

**Functional approach of flexibly modelling generalized  
longitudinal data and survival time**

Fang Yao

Department of Statistics

University of Toronto

100 St. George Street

Toronto, ON, M5S 3G3

Canada

Email: *fyao@utstat.toronto.edu*

Phone: +1(416)9465790

Fax: +1(416)9785133

March 29, 2007

## Summary

We propose a flexible functional approach for modelling generalized longitudinal data and survival time using principal components. In the proposed model the longitudinal observations can be continuous or categorical data, such as Gaussian, binomial or Poisson outcomes. We generalize the traditional joint models that treat categorical data as continuous data by using some transformations, such as CD4 counts. The proposed model is data-adaptive, which does not require pre-specified functional forms for longitudinal trajectories and automatically detects characteristic patterns. The longitudinal trajectories observed with measurement error or random error are represented by flexible basis functions through a possibly nonlinear link function, combining dimension reduction techniques resulting from functional principal component analysis. The relationship between the longitudinal process and event history is assessed using a Cox regression model. Although the proposed model inherits the flexibility of nonparametric methods, the estimation procedure based on the EM algorithm is still parametric in computation, and thus simple and easy to implement. The computation is simplified by dimension reduction for random coefficients or functional principal component scores. An iterative selection procedure based on Akaike Information Criterion (AIC) is proposed to choose the tuning parameters, such as the knots of spline basis and the number of functional principal components, so that appropriate degree of smoothness and fluctuation can be addressed. The effectiveness of the proposed approach is illustrated through a simulation study, followed by an application to longitudinal CD4 counts and survival data which were collected in a recent clinical trial to compare the efficiency and safety of two antiretroviral drugs.

*Key words:* Cox regression; EM algorithm; Functional principal components; Generalized linear mixed effects model, Longitudinal data; Survival.

# 1 Introduction

Many scientific investigations, such as clinical trials, collect longitudinal data with repeated measurements for a sample of subjects, and event history data that are possibly censored time-to-event, i.e., “failure” or “survival”. Additional covariate information may be also recorded. A complication that often occurs is that the longitudinal process is usually unobservable due to measurement error or random error. Because of this, if the Cox regression model is used to analyze the survival process, the required longitudinal information at each failure time for all members in the corresponding risk set is not available. It is well known that conventional partial likelihood approaches used for the Cox model cannot avoid biased inference by using some sort of imputation of the latent longitudinal process, such as last-value-carried-forward method (Prentice, 1982), smoothing techniques (Raboud *et al.*, 1993), or “two-stage” approaches (Tsiatis *et al.*, 1995). This invokes the consideration of longitudinal and event processes simultaneously, i.e., the “so-called” joint models, that have attracted substantial research interest and make more efficient use of data by jointly maximizing the likelihood of both processes.

There has been substantial recent work on jointly modelling a continuous longitudinal process and survival history. Typical examples are HIV trials, where a biomarker such as CD4 lymphocyte counts are measured intermittently and time to progression to AIDS or death is recorded, with possible early dropout or failing to experience event by the end of study. Tsiatis *et al.* (1995), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997), Bycott and Taylor (1998) and Dafni and Tsiatis (1998) characterized the longitudinal process by parametric random effects models focusing on smooth trends determined by a small number of random effects. Alternative models consisting of random effects and some mean-zero stochastic processes were proposed by Taylor, Cumberland and Sy (1994), Henderson *et al.* (2000), Wang and Taylor (2001) and Xu and Zeger (2001), and investigated “wiggly” fluctuations that may be caused by a biological mechanism.

