the implicit profile score equation (1.12) with respect to β by treating the conditional expectation $E_i[\cdot]$ as if it does not involve β. This induces two gaps. First, E_i does involve β through

the posterior density, so the proper way to take the derivative of (1.12) should involve the multiplication rule. Secondly, to obtain the correct Fisher information of β, projection method needs to be employed on the Hessian matrix, which has 4x4 block matrices corresponding to the four parameter components. To be more specific and take a simpler case for illustration, assume that α and σ_e^2 are known for the moment so there are only two parameters β and λ_0. Let $I_{\beta\beta} = -\frac{\partial^2}{\partial\beta^2}\log L_2(\theta)$ $I_{\beta\lambda_0} = -\frac{\partial^2}{\partial\beta\partial\lambda_0}\log L_2(\theta)$, and define $I_{\lambda_0\lambda_0}$ and $I_{\lambda_0\beta}$ similarly. The correct projection to reach the Fisher information for β would be:

$$I_{\beta\beta} - I_{\beta\lambda_0}[I_{\lambda_0\lambda_0}]^{-1}I_{\lambda_0\beta}.$$

It is now clear that the correct Fisher information in the real situation is difficult to compute as the projection would be with four parameters involved and the Hessian matrix would be huge due to the high dimension of λ_0. We recommend bootstrapping to estimate the standard deviation for all finite dimensional parameters $(\alpha, \sigma_e, \beta)$ and illustrated it in Section 6.

We close this section by demonstrating that the working SE, I_β^w, is smaller than the sample Fisher information of β. Hence a statistical inference based on I_β^w would be too optimistic, and the discrepancy could be large as illustrated numerically in section 5.

Note that the true Hessian is

$$I_{\beta\beta} = -\frac{\partial^2}{\partial\beta^2}\log L_2(\theta) = \sum_{i=1}^{n} I_{i(\beta,\beta)}, \tag{1.15}$$

where

$$\begin{aligned} I_{i(\beta,\beta)} &= -\frac{\partial^2}{\partial\beta^2}\log L_{2i}(\theta) = -\frac{\partial}{\partial\beta}E_i[S^c(\beta;\lambda_0,\mathbf{b}_i)] \\ &= E_i[-\frac{\partial}{\partial\beta}S^c(\beta;\lambda_0,\mathbf{b}_i) - E_i[S^c(\beta;\lambda_0,\mathbf{b}_i)S^h(\beta;\lambda_0,\mathbf{b}_i)], \end{aligned} \tag{1.16}$$

where

$$\begin{aligned} S^h(\beta;\lambda_0,\mathbf{b}_i)) &= \frac{\partial}{\partial\beta}\log h(\mathbf{b}_i|V_i,\Delta_i,\mathbf{W}_i,\mathbf{t}_i;\theta) \\ &= S^c(\beta;\lambda_0,\mathbf{b}_i) - E_i\{S^c(\beta;\lambda_0,\mathbf{b}_i)\} \end{aligned} \tag{1.17}$$

is the the score function pertaining to posterior density $h(\mathbf{b}_i|V_i, \Delta_i, \mathbf{W}_i, \mathbf{t}_i; \theta)$. With this equation, it is interesting to note that the second term in (1.16) is equal to $Var_i[S^c(\beta; \lambda_0, \mathbf{b}_i)]$, the conditional variance with respect to the posterior density, which is also equal to the amount of Fisher information of β contained in this posterior density.

The first term in (1.16) is the sample Fisher information of β pertaining to the likelihood of $f(X_i, \Delta_i|b_i, \beta)$ in the parametric Cox model, and summing them over i yields the working SE in (1.14). That is,

$$I_\beta^W = \sum_{i=1}^{n} E_i\{-\frac{\partial}{\partial\beta} S^c(\beta; \lambda_0, \mathbf{b}_i)\}.$$

It now follows from (1.15)-(1.17) that

$$I_{\beta\beta} = I_\beta^W - \sum_{i=1}^{n} Var_i\{S^c(\beta; \lambda_0, \mathbf{b}_i)\}. \tag{1.18}$$

The implication of this relation is that under the joint modelling with random effect structure on the covariate $\mathbf{X}_i$, there is a lost and unrecoverable amount of information. Therefore, I_β^W is not the true Hessian, and would be too small comparing to the true one, $I_{\beta\beta}$.

