Joint Modelling of Accelerated Failure Time and Longitudinal Data

Yi-Kuan Tseng, Fushing Hsieh, and Jane-Ling Wang

Department of Statistics, University of California
Davis, CA 95616, U.S.A.
email: wang@wald.ucdavis.edu

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

KEY WORDS: EM algorithm; Measurement errors; Missing data; Monte Carlo integration; Random effects; Survival data.

1 Introduction

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

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

at discrete time points. For simplicity we assume that there is only one longitudinal covariate; as the case of multiple longitudinal covariates and additional time-independent covariates can easily be adapted.

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

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

where $\bar{X}(t) = \{X(s) : 0 \leq s < t\}$ is the covariate history up to time t, β is the regression parameter, and $\lambda_0(t)$ is the unspecified baseline hazard rate function.

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

However, the proportionality assumption may fail and we assume, as an alternative, the accelerated failure time model as described in Cox and Oakes (1984, Ch. 5, pp. 64-5):

$$U \sim S_0, \qquad \text{where} \quad U = \psi\{X(T);\beta\} = \int_0^T \exp\{\beta X(s)\}ds. \tag{3}$$

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

S_0, the hazard rate function for an individual with covariate history $\bar{X}(t)$ can thus be expressed as

$$\lambda\{t|\bar{X}(t)\} = \lambda_0[\int_0^t \exp\{\beta X(s)\}ds]\exp\{\beta X(t)\} = \lambda_0[\psi\{X(t);\beta\}]\dot{\psi}\{X(t);\beta\}, \tag{4}$$

where $\lambda_0(\cdot)$ is the hazard function for S_0 and $\dot{\psi}$ is the first derivative of ψ. Here, U plays the role of a baseline failure-time variable and we thus refer to $\lambda_0(\cdot)$ as the baseline hazard function, which is usually left unspecified. Thus, (4) corresponds to a semiparametric model, first studied by Robins and Tsiatis (1992) using a certain class of rank estimating equations for β. These rank estimators were shown to be consistent and asymptotically normal by Lin and Ying (1995). Recently, Hsieh (2003, manuscript) proposed an over-identified estimating equation approach for achieving semiparametric efficiency and to extend (4) to a heteroscedastic version. All this aforementioned work assumes, however, that the entire covariate process, $X(t)$, can be observed without measurement errors.

2 The joint model

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

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

$$X_i(t) = b_i'\, \rho(t), \tag{5}$$

where $\rho(t) = \{\rho_1(t), \cdots, \rho_p(t)\}'$ and $\rho(t)$ are known functions; b_i'= $(b_{1i}, \cdots, b_{pi})$ are p-dimensional multivariate normal, $N_p(\mu, \Sigma)$, independent of the measurement errors e_i. The measurement errors, e_i, are also assumed to be multivariate normal, with independent and identically distributed components $e_{ij} \sim N(0, \sigma_e^2)$. The random effect vectors b_i, which are not observed and are treated as missing data in the likelihood approach to follow, are estimated by the EM algorithm. If $p = 2$ and $\{\rho_1(t), \rho_2(t)\} = (1, t)$, then (5) 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-curve models at high computational cost, as the EM steps involve evaluation of p-dimensional integrals.

Under the accelerated failure time assumption and the parametric longitudinal model (5), the hazard function

in (4) now takes the form

$$\lambda\{t|\bar{X}(t)\} = \lambda(t|\beta, b_i) = \lambda_0\{\psi(t;\beta,b_i)\}\dot{\psi}(t;\beta,b_i), \tag{6}$$

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

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

corresponds to the transformation in (3) and (4) with derivative

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

To construct the likelihood function, we assume noninformative censoring and measurement schedule t_{ij}, which is also independent of the future covariate history and random effects b_i. Under these assumptions, the probability mechanisms of both censoring and the measurement schedule can be factorized out of the likelihood function, and the joint observed likelihood for the model made up of (1) and (6) can be expressed as

$$\begin{aligned} L(\theta) &= L(\beta,\mu,\Sigma,\sigma_e^2,\lambda_0) \\ &= \Pi_{i=1}^n[\int\{\Pi_{j=1}^{m_i} f(W_{ij}|b_i,t_i,\sigma_e^2)\}f(V_i,\Delta_i|b_i,t_i,\lambda_0,\beta)f(b_i|\Sigma,\mu)db_i], \end{aligned} \tag{7}$$