The research of this paper is motivated by the fact that in many clinical trials longitudinal observations are not necessarily continuous, but may be categorical, such as binomial or Poisson outcomes. For example, in some cancer clinical trials smoking status, a binary longitudinal covariate of interest, is often related to the progression to cancer. The binary observations are obviously subject to random error that yields dichotomous outcomes and plays a significant role to mask the latent process from observed values. In such cases the mean or probability of the binary observations is more appropriate to be used as the covariate process for modelling the survival time instead of the observed dichotomous outcomes. Another important example is the CD4 count data. It is well known that CD4 counts are usually transformed by fourth-root power or logarithm to achieve normality

and homogeneity of within-subject variation (Taylor *et al.*, 1991). The transformed data are then modelled as continuous variable using linear mixed effects models, and the inference is usually based on Gaussian assumption of the transformed data. An alternative approach proposed in this paper is to model the original CD4 counts as Poisson outcomes. It is clear that linear mixed effects models may not be adequate for modelling a variety of outcome measures. Therefore it is natural and necessary to extend the regression analysis to a general class of models, and simultaneously take the event process into account.

For the modelling of generalized longitudinal outcomes that can be continuous or categorical outcomes, generalized linear mixed models (GLMMs, see Diggle *et al.*, 2002, for introduction and Section 2.2 below for general formulation) are a natural outgrowth of both linear mixed effects models and generalized linear models (McCullagh and Nelder, 1989). They are of wide applicability and practical importance (Breslow and Clayton, 1993). GLMMs enable the accommodation of non-normally distributed outcomes. Specifically they can model within-subject correlation by incorporating random effects for longitudinally measured outcomes. Although GLMM is a rich class of models, its use in practice has been limited by the complexity of the likelihood function. The approaches involving analytical approximation to the likelihood (Goldstein, 1991; Schall, 1991; Breslow and Clayton, 1993; Wolfinger and O’Connell, 1993) are known to be inconsistent under standard (small domain) asymptotic assumptions (Breslow and Lin, 1995; Lin and Breslow, 1996). McCulloch (1994, 1997) explored Monte Carlo EM algorithms using Gibbs chain and Metropolis-Hastings steps to approximate the E-step, while Zeger and Karim (1991) employed a Gibbs sampling approach. An automated Monte Carlo EM algorithm was developed by Booth and Hobert (1999) that used rejection or importance sampling to simulate random samples in the E-step and yields approximately unbiased estimation.

Due to the numerical challenges of evaluating the intractable integrals in both GLMM and the joint model, there is a lack of research in such a generalized joint model framework. Molenberghs *et al.* (1997) combined the multivariate Dale model for longitudinal ordinal data with a logistic regression model for drop-out instead of a survival model. Faucett *et al.* (1998) proposed a joint model to analyze the survival data with binary longitudinal covariate using a Markov model that is not capable of incorporating random effects. Larsen (2005) proposed a joint approach with a two-parameter logistic model that was applied to the Women’s Health and Aging Study. In this paper we develop a framework which jointly models generalized longitudinal outcomes and survival time by combining GLMM and Cox regression into a mega-model. The joint models that consider continuous longitudinal covariate can be viewed as a special case. The joint likelihood is maximized using the Monte Carlo EM algorithm which yields approximately unbiased parameter estimation, as illustrated by the simulation study in Section 4.

As a parametric model will only find those features in the data that have been pre-specified, this may not be adequate if the time course is not well defined and does not fall into the preconceived class. In such situation an analysis through semi- or nonparametric methods is advisable. Functional data analysis attracted substantial interest recently for modelling a sample of trajectories semi- or nonparametrically, see Ramsay and Silverman (1997, 2002) for a summary. In particular, functional principal component (FPC) analysis attempts to find the dominant modes of variation around the overall trend in the data, and is thus a key technique in functional data analysis (Berkey and Kent, 1983; Besse and Ramsay, 1986; Castro *et al.*, 1986; Rice and Silverman, 1991; Silverman, 1996; James *et al.*, 2000; Yao *et al.*, 2003, 2005).

Based on the proposed general framework of the joint model, we extend the standard GLMM with parametric regression of predictors to its semiparametric version with FPCs represented by flexible basis splines and random coefficients (James *et al.*, 2001). The model is data-driven and automatically captures important features using leading FPCs. Another advantage of using FPCs in GLMM is that the dominant modes of variation can often be described by the first few eigenfunctions in practice, i.e., the mean trajectories of longitudinal outcomes can be well approximated by a few leading principal components. This means that only significant variance components need to be included in GLMM, thus avoiding the over-parametrization caused by small variance components and reducing the computational intensity. The tuning parameters such as the numbers of knots of splines and the number of principal components are chosen by an iterative selection procedure using Akaike information criterion (AIC).