This missing information principle, termed by Orchard and Woodbury (1972), applies to all the finite dimensional parameters: α and σ_e^2, as well as $\lambda_0(t)$ in the setting considered in this paper. Hence Fisher information could not be evaluated along the iterative procedure of EM-algorithm. The implicit profiled estimates serve as base for the EM-algorithm to lead to a maximum likelihood estimate of θ. But, this likelihood is not a ground for proper statistical inferences to be derived. Unfortunately, the actual amount of loss, $\sum_{i=1}^{n} Var_i\{S^c(\beta; \lambda_0, \mathbf{b}_i)\}$ in (1.18) for β, is complicated and traditional asymptotics for Maximum likelihood approach and the profile likelihood theory that prescribe the asymptotic precision via Fisher information are not directly applicable in the joint modelling literature. A valid profile likelihood theory under this semiparametric setting needs to be developed and this is a rather challenging task.

1.4 Robustness of the Likelihood Approach

.

Although no model assumption needs to be imposed on the baseline hazard function in **WT**, both $f(W_{ij}|\mathbf{b}_i, \sigma_e^2)$ and the random effects distribution g_α in (1.4) are assumed to have normal distribution there. Such a parametric assumption is required for a likelihood based procedure and the normality assumption in fact simplified the computations in the joint modelling setting. Typically, one would expect that a parametric procedure be sensitive to the model assumption. However, simulation results reported in Song et al. (2002) and Tsiatis and Davidian (2002) repeatedly demonstrate robustness of the procedure in **WT** for the survival parameter, β, and this robust feature goes beyond the local contamination scenario in conventional robust statistics settings. The procedure in **WT** is nearly as efficient as the semiparametric random effects procedure in Song et al. (2002) if the actual random effects is a mixture of multivariate normal distributions. Our simulation in Section 5 for random effects from a truncated normal distribution also reveals satisfactory performance of the likelihood based procedure in **WT**. In fact, as long as the longitudinal data is not very sparse, the robustness feature persists for all kinds of alternative random effects distributions, such as those with heavy tails, are skewed or bimodal. Tsiatis and Davidian (2004) called for further investigation of this intriguing robustness feature, and we show in this section that this phenomenon in fact relies on the richness of the longitudinal measurements. The likelihood based procedure might be sensitive to the distributional assumption of random effects if only a few longitudinal measurements are available for each subject.

To illuminate on this, we refer to the likelihood function decomposition in (1.5)-(1.7). It follows from (1.6) and (1.7) that the likelihood contributed by the ith individual is

$$L_i^*(\theta) = f(\mathbf{W}_i; \alpha, \sigma_e) E_i^*\{f(V_i, \Delta_i)|\mathbf{b}_i; \theta)\}, \tag{1.19}$$

where $f(\mathbf{W}_i; \alpha, \sigma_e)$ is the marginal density of $\mathbf{W}_i$ and $E_i^*\{f(V_i, \Delta_i|b_i; \theta)\}$ denotes the

conditional expectation with respect to the posterior density of b_i given $\mathbf{W}_i$, which is:

$$g(\mathbf{b}_i|\mathbf{W}_i;\alpha,\sigma_e^2) = f(\mathbf{W}_i|\mathbf{b}_i;\sigma_e^2)g_\alpha(\mathbf{b}_i)/f(\mathbf{W}_i;\alpha,\sigma_e^2).$$

Note that this posterior density is different from the one in (1.9) in the EM-step, which is the posterior density of b_i given all the data, not just $\mathbf{W}_i$. Two facts emerge from this likelihood function (19). First, $E_i^*\{f(V_i,\Delta_i|\mathbf{b}_i;\theta)\}$ carries information on the longitudinal data. If it is ignored, the marginal statistical inference based on $f(\mathbf{W}_i;\alpha,\sigma_e^2)$ alone would be inefficient or even biased, the latter if there is informative missing longitudinal data as discussed in section 1. This shed light on how a joint modelling approach eliminate the bias incurred by a marginal approach, and why it is still preferable even in the absence of such a bias. The latter was also demonstrated numerically via the simulations presented in Table 1.1 of Henderson et al. (2000), but (1.19) provides a theoretical insight.