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

$$f(V_i,\Delta_i|b_i,t_i,\lambda_0,\beta) = [\lambda_0\{\psi(V_i;\beta,b_i)\}\dot{\psi}(V_i;\beta,b_i)]^{\Delta_i}\exp\{-\int_0^{\psi(V_i;\beta,b_i)}\lambda_0(t)dt\}. \tag{8}$$

3 EM Algorithm

The joint likelihood in (7) will be maximized via the EM algorithm. The complete data for the ith subject are $(V_i,\Delta_i,W_i,t_i,b_i)$ and the complete-data likelihood is

$$L^*(\theta) = \Pi_{i=1}^n[\{\Pi_{j=1}^{m_i} f(W_{ij}|b_i,t_i,\sigma_e^2)\}f(V_i,\Delta_i|b_i,t_i,\lambda_0,\beta)f(b_i|\Sigma,\mu)]. \tag{9}$$

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

3.1 M-step

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

$$\hat{\mu} = \sum_{i=1}^n E_i(b_i)/n, \tag{10}$$

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

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i} E_i\{W_{ij} - b_i'\ \rho(t_{ij})\}^2 / \sum_{i=1}^{n} m_i. \tag{12}$$

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 (3). Ideally, we could approximate λ_0 by step functions, which leads to a natural parameterization of the baseline hazard function. Since we cannot observe the baseline failure times, we estimate them through (3). 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 (3), for these d subjects are $u_k = \int_0^{T_k} \exp\{\beta\ b_k'\rho(s)\}ds, k = 1, ..., d$. They can then be estimated by plugging in the current estimate of β and the current empirical Bayes estimate of 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 parameterization 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

$$\lambda_0(u) = \sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} \leq u < \hat{u}_{(j)}\}}. \tag{13}$$

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

$$\int_0^{\psi(V_i;\beta,b_i)} \lambda_0(s)ds = \int_0^{u_i} \lambda_0(s)ds = \sum_{j=1}^{d} C_j(\hat{u}_{(j)} - \hat{u}_{(j-1)})1_{\{\hat{u}_{(j)} \leq u_i\}}. \tag{14}$$

Therefore, the maximum likelihood estimate for C_k is

$$\hat{C}_k = \frac{\sum_{i=1}^{n} E_i(\Delta_i 1_{\{\hat{u}_{(k-1)} \leq u_i < \hat{u}_{(k)}\}})}{\sum_{i=1}^{n} E_i\{(\hat{u}_{(k)} - \hat{u}_{(k-1)})1_{\{\hat{u}_{(k)} \leq u_i\}}\}}. \tag{15}$$

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

$$\sum_{i=1}^{n} E_i \left\{ \Delta_i \log\left(\sum_{j=1}^{d} C_j 1_{\{\hat{u}_{(j-1)} < u_i \leq \hat{u}_{(j)}\}} \right) + \Delta_i \beta\{b_i'\ \rho(V_i)\} - \sum_{j=1}^{d} C_j(\hat{u}_{(j)} - \hat{u}_{(j-1)})1_{\{\hat{u}_{(j)} \leq u_i\}} \right\}$$
$$+ \sum_{i=1}^{n} E_i\{\log f(b_i|\Sigma, \mu)\} + \sum_{i=1}^{n} E_i\{\sum_{j=1}^{m_i} \log f(W_{ij}|b_i, \sigma_e^2)\}. \tag{16}$$

There is no closed-form expression for the maximum likelihood estimate $\hat{\beta}$ in (16) since the u_i's involve β. Furthermore, it is not easy to derive the score for β because of the complexity of u_i's and the indicator functions that are involved in β in (16). Therefore, instead of using the Newton-Raphson method to obtain the slope for $\hat{\beta}$, one can estimate β by directly maximizing the likelihood when β is low-dimensional.