The remainder of the paper is organized as follows. In Section 2 we present the framework of the proposed joint model that considers generalized longitudinal data and survival time, the FPC model of the latent process for the generalized longitudinal process is presented in Section 3. Simulation results that illustrate the effectiveness of the proposed method are reported in Section 4, while an application of the proposed model to longitudinal CD4 counts and survival records which were collected in a recent clinical trial to compare the efficacy and safety of two antiretroviral drugs, are provided in Section 5. Technical details are deferred to the Appendix.

## 2 General Method of Modelling Generalized Longitudinal Outcomes and Survival Data

### 2.1 Motivations, notations and assumptions

Our method is motivated by the need to analyze data consisting of time-to-event and longitudinal covariates. The outcome variable may be categorical, such as the simulated data in Section 4 where one has longitudinally sampled binary observations, and the CD4 counts data in Section 5 which are treated as Poisson outcomes. To deal with the generalized longitudinal outcome, there is the need to extend the linear mixed models used in existing joint modelling approaches to a more general classes of models. As the subject-specific pattern is to be characterized in the joint model, this typically needs models with hidden random effects. Therefore GLMM is a natural choice over marginal approaches, such as the generalized estimating equations (GEE) (Diggle *et al.*, 2002).

It is noticed that the conventional partial likelihood approaches used in the Cox model cannot avoid biased inference by using observed or some sort of imputation of the latent longitudinal process. This is due to the fact that the required information may not be available for necessary failure times or is contaminated with measurement error or random error. For generalized longitudinal observations, such bias is also expected and may be more serious, since the random error plays a more significant role to mask the latent process. In this section we develop a general framework of the joint model, combining the GLMM approach for generalized longitudinal covariates and Cox model for survival time. For convenience we assume that the survival time is subject to right censoring, and that the censoring is independent of all other survival and covariate information of interest. The survival time for the  $i$ th individual is denoted by  $S_i$  and the potential censoring time by  $C_i$ . One can only observe  $T_i = \min(S_i, C_i)$  and the failure indicator  $\Delta_i$  which equals to 1 if the failure is observed, i.e.,  $S_i \leq C_i$ , and 0 otherwise.

Here we consider the case of a single underlying process  $\mu(t)$  that is the mean process of the observed generalized longitudinal outcomes, and  $\mu_i(t)$  is the realization of the  $i$ th subject, where  $t \in [0, \tau]$  and  $\tau$  is usually the duration of study. The observations  $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$  for the  $i$ th subject are assumed to be sampled from the latent process  $\mu_i(t)$  intermittently at  $t_i = (t_{i1}, \dots, t_{in_i})^T$  subject to random error  $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^T$ . Then one has  $Y_{ij} = \mu_i(t_{ij}) + \epsilon_{ij}$ ,  $j = 1, \dots, n_i$ , where  $t_{ij} \in [0, \tau]$ ,  $n$  is the number of subjects, and  $n_i$  is the number of longitudinal observations available from the  $i$ th subject. This implies that  $T_i \leq \tau$ , and  $t_{ij} \leq T_i$  for  $1 \leq i \leq n$  and  $1 \leq j \leq n_i$ . Besides the longitudinal and time-to-event data, there might be other covariates that possibly have significant effects on longitudinal or survival processes. Let  $Z_i(t) = (Z_{i1}(t), \dots, Z_{ir}(t))^T$  denote the

vector of covariates valued at time  $t \leq T_i$  associated with the longitudinal process, and  $V_i(t) = (V_{i1}(t), \dots, V_{is}(t))^T$  the vector of covariates associated with the event process. Note that vectors  $Z_i(t)$  and  $V_i(t)$  are possibly time-dependent, and may or may not have elements in common. Assume that the true values of these covariates for the  $i$ th subject can be observed at any time  $t \leq T_i$ , and particularly  $V_i(T_i)$  are available for all subjects. In contrast, the longitudinal covairates  $Y_i$  are assumed to be subject to random error and are only observed on time points  $t_i$ . Denote the  $n_i \times r$  design matrix formed by the covariate  $Z_i(t)$  on  $t_i$  by  $Z_i = (Z_i(t_{i1}), \dots, Z_i(t_{in_i}))^T$ .