Secondly, the random effect structure, g_α, and $\mathbf{W}_i$ are relevant to the information of survival parameters, β and λ_0, only through the posterior density $g(\mathbf{b}_i|\mathbf{W}_i,\alpha,\sigma_e^2)$. Analytically, the shape of the posterior density $g(\mathbf{b}_i|\mathbf{W}_i,\alpha,\sigma_e^2)$ should be unimodal when $\mathbf{W}_i$ provides relative large amount of information for $\mathbf{b_i}$, and this occurs when reasonably large number of longitudinal measurements were taken per subject. By applying the Laplace approximation, (Tierney and Kadane (1986)), $E_i^*\{f(V_i,\Delta_i|\mathbf{b}_i;\theta)\}$ can be approximated to a certain precision by a normal density having the same location of the mode and curvature of $g(\mathbf{b}_i|\mathbf{W}_i,\alpha,\sigma_e^2)$ at the mode. This explains why **WT**'s procedure was robust against departure from the prior assumption as observed in several reports, see for example, the simulation results in Song et al. (2002) and Tsiatis and Davidian (2004). There were only at most 13 repeated measurements per subject in these simulations, but apparently they carry sufficient information for the 2-dimensional $\mathbf{b}_i$ to warrant the robustness observed there. Note, however, that the prior assumption may be crucial when the longitudinal measurements are sparse per subject.

1.5 Simulation Study

In this section, we examine the gap between the working SE formula, $(I_\beta^W)^{-1/2}$, given in (1.14) and the sample standard deviation in a simulation study.

The longitudinal covariate is set to be $X_i(t) = b_{1i} + b_{2i}(t)$, with $E(b_{1i}) = \mu_1$, $E(b_{2i}) = \mu_2$, $var(b_{1i}) = \sigma_{11}$, $var(b_{2i} = \sigma_{22}$, and $cov(b_{1i}, b_{2i}) = \sigma_{12}$. The time schedule for the longitudinal measurements of each subject was set at $t_{ij} = (0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80)$ in the simulation. Constant baseline hazard function is used for $\lambda_0(t) = \lambda_0$. Parameters for the longitudinal and survival parts are considered under three settings.

(1) $\beta = 1, \lambda_0 = 1, (\mu_1, \mu_2) = (4.173, -0.0103)$ and the variance components are: $\sigma_e^2 = 0.6$ and $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (4.96, -0.0456, 0.012)$.

(2) $\beta = -1, \lambda_0 = 0.001, (\mu_1, \mu_2) = (2, 0.05)$ and the variance components are $\sigma_e^2 = 0.3$ and $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.5, -0.001, 0.025)$.

(3) Same as (2) above except that $\sigma_e^2 = 0.15, \beta = 1, \sigma_e^2 = 0.15, \lambda_0 = 0.0001, E(\theta) = (2, 0.05)^T$ and
$\{\sigma_{11}, \sigma_{12}, \sigma_{22}\} = \{0.5, -0.001, 0.025\}$.

The first choice of Ω values are same as the one used in Wulfshon and Tsiatis(1997). It is noted that the negativeness of β value renders that the cumulative hazard for survival time T_i is not increasing to ∞. This phenomenon will cause the true model under the simulation setting specified by Ω being very different from Cox's model assumption.

In each setting, the variance of measurement σ_ϵ is change toward very large and very small comparing the original choice. These changes would correspondingly make the conditional variance of the posterior density of θ_i becomes bigger or smaller. Then we examine the differences between the $(I_\beta^w)^{-1/2}$ and the sample standard deviation.

Also the variation σ_{22} of the θ_{i2}, the slope of of $Z_i(t)$, is also changed in both direction to increase or reduce the net information content of β in the simulated data .

In each replication, the data is consisting of 200 samples. Each simulated study has 100 replication. The Censoring time for each subject is generated from Exponential distribution with mean 110 and is independent of all other variables. The EM-algorithm procedure used here is same as that in Wulfshon and Tsiatis(1997) in estimate Ω. For numerical integration for evaluating the conditional expectation in E-step, instead of using Guass-Hermite quadrature formula, we apply Monte Carlo integration as suggested by Henderson et al.(2000) to calculate conditional expectations.

Table 1.1

Simulated results of longitudinal and survival joint model (of case I) via EM-algorithm.

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

Table 1.2

Simulated results of longitudinal and survival joint model (of case II) via EM-algorithm.

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

Table 1.3

Simulation on sparse case

	β
target	-1
mean of $\widehat{\beta}$	-1.3498
SD for $\widehat{\beta}$	0.1647
SE for $\widehat{\beta}$	0.1163

Simulation results are summarized into table 1.1-1.3. From the three tables, it is seen that the differences between the SD (the sample standard deviation from 100 Monte Carlo samples by EM algorithm) and SE (sample average of $(I_\beta^w)^{-1/2}$ with I_β^w obtained as the slope of Newton-Raphson method used in M-step at the converging iteration) are coherent with changes of sizes of measurement error. This fact empirically confirms that the working slope I_β^w should not be used in statistical inferences for β. The 2nd and 3rd columns also reveal that other factors, which have influences on the information content of the data, might as well affect the validity of $(I_\beta^w)^{-1/2}$ as a precision estimate to different degree.