3.2 E-step

The M-step above involved E_i, which requires knowledge of $f(b_i|V_i, \Delta_i, W_i, t_i, \hat{\theta})$. This can be obtained through the Bayes rule, and b_i is estimated by BLUP or equivalently, the empirical Bayes estimate. To be more specific, let $\rho^*= \{\rho'(t_{i1})\mu, \cdots, \rho'(t_{im_i})\mu\}'$ and $A = \{\rho(t_{i1}), \cdots, \rho(t_{im_i})\}'$. Given t_i, we have

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

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

$$b_i|W_i, t_i, \widehat{\theta} \sim N\{\mu + \Sigma_{21}\Sigma_{11}^{-1}(W_i - A\mu),\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\}. \tag{17}$$

The empirical Bayes estimate for b_i is thus the estimated mean of (17). Moreover, Monte Carlo integration is used to derive all $E_i(\cdot)$, similarly to Henderson et al. (2000), by generating a number, M, of multivariate normal sequences for $b_i|W_i, t_i, \widehat{\theta}$, denoted by $N_i = (N_{i1}, \cdots N_{iM})'$. Then, for any function, $h(\cdot)$ of b_i, we have

$$\begin{aligned} E_i\{h(b_i)\} &= \frac{\int_{-\infty}^{\infty} h(b_i) f(V_i, \Delta_i|b_i, t_i, \widehat{\theta}) f(b_i|W_i, t_i, \widehat{\theta}) db_i}{\int_{-\infty}^{\infty} f(V_i, \Delta_i|b_i, t_i, \widehat{\theta}) f(b_i|W_i, t_i, \widehat{\theta}) db_i} \\ &\simeq \frac{\sum_{j=1}^{M} h(N_{ij}) f(V_i, \Delta_i|N_{ij}, t_i, \widehat{\theta})}{\sum_{j=1}^{M} f(V_i, \Delta_i|N_{ij}, t_i, \widehat{\theta})}, \end{aligned}$$

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

When estimating the standard error of $\widehat{\beta}$, we encounter two difficulties. The first is that the exact information matrix of parameters of interest cannot be obtained directly in the EM algorithm. Remedies proposed in Louis (1982) and McLachlan and Krishnan (1997, Ch. 4) approximate the observed Fisher information matrix. These approximations are asymptotically valid for a finite-dimensional parameter space, but we consider the baseline hazard to be unspecified, and the asymptotic validity of such approximations is dubious for an infinite-dimensional parameter space. The second difficulty is that a promising way of deriving the information matrix is provided by profile likelihood. However, the mixture structure of the joint accelerated failure time model does not allow an explicit profile likelihood. Hence we need to project on to all other parameters, including the infinite-dimensional parameter, λ_0, in order to derive estimated standard errors for $\hat{\beta}$. It is very difficult to derive this projection, which involves the infinite-dimensional parameter λ_0.

In view of the above difficulties, we suggest the use of Efron's (1994) bootstrap technique for missing data to derive the standard error estimates. Tseng et al. (2005) provide the detail of implementation of the bootstrap technique in joint modelling. We will apply their procedure on medfly data to derive standard error estimates.

4 Simulation studies

We study the performance of the EM-procedures in § 3 through simulations with n=100 subjects and 100 simulated samples. In the survival model (4), 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 (5) with $\rho_1(t) = 1$ and $\rho_2 = t$, normal random effects with mean $\mu = (1, 0.5)'$, 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 is available after death or censoring time. Three different settings are considered for the variance components, Σ and censoring schemes: (i) $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.01, -0.001, 0.001)$, and no censoring on scheduled measure times; (ii) the σ_{ij} take the same values as (i), but the lifetime is subject to censoring according to the exponential distribution with mean 25, which resulted in about 20% censoring among all subjects; (iii) the same setting as (ii) except that $\sigma_{22} = 0.3$. As a result of the larger variation, b_{2i} may become negative in (iii), leading to improper survival distributions with positive point mass at ∞. While this causes no problem as the data would be censored at the censoring time in such a case, they are unnatural in that this assumes infinite survival time, as in the cure model setting. We choose to discard the negative values and the resulting b_i is thus actually generated from a truncated bivariate normal distribution with 35% of the bivariate vectors truncated. This deviation from the normality assumption allows us to check the robustness of our procedure, which assumes a normal random effect.