To validate the specification of the proposed method, one requires that the timing of measurement process might depend on the observable covariate history and latent longitudinal process, but not additionally on the unobserved future event time itself. For more detail, we refer to Tsiatis and Davidian (2004). For simplicity, in what follows we assume that the measurement process is non-informative. The observed longitudinal covariate is assumed to be independent of event time conditional on the latent longitudinal process  $\mu_i(t)$  and covariates  $Z_i(t)$  and  $V_i(t)$ , and the data from different subjects are generated by independent realizations.

## 2.2 Joint model for generalized longitudinal covairate and survival time

The subject-specific modelling of generalized longitudinal data typically requires models with unobservable random effects as in GLMM. We thus need to specify an additional latent process  $X(t)$  that is related to the latent longitudinal process  $\mu_i(t)$  through a possibly nonlinear link function  $g(\cdot)$ . The function  $g(\cdot)$  is usually a smooth and strictly monotone increasing function. Assume that the latent process  $X$  can be modelled by a linear form with subject-specific random effects for now, i.e., letting  $x_{i,t}^T$  and  $z_{i,t}^T$  be  $p$ - and  $q$ -vectors of covariates associated with  $X_i(t)$ ,  $i = 1, \dots, n$ , then

$$X_i(t) = x_{i,t}^T \beta + z_{i,t}^T u_i, \quad (1)$$

where  $\beta$  is a  $p$ -vector of unknown regression coefficients and  $u_i$  is the  $q$ -vector of unobservable random effects. The assumption in (1) will be relaxed when we discuss FPC representation of  $X_i$ . Note that  $x_{i,t}^T$  and  $z_{i,t}^T$  are not necessarily related to  $Z_i(t)$  that appears in FPC model (8).

Assumptions for modelling the generalized longitudinal observations are that the dependence between the observations  $Y_{ij}$  is inherited from the the unobserved process  $X_i$ , and that the  $Y_{ij}$  arise from a generalized liner model with linear predictor  $X_i(t) = x_{i,t}^T \beta + z_{i,t}^T u_i$ , and  $\mu_i(t_{ij}) = E\{Y_{ij}|X_i(t_{ij})\}$  satisfying  $g(\mu_i(t)) = X_i(t)$ , for some link function  $g(\cdot)$ . Therefore the longitudinal observations  $Y_{ij}$  are conditionally independent with density functions from the canonical exponential family

$$f(Y_{ij}|X_i(t_{ij}), \beta, \phi) = \exp \left[ \frac{w_{ij}}{\phi} \{Y_{ij} \theta_{ij}^* - b(\theta_{ij}^*)\} + c(Y_{ij}, \frac{w_{ij}}{\phi}) \right], \quad (2)$$

where the  $w_{ij}$  are known weights and the conditional mean and canonical parameters  $\theta_{ij}^*$  are related through the equation  $\mu_i(t_{ij}) = b'(\theta_{ij}^*)$  (McCullagh and Nelder, 1989, chapter 2). The specification of the GLMM is completed by assuming that  $u_i$  is a  $q$ -vector random variable with a parametric density  $f(u|\sigma_0^2)$  that depends on an unknown  $q^* \times 1$  vector of variance components  $\sigma_0^2$ .