It should re-emphasized here that the phenomenon we have discussed so far applies equally to all other parameters in $(\alpha, \sigma_e, \lambda_0)$.

1.6 Egg-laying Data Analysis

In this section, it is proposed that a valid precision estimate of the EM-estimates could be calculated via bootstrapping method when computation is feasible. That is, the bootstrapping sample variance, denoted as SD_B, is suggested to be used in constructing confidence interval and testing purposes.

Two subsets of egg-laying data were taken from Carey, et al. (1998) to illustrate the bootstrap procedures. The original data consists of the lifetime (in days) and complete reproductive history, in terms of number of eggs laid daily until death, of 1,000 female medflies (Mediterranean fruit fly). The goal was to explore the relation between reproduction and longevity. The proportional hazards assumption failed for flies that are highly productive, so we restrict our attention to the less productive half of the flies. This includes female medflies that produced less than 799 eggs in total during their lifetimes. Moreover, to get a dynamic of the effects of wrongly employing I_β^W to estimate the standard errors of the estimating procedures in section 3, we divided these flies further into two groups. The first group includes the 247 female medflies (Mediterranean fruit fly) which have produced eggs, but the total lifetime reproduction is less than 300 eggs, while the second group includes the 249 female medflies producing more than 300, but less than 799 eggs in lifetime.

For each individual, the life span, L_i, and daily egg reproduction schedule, $m(j)$, for day $j = 1, 2, ..., 99$, were recorded without missing, nor censoring. This affords us the opportunity to focus on the missing information issue in Section 4 without attributing the information loss to other missing data structure such as random censoring of lifetime and irregular sampling of the longitudinal process that are common for medical follow-up data. It also allow us to use standard software to check the proportional hazards assumption since we have the complete egg-laying history.

Since daily egg production is subject to random fluctuation but the underneath reproductive process can be reasonably assumed to be a smooth process, the measurement error model (1.1) is a good prescription to link the underneath reproductive process to longevity. Because we are dealing with count data, it is common to take the log transformation, so we take $W_(t) = log\{m(t) + 1\}$, to avoid days where no eggs were laid. An examination of the individual egg-laying trajectories in Carey et al. (1998) suggests the

following parametric longitudinal process for $X(t)$

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

where the prior distribution of (b_1, b_2) is bivariate Normally distributed with mean (μ_1, μ_2), and $\sigma_{11} = var(b_1), \sigma_{12} = cov(b_1, b_2), \sigma_{22} = var(b_2)$.

The Cox's proportional hazard regression model assumption were checked via martingale residues for both data sets and reasonably satisfied with p-values 0.9 and 0.6, respectively. We thus proceeded with the joint models in (1.1) and (1.2) and Tables 1.4 and 1.5 summarize the EM-algorithm estimates of θ together with the working precision $(I_\beta^W)^{-1/2}$ under the joint model setting. Furthermore, the bootstrapping method with 100 replication is applied on both data sets, and the results of sample mean and standard deviation of the bootstrap samples are also reported.

From Tables 1.4 and 1.5, we can see that the sample bootstrap means are close to the corresponding EM-estimates, suggesting the feasibility of bootstrap approach. Therefore, the bootstrap SD could provide reliable estimates for the standard errors of the EM-estimates. On the other hand, the working SE, corresponding to $(I_\beta^w)^{-1/2}$ in the last step of the EM-algorithm, produced estimates that are noticeably smaller than the bootstrap SD estimate. It yielded an estimate of 0.0801 under the setting of Table 1.4, which is about 15% departure from the bootstrap SD (0.0921). The discrepancy increases to 25% (working SE is 0.0672) for the second data set in Table 1.5 due to a higher measurement error there (1.4344 in Table 1.5 vs. 0.9883 in Table 1.4).

From these results, it seems that the bootstrapping method indeed gives empirically reasonable precision estimate for the EM-estimate in the joint modelling setting. Until further theoretical advances afford us reliable standard error estimates, we recommend to use the bootstrap SD estimate instead of the working SE adopted in the literature to estimate the standard errors of the EM-estimators.

Table 1.4

Analysis of lifetime and log-fecundity of medflies with lifetime reproduction less than 300 eggs. The bootstrap mean and SD are reported in the last two rows.