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

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

The robustness property exemplified with setting (iii) was also observed in Song et al. (2002) and Tsiatis

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

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

5 Application to Medfly fecundity

We apply our procedures to the egg-laying data in Carey et al. (1998), which motivated our model. The original dataset consist of 1000 female Mediterranean fruit flies (medflies), for which the numbers of eggs produced daily until death were recorded without missing values. The goal there was to explore the relationship of the pattern of these fecundity curves, $X(t)$, to longevity, as measured by the associated lifetimes of the medflies. Such information is important because reproduction is considered by evolutionary biologists to be the single most important life history trait besides lifetime itself. This dataset is unusual and is selected for illustration for several reasons.

First, the proportional hazards assumption fails for the most fertile medflies; we use data from the 251 flies that produced more than 1150 eggs in their lifetime. The proportional hazards assumption was rejected by the test based on Schonfeld residuals in S-Plus, as described later. This is not surprising because of the complexity of the reproductive dynamics and its association with lifetime. On the other hand, an accelerated failure time model, as defined in (4), 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 dataset contains the complete event history, the reproductive history in this case, for all experimental subjects, which is rare for data collected in medical longitudinal studies. The complete data setting allow us to discard most of the original data artificially and to apply our procedure to both the complete and incomplete datasets. This allows us to check the stability of the joint accelerated failure time procedure.

5.1 Fitting the model to the complete medfly data

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

'random' shape and scale parameters for the ith fly:

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

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

$$\log(W_{ij}+1) = X_{ij} + e_{ij}, \tag{18}$$

$$X_{ij} = b_{1i}\log(t_{ij}) + b_{2i}(t_{ij}-1), \tag{19}$$

where $e_{ij} \sim N(0, \sigma_e^2)$ and $b_i = (b_{1i}, b_{2i})' \sim N(\mu_{2\times1}, \Sigma_{2\times2})$, for $i = 1, ..., 251$, $j = 1, ..., m_i$ and $22 \le m_i \le 99$. 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 hazards 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 accelerated failure time survival model is thus proposed, based on its aforementioned biological appealing feature. The results of the joint accelerated failure time procedure developed in § 3 are summarized in Table 2 (a), where the standard error estimate for each parameter is derived from 100 bootstrap samples as described in Tseng et al. (2005). 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 of the reliability of the bootstrap procedure under the joint modelling framework. Based on the bootstrap standard deviations, all the parameters are highly significant, and the 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 of genetic fitness for those flies.

Fig. 3 provides the empirical Bayes's estimates of the four individual $X(t)$, with b_i estimated from the mean of the bivariate normal distribution in (17). The four fitted curves , denoted by dashed lines, capture the egg-laying trajectories quite well. Fig. 4 shows the cross-sectional sample mean of the log daily egg-laying and the mean of the 251 fitted curves. The fitted mean curve, denoted by dashed, is very close to the sample means up until day 60, at which time only 10% of the medflies are still alive. The variation becomes larger afterwards, as expected. We have thus demonstrated the feasibility of the joint models (18) and (19) for female medfly fecundity and survival data.

5.2 Fitting incomplete medfly data

We now test our procedure in the presence of censoring and irregular sampling plans. We randomly select 1 to 7 days as the corresponding schedule times for each individual and then add the day of death as the last schedule time. Therefore, a minimum of 2 and a maximum of 8 repeated measurements 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 data are further censored by an exponential distribution with mean 500, which resulted in censoring of lifetimes for 20 % of the medflies and many fewer longitudinal measurements for the censored subjects. The joint accelerated failure time procedure is then applied to this incomplete dataset, and the results are presented in Table 2 (b).

Here again, the bootstrap procedures seem to be effective, all parameters are highly significant, and the point estimates based on the incomplete data are close to those based on the complete data.