The hazard of failure is modelled by the original Cox model formulation (Cox, 1972, 1975), where the hazard depends on the longitudinal process  $X_i$  through its current value and other time-dependent or time-independent covariates  $V_i$ . Other aspects of longitudinal trajectories can also be considered. Then the framework for characterizing the association between the generalized longitudinal and survival processes as well as other covariates is given by

$$\begin{aligned} h_i(t) &= \lim_{dt \rightarrow 0} P\{t \leq T_i < t + dt | T_i \geq t, \mu_i^H(t), V_i(t)\} / dt \\ &= h_0(t) \exp\{\gamma \mu_i(t) + V_i(t)^T \eta\}, \end{aligned} \quad (3)$$

where the  $\gamma$  and  $\eta = (\eta_1, \dots, \eta_s)^T$  are the regression coefficients, reflecting the association between the latent longitudinal process (on the observed scale) and survival time, and  $\mu_i^H(t) = \{\mu_i(u) : 0 \leq u < t\}$  is the history of the unobserved longitudinal process  $\mu_i$  up to time  $t$ . The inference is complicated by the fact that  $\mu_i(t)$  is only observed intermittently at  $t_i = (t_{i1}, \dots, t_{in_i})^T$  and subject to random error.

We next combine the generalized longitudinal and survival processes into a joint model. A critical assumption is that the observed values  $Y_{ij}$  of the longitudinal process and the failure times  $(T_i, \Delta_i)$  are conditionally independent given the latent process  $\mu_i(t)$  as well as possible covariates  $Z_i(t)$  and  $V_i(t)$ . Denote the observed data for each individual by  $O_i = \{T_i, \Delta_i, Y_i, Z_i, V_i, t_i\}$ , the vector of  $\mu_i(t)$  valued at  $t_i$  by  $\tilde{\mu}_i = (\mu_i(t_{i1}), \dots, \mu_i(t_{in_i}))^T$ , and the history of  $\mu_i(t)$  prior to  $T_i$  by  $\mu_i^H(T_i) = \{\mu_i(t) : 0 \leq t < T_i\}$ . Note that the trajectories  $\mu_i(t) = g^{-1}(x_{i,t}^T \beta + z_{i,t}^T u_i)$  are determined by the random effects  $u_i$ , i.e., the conditional distributions of  $\{Y_i, T_i, \Delta_i\}$  given  $\mu_i^H(T_i)$  and  $\tilde{\mu}_i$  are in fact only determined by the random effects  $u_i$ . By assuming the conditional independence of  $(T_i, \Delta_i)$  and  $Y_i$  given  $u_i$ , the likelihood of the observed data for the full set of parameters of interest, denoted by  $\Omega = \{\gamma, \eta, h_0(\cdot), \beta, \phi, \sigma_0^2\}$ , is given by

$$L_O = \prod_{i=1}^n \left\{ \int f(T_i, \Delta_i | \mu_i^H(T_i), V_i(T_i), \gamma, \eta, h_0) f(Y_i | \tilde{\mu}_i, t_i, \beta, \phi) f(u_i | \sigma_0^2) du_i \right\}, \quad (4)$$

where

$$\begin{aligned} & f(T_i, \Delta_i | \mu_i^H(T_i), V_i(T_i), \gamma, \eta, h_0) \\ &= [h_0(T_i) \exp\{\gamma \mu_i(T_i) + V_i(T_i)^T \eta\}]^{\Delta_i} \exp \left[ - \int_0^{T_i} h_0(u) \exp\{\gamma \mu_i(u) + V_i(u)^T \eta\} du \right], \end{aligned} \quad (5)$$

$$f(Y_i|\tilde{\mu}_i, t_i, \beta, \phi) = \prod_{j=1}^{n_i} f(y_{ij}|X_i(t_{ij}); \beta, \phi), \quad (6)$$

and  $f(Y_{ij}|X_i(t_{ij}); \beta, \phi)$  is defined in (2). For details of the above formulation, please refer to Wulfsohn and Tsiatis (1997).

A common assumption is that  $u_i$  is a multivariate Gaussian random vector with mean zero and covariance matrix  $\Sigma = \Sigma(\sigma_0^2)$ . However, the Monte Carlo EM algorithm we use for model estimation is not restricted to the models with normally distributed random effects, and can be modified to adapt to the random effects under other distributional assumptions. Nevertheless, from the simulation study reported in Section 4, the procedure implemented with Gaussian assumption under the non-normal scenario yields results comparable to those obtained from the normal scenario. Similar ‘‘robustness’’ in non-normal situation has also been observed in Tsiatis and Davidian (2004) for joint modelling of continuous data.