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
EM estimate	0.5597	0.6950	-0.0542	0.7656	-0.1123	0.0196	0.9883
bootstrap mean	0.5676	0.6932	-0.0541	0.7621	-0.1121	0.0198	0.9858
SD_B	0.0921	0.0736	0.0156	0.1056	0.0207	0.0043	0.0606

Table 1.5

Analysis of lifetime and log-fecundity of medflies with lifetime reproduction between 300 and 799 eggs. The bootstrap mean and SD are reported in the last two rows.

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
EM estimate	-0.1997	1.7618	-0.1730	1.0540	-0.1734	0.0301	1.4344
bootstrap mean	-0.2036	1.7659	-0.1739	0.7621	-0.1766	0.0309	1.4333
SD_B	0.0836	0.0800	0.0145	0.0784	0.0151	0.0033	0.0356

1.7 Discussion and Conclusion

We have accomplished three goals in this paper:

1. Reinforce the merit of the joint modelling approach in **WT** by providing a theoretical explanation of the robust feature observed in the literature. This answers a questions raised in Tsiatis Davidian (2004) and suggests that the likelihood based procedure with normal random effects in **WT** can be very efficient even if this normality assumption has been violated as long as there is rich enough information available on the longitudinal data. Generally, this means that the longitudinal data should not be too sparse or with too large measurement errors.

2. Demonstrate the missing information in joint modelling, both theoretically and empirically, to alert practitioners. The efficiency loss can be quite substantial as observed in the simulations and data analysis. We recommend to use bootstrap procedure to estimate the standard errors of the estimates resulted from the EM-algorithm, and illustrated this in the fecundity data in Section 6.

3. Connect the approach in **WT** to nonparametric maximum likelihood approach. This enhances the speculation in Tsiatis and Davidian (2004) that they might indeed be efficient as traditional MLEs, although a rigorous proof is not yet available.

However, theoretical gaps remain to validate the asymptotic properties of the estimates in **WT**. It is known that the theory of profile likelihood approach is well developed for parameteric setting (Patefield (1977)), but not immediately applicable in semiparametric setting (Murphy and Van der Vaart (2000) and Fan and Wong (2000)). The complexity caused by profiling on nonparametric parameter with only implicit structure creates additional difficulties in theoretical and computational developments in the joint modelling. Much further efforts are required to resolve these issues and to provide reliable precision estimates for statistical inferences under the joint modelling setting.

Chapter 2

Accelerated Failure Time Models For The Event Times

2.1 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

$$W(t) = X(t) + e(t) \tag{2.1}$$

, 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:

$$\lambda\{\ t|\bar{X}(t)\} = \lambda_0(t)\ exp\{\beta X(t)\}, \tag{2.2}$$

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 (2.1) and the entire history of $X(t)$ were available, one could use Cox's partial likelihood to estimate the regression parameter β in (2.2). 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 encounter 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 (2.1) and (2.2) 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 (2.2) 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:

$$log\ T = -\beta' X + e \tag{2.3}$$

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:

$$U \sim S_0, \qquad where \quad U = \psi\{\bar{X}(T); \beta\} = \int_0^T exp\{\beta X(s)\} ds. \tag{2.4}$$

With this transformation, the survival function for an individual with covariate history $\bar{X}(t)$, is $S\{t|\bar{X}(t)\} = S_0[\psi\{\bar{X}(t); \beta\}]$. This means that individuals age on an accelerated schedule , $\psi\{\bar{X}(t); \beta\}$, under a baseline survival function $S_0(\cdot)$. 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

$$\lambda\{t|\bar{X}(t)\} = \lambda_0[\int_0^t exp\{\beta X(s)\} ds] exp\{\beta X(t)\} = \lambda_0[\psi\{\bar{X}(t); \beta\}] \psi'\{\bar{X}(t); \beta\}, \tag{2.5}$$

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, (2.5) 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 (2.5) 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 chapter 2, we consider the joint AFT model as specified by (1) and (2.5) (or equivalently, (1) and (2.4)), 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 (2.4) (or (2.5)); 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 (2.5), 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 the second part of the thesis.