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

6 Discussion

We have demonstrated that our procedure can be insensitive to the normality assumption, but this must not be mistaken for 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 basis functions, ρ_k. A misspecified functional form of the longitudinal covariates could induce large bias. For example, if instead of (18) and (19) we fit the longitudinal covariates for the medfly data by a simple linear mixed model given by (5) with $\rho(t) = (1, t)'$ and $b_i = (b_{1i}, b_{2i})'$, the estimate of β becomes -0.021 with standard deviation 0.14, which results in nonsignificance of the fecundity curve for the medfly data.

It is straightforward to extend our procedure to accommodate multivariate time dependent covariates and/or baseline covariates. Instead of (3) we have

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Song, X., Davidian, M. and 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. andGrambsch, P. M. (2000). *Modeling Survival Data* . New York: Spinger.

Tseng, Y. K., Hsieh, F. and Wang, J. L. (2005) Joint modelling of Accelerated failure time and longitudinal data. *Biometrika*, In press.

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

Tsiatis,A. A. and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809-34.

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

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

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

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

Yu, M., Law, N.J., Taylor, J. M. G. and Sandler H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, **14**, 835-62.

Table 1. *Simulation study. Results for three settings.*

(a) *No censoring, normal random effects*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0.5	0.01	-0.001	0.001	0.25
Mean	1.0075	0.9955	0.5013	0.0087	-0.0011	0.0009	0.2528
SD	0.0945	0.0163	0.0055	0.0015	0.0002	0.0002	0.0135

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

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0.5	0.01	-0.001	0.001	0.25
Mean	0.9918	0.9944	0.5015	0.0083	-0.0011	0.0009	0.2516
SD	0.1272	0.0249	0.0056	0.0023	0.0004	0.0002	0.0198

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

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Target	1	1	0.5	0.01	-0.001	0.3	0.25
Empirical target	1	0.9993	0.6758	0.0104	-0.0058	0.1358	0.2753
Mean	0.9950	1.0007	0.6682	0.0099	-0.0006	0.1627	0.2500
SD	0.1091	0.0140	0.0535	0.0004	0.0036	0.0318	0.0223

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

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

(a) *Complete data*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Fitted value	-0.4340	2.1227	-0.1442	0.3701	-0.0482	0.0068	0.8944
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

(b) *Incomplete data*

	β	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Fitted value	-0.3890	2.2011	-0.1665	0.2833	-0.0382	0.0051	0.9775
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

SD, standard deviation of 100 bootstrap estimates.