The EM algorithm described by Wulfsohn and Tsiatis (1997) can be extended to the proposed model. Although the conditional distribution of the random effects  $u_i$  given the generalized longitudinal data  $Y_i$  involves intractable integrals and can not be written in closed form, we can obtain random samples of  $u_i$  in E-step from its exact distribution by rejection sampling using the marginal distribution  $f(u|\sigma_0^2)$  as a candidate, as employed by Booth and Hobert (1999). The M-step for estimating  $\beta$ ,  $\phi$  and  $\sigma_0^2$  is similar to that in GLMM context, as  $Y_i$  and  $(T_i, \Delta_i)$  are conditionally independent given the latent process  $\mu_i$ . Then  $\beta$  and  $\phi$  can be estimated by iteratively weighted least squares conditional on  $u_i$ , and the estimate of  $\sigma_0^2$  can sometimes be written in a closed form, depending on the distribution of  $u_i$ , such as Gaussian distribution. The estimation of the Cox regression coefficients  $\gamma$  and  $\eta$  is achieved by maximizing the Monte Carlo approximation of the partial likelihood of  $(T_i, \Delta_i)$  using Newton-Raphson algorithm and the baseline hazard  $h_0(\cdot)$  is then estimated by the Breslow estimator. See the Appendix for details of the estimation procedure.

### 3 Flexibly Modelling Generalized Longitudinal data through Functional Principal Components

As the parametric model (1) can only find features that have been incorporated *a priori*, it may not be adequate if the  $X_i$  are not well defined or do not fall into the pre-specified class. We now extend the standard GLMM with a parametric form (1) to a semiparametric version. The underlying processes  $X_i$  will be modelled by a set of FPCs through flexible spline basis and random coefficients (James *et al.*, 2001) in joint model context. The advantage of FPC model is that one does not need to

pre-specify a parametric form for the mean process  $\mu_i(t)$  of the longitudinal trajectories  $X_i(t)$ , as the model itself is data-driven and automatically captures important features by the leading principal components.

Recall that  $\mu_i(t)$  is the mean trajectory of the generalized data  $Y_i$ ,  $X_i(t)$  is the  $i$ th realization of the latent process  $X(t)$ , satisfying  $g(\mu_i(t)) = X_i(t)$ . Let  $\mu_X$  be the overall mean function of  $X_i$ . Considering the vector of covariates  $Z_i(t)$ , let

$$\mu_{X_i}(t) = \mu_X(t|Z_i) = \mu_X(t) + Z_i(t)^T \alpha, \quad t \in [0, \tau], \quad (7)$$

where  $\alpha = (\alpha_1, \dots, \alpha_r)^T$  and  $Z_i(t) = (Z_{i1}(t), \dots, Z_{ir}(t))^T$ . The covariance structure of  $X_i(t)$  might also depend on components of  $Z_i(t)$ , e.g., if  $Z_i(t)$  contains a treatment indicator. For convenience, we take the common specification that the same covariance structure holds for all subjects, denoted by  $G(s, t)$ , i.e.,  $G(s, t) = \text{cov}(X_i(s), X_i(t))$ . Assume that there exists an orthogonal expansion (in the  $L^2$  sense) of  $G$  in terms of eigenfunctions  $\{\phi_k\}_{k=1,2,\dots}$  and non-decreasing eigenvalues  $\{\lambda_k\}_{k=1,2,\dots}$ , i.e.,  $G(s, t) = \sum_k \lambda_k \phi_k(s) \phi_k(t)$ ,  $s, t \in [0, \tau]$ . Karhunen-Loéve representation in classical FPC analysis implies that the individual trajectories can be expressed as  $X_i(t) = \mu_{X_i}(t) + \sum_k \xi_{ik} \phi_k(t)$ , where  $\mu_{X_i}(t)$  is the mean function of  $X_i(t)$ , the coefficients  $\xi_{ik} = \int_0^\tau \{X_i(t) - \