2.2 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,

$$X_i(t) = \mathbf{b}_i^T \boldsymbol{\rho}(t), \tag{2.6}$$

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 (2.6) 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 (2.6). 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 (2.6), the hazard function in (2.5) now takes the form:

$$\lambda(t|\bar{X}(t)) = \lambda(t|\beta, \mathbf{b}_i) = \lambda_0\{\psi(t; \beta, \mathbf{b}_i)\}\psi^{'}(t; \beta, \mathbf{b}_i), \tag{2.7}$$

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 (2.4) and (2.5) 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 (2.1) and (2.7) can be expressed as:

$$\begin{aligned} L(\theta) &= L(\beta, \boldsymbol{\mu}, \Sigma, \sigma_e^2, \lambda_0) \\ &= \Pi_{i=1}^{n}[\int \{\Pi_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \mathbf{t}_i, \sigma_e^2)\} f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) f(\mathbf{b}_i|\Sigma, \boldsymbol{\mu}) d\mathbf{b}_i], \end{aligned} \tag{2.8}$$

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

$$f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) = [\lambda_0\{\psi(V_i; \beta, \mathbf{b}_i)\}\psi'(V_i; \beta, \mathbf{b}_i)]^{\Delta_i} exp\{-\int_0^{\psi(V_i;\beta,\mathbf{b}_i)} \lambda_0(t)dt\}. \tag{2.9}$$

Difficulties in Baseline Estimation: The expression in (2.9), 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 (2.5) or (2.9) excludes discrete survival times and hence the point mass approach for baseline hazards. Moreover, direct maximum likelihood estimate for baseline hazard function fails for (2.9), 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 (2.8) and (2.9). 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.

2.3 EM Algorithm

The joint likelihood in (2.8) 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

$$L^*(\theta) = \Pi_{i=1}^{n}[\{\Pi_{j=1}^{m_i} f(W_{ij}|\mathbf{b}_i, \mathbf{t}_i, \sigma_e^2)\} f(V_i, \Delta_i|\mathbf{b}_i, \mathbf{t}_i, \lambda_0, \beta) f(\mathbf{b}_i|\Sigma, \boldsymbol{\mu})]. \tag{2.10}$$

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.

2.3.1 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:

$$\hat{\mu} = \sum_{i=1}^{n} E_i(\mathbf{b}_i)/n, \tag{2.11}$$

$$\hat{\Sigma} = \sum_{i=1}^{n} E_i(\mathbf{b}_i - \hat{\mu})(\mathbf{b}_i - \hat{\mu})^T/n, \tag{2.12}$$

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i} E_i\{W_{ij} - \mathbf{b}_i^T\ \boldsymbol{\rho}(t_{ij})\}^2 / \sum_{i=1}^{n} m_i. \tag{2.13}$$

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 (2.4). 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 (2.4). 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 (2.4), 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 :

$$\lambda_0(u) = \sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} \leq u < \hat{u}_{(j)}\}}. \tag{2.14}$$

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

$$\int_0^{\psi(V_i;\beta,\mathbf{b}_i)} \lambda_0(s)ds = \int_0^{u_i} \lambda_0(s)ds = \sum_{j=1}^{d} C_j\{\hat{u}_{(j)} - \hat{u}_{(j-1)}\}1_{\{\hat{u}_{(j)} \leq u_i\}}. \tag{2.15}$$

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

$$\begin{aligned}
&\frac{\partial}{\partial C_k} E_i\{logL^*(\theta)\} \\
&= \frac{\partial}{\partial C_k} \sum_{i=1}^{n} E_i[\Delta_i log\ \lambda_0(u_i) - \Lambda_0\{\psi(V_i;\beta,\mathbf{b}_i)\}] \\
&= 0. \qquad (2.16)
\end{aligned}$$

Substituting $\lambda_0(u_i)$ and $\int_0^{u_i} \lambda_0(t)dt$ in (2.16) by (2.14) and (2.15) respectively, (2.16) 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

$$\hat{C}_k = \frac{\sum_{i=1}^{n} E_i[\Delta_i 1_{\{\hat{u}_{(k-1)} \leq u_i < \hat{u}_{(k)}\}}]}{\sum_{i=1}^{n} E_i[\{\hat{u}_{(k)} - \hat{u}_{(k-1)}\}1_{\{\hat{u}_{(k)} \leq u_i\}}]}. \qquad (2.17)$$

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

$$\begin{aligned}
&\sum_{i=1}^{n} E_i\left(\Delta_i log\left[\sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} < u_i \leq \hat{u}_{(j)}\}}\right] + \Delta_i\beta\{\mathbf{b}_i^T\boldsymbol{\rho}(V_i)\} - \sum_{j=1}^{d} C_j\{\hat{u}_{(j)} - \hat{u}_{(j-1)}\}1_{\{\hat{u}_{(j)} \leq u_i\}}\right) \\
&+ \sum_{i=1}^{n} E_i\{logf(\mathbf{b}_i|\Sigma,\boldsymbol{\mu})\} + \sum_{i=1}^{n} E_i\{\sum_{j=1}^{m_i} log\ f(W_{ij}|\mathbf{b}_i,\sigma_e^2)\}. \qquad (2.18)
\end{aligned}$$

There is no closed form expression for the maximum likelihood estimate $\hat{\beta}$ in (2.18) 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 (2.18). 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.