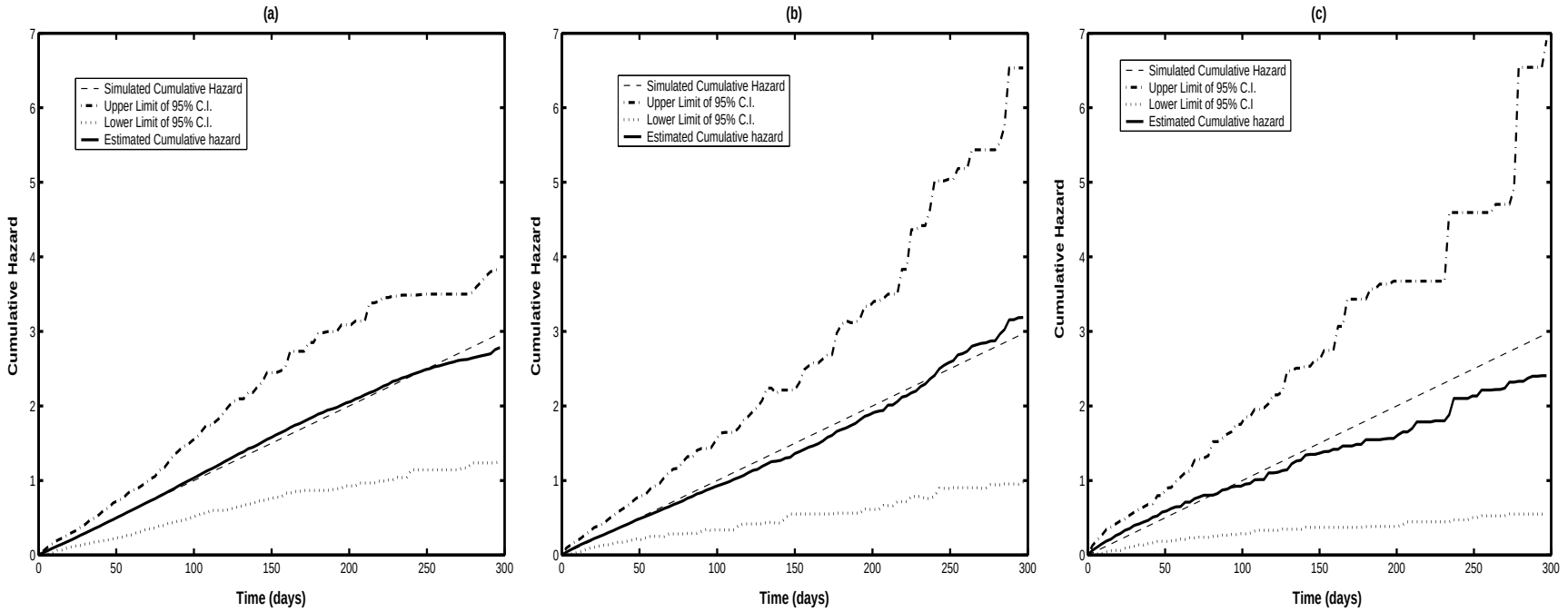

Fig. 1. Simulation study. Estimated cumulative baseline hazard function. (a)-(c) correspond to settings (i)-(iii) respectively. Each simulated cumulative baseline hazard function, dashed, the estimate of the function, solid, and bands made up of the upper, dot-dashed, lower, dotted, limits of the 95% confidence intervals.

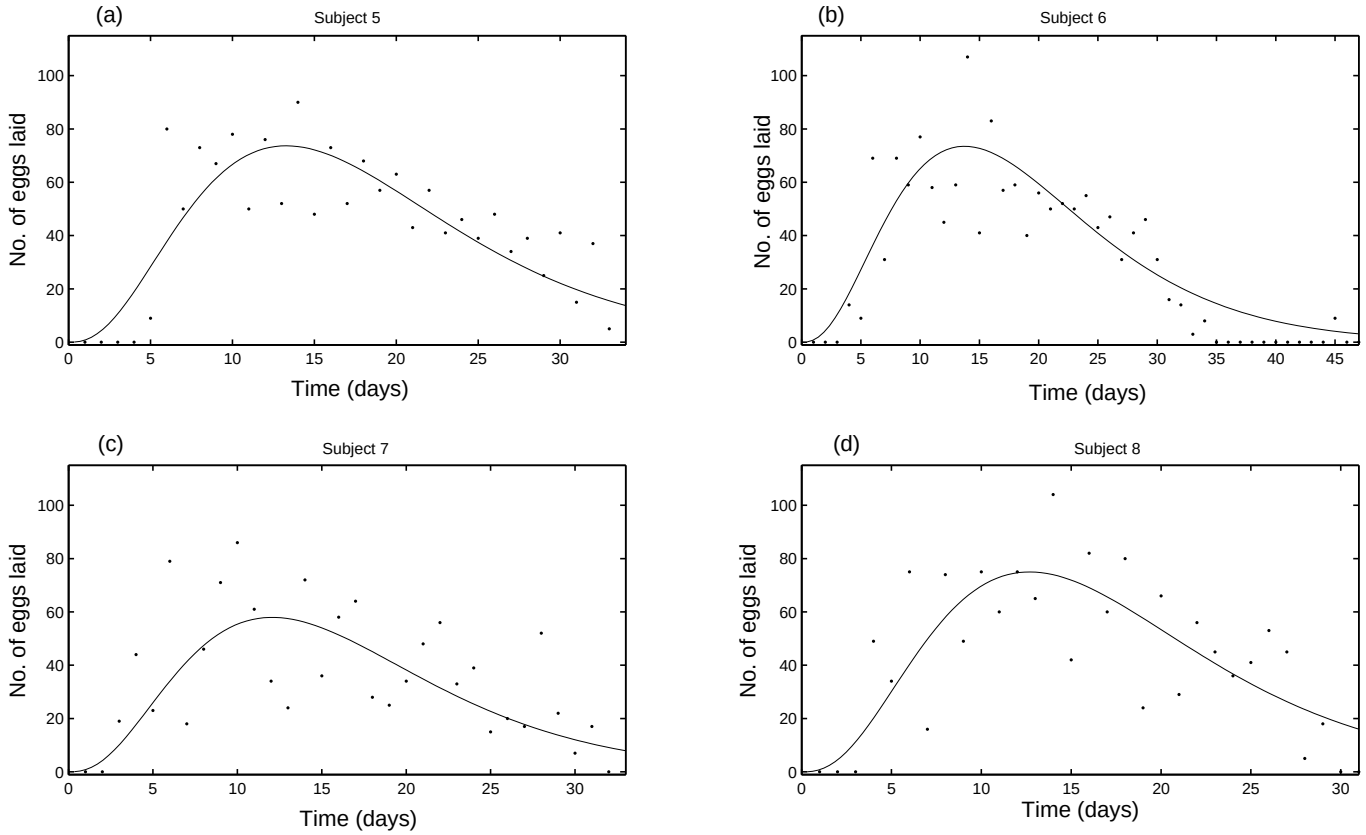

Fig. 2. Individual profiles are fitted by the gamma function. Number of eggs laid of subject 5 is fitted by $t^{2.710}e^{-0.204t}$, subject 6 by $t^{2.652}e^{-0.193t}$, subject 7 by $t^{2.725}e^{-0.226t}$, and subject 8 by $t^{2.803}e^{-0.221t}$. The obtained by least squares.

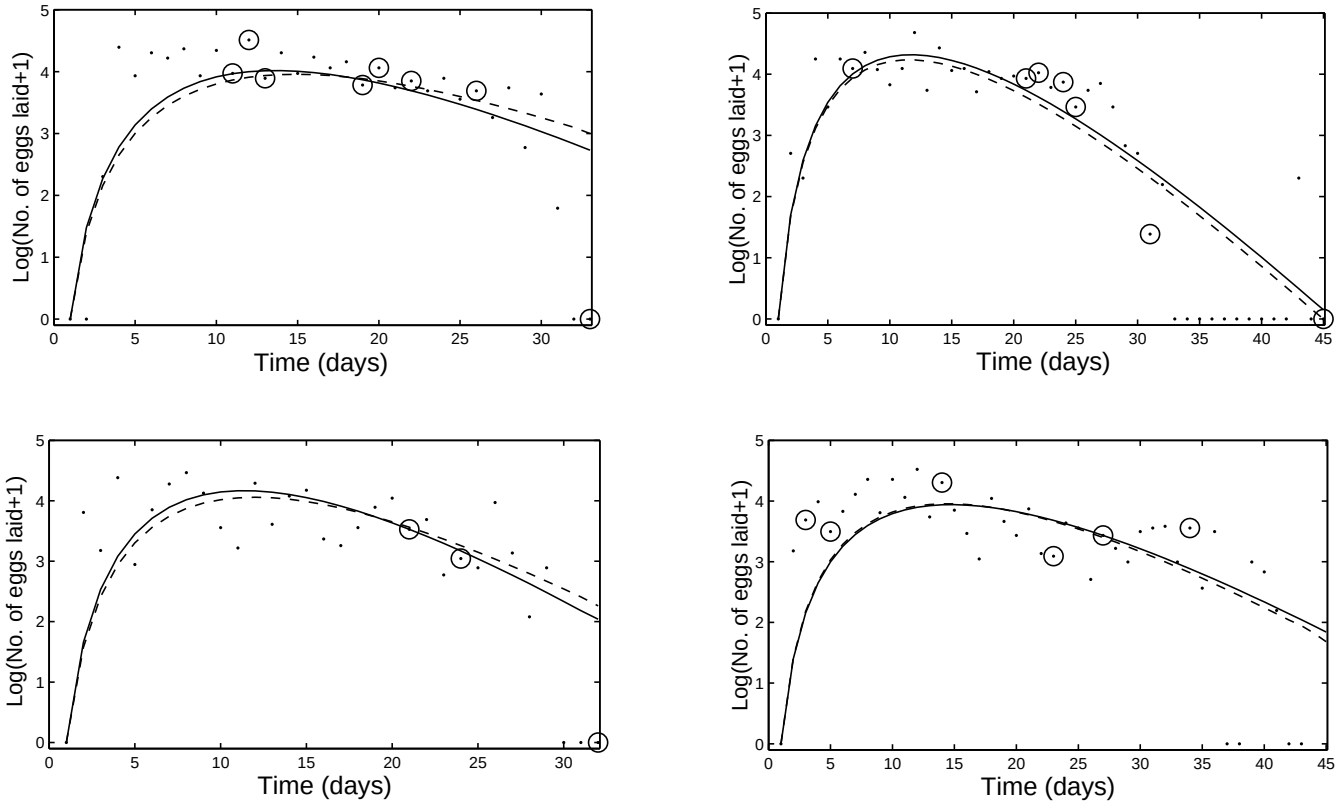

Fig. 3. Fitted fecundity curves for four medflies based on complete (dots) and incomplete (circled dots) data. The dashed lines are the fitted curves based on complete data, and the solid lines for incomplete data.

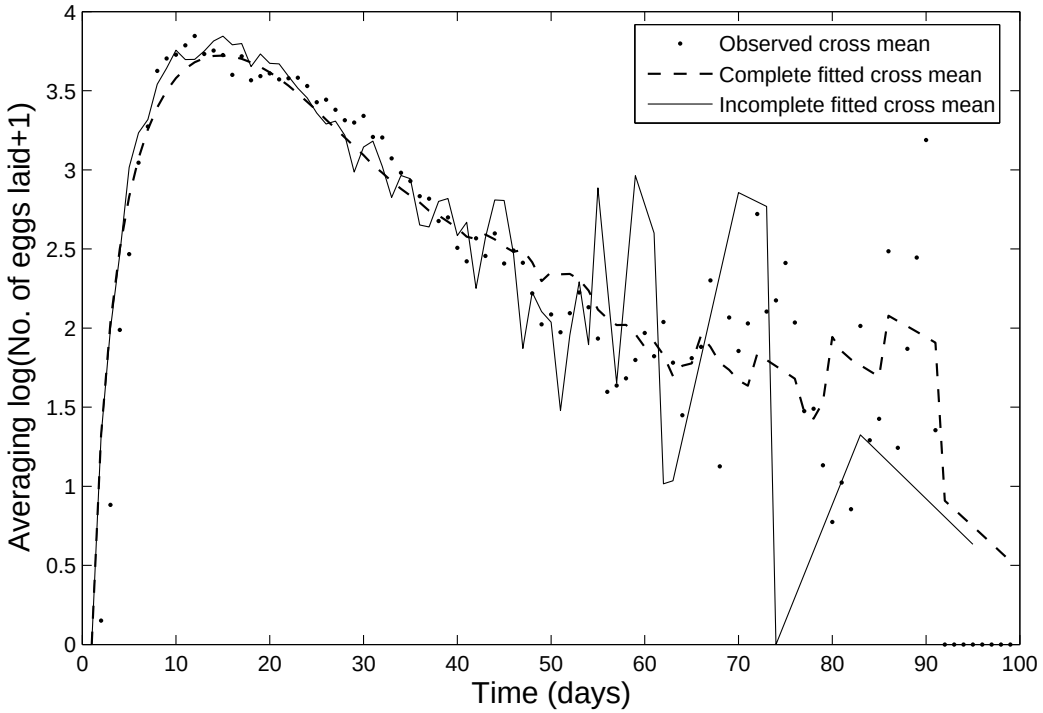

Fig. 4. Fitted Cross-sectional mean curves for complete and incomplete data.

The dots represents the daily mean eggs of those that are still alive.