2.3.2 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 (2.10) 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

$$\mathbf{b}_i|\boldsymbol{W_i}, \boldsymbol{t_i}, \widehat{\theta} \sim N\{\boldsymbol{\mu} + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{W_i} - A\boldsymbol{\mu}),\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\}. \tag{2.19}$$

The empirical Bayes estimate or BLUP for $\mathbf{b}_i$ is thus the estimated mean of (2.19). 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.

2.3.3 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 (2.11), (2.12), (2.13) and (2.17) 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 (2.6). 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.

2.3.4 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_o^* from original observed data w_o.
2. The EM algorithm is applied to the bootstrap sample w_o^* 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.

2.4 Simulation Study

We study the performance of the EM-procedures in section 10 through simulations with n=100 subjects and 100 simulated samples. In the survival model (2.5), 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 (2.6) 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. For this reason, the negative values

are discarded and the resulting $\mathbf{b}_i$ is thus actually generated from a truncated bivariate normal distribution with 35% of the bivariate vectors truncated.

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 2.1

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 2.2

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 2.3

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 2.1 and Table 2.2 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 2.3, 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 2.3, 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 2.1 and 2.2, this is probably due to the increase in the target variance components rather than the stability of the procedures. Comparing Table 2.3 to Table 2.2, 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 2.1 , 2.2 and 2.3 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 2.3), 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.

2.5 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 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 (2.5), 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.

2.5.1 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 2.1. 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 2.1.

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 val-

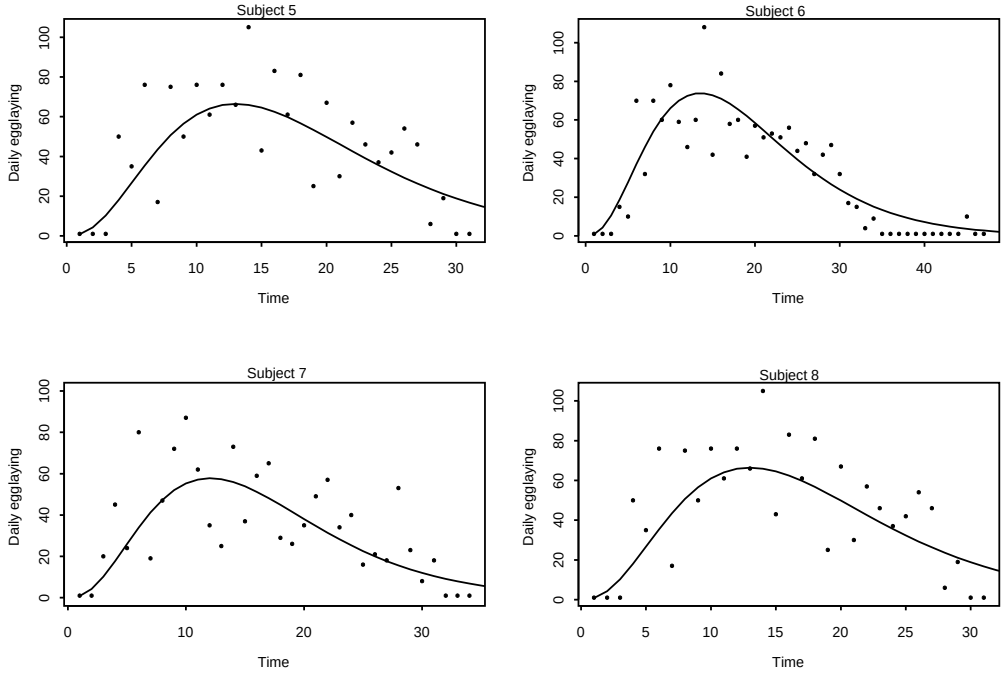


Figure 2.1: *Individual profiles are fitted by the gamma function. Daily egg-laying of subject 5 is fitted by* $t^{2.710}e^{-0.204}$*, subject 6 by* $t^{2.652}e^{-0.193}$*, subject7 by* $t^{2.725}e^{-0.226}$*, and subject 8 by* $t^{2.803}e^{-0.221}$.

ues, since daily egg-laying of each individual could be zero. Consequentially, the final longitudinal model for the ith individual becomes:

$$log(W_{ij} + 1) = X_{ij} + e_{ij}, \tag{2.20}$$

$$X_{ij} = b_{1i}log(t_{ij}) + b_{2i}(t_{ij} - 1), \tag{2.21}$$

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 2.4, where the standard error estimate for each parameter is derived by 100 bootstrap samples as described in Section 3·5.

Table 2.4

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.2 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 (2.19). The four fitted curves (dashed lines) capture the egg-laying trajectories quite well. Fig. 3 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

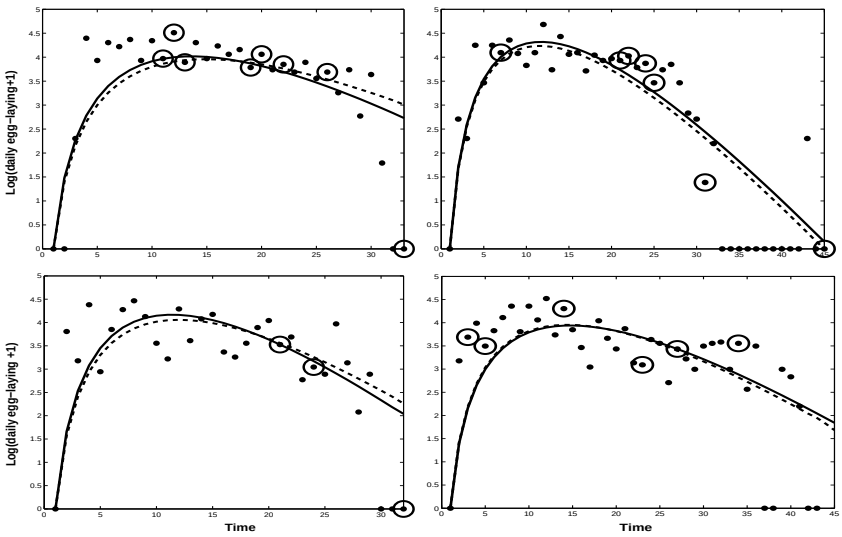


Figure 2.2: *Fitted fecundity curves for four medflies based on complete (dots) and incomplete (circled dots) data. The dashed lines are the fitted curves based on complete data, and the solid lines for incomplete data.*

(2.20) and (2.21) for female medfly fecundity and survival data.

2.5.2 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 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

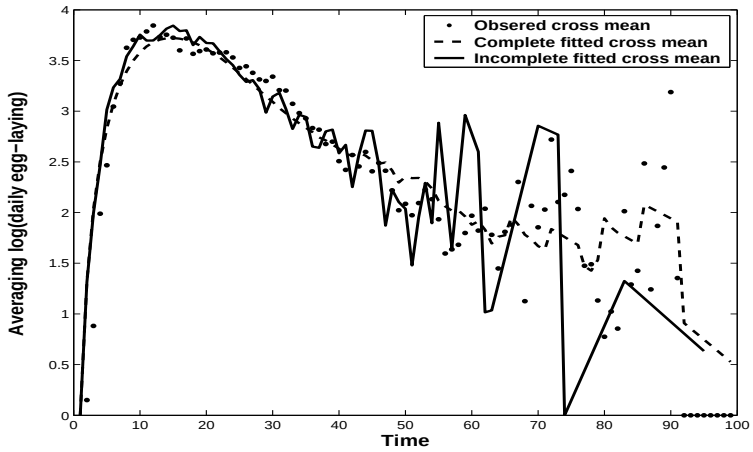


Figure 2.3: *Fitted Cross-sectional mean curves for complete and incomplete data. The dots represents the daily mean eggs of those that are still alive.*

% 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 2.5.

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 2.5 are all much larger than those in Table 2.4 because a large proportion of information is lost due to the unavailability of the measurements.

Table 2.5

The parameter estimates derived from incomplete 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	-0.3890	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

The fitted individual curves for the four subjects based on the incomplete data are also shown in Figure 2.2 (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 2.3. 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.

2.6 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 (2.20) and (2.21), we fit the longitudinal covariates for the medfly data by a simple linear mixed model which is (2.6) 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 (2.4) 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.

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.

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

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.

Johansen, S. (1983). An extension of Cox's regression Model. *International Statistics Review* **51**, 258-262.

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

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

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

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.

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

Murphy, S. & van der Vaart, A. W. (2000). On profile likelihood (with Discussion). *Journal of American Statistical Association* **95**,449-465.

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.

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

Pawitan, Y. & Self, S. (1993). Modeling disease marker process in AIDS.*Journal of American Statistical Association* **88**, 719-726.

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, New York.

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

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**, 809-34.

Tsiatis, A. A., Degruttola, V. & 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. & 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**, 835-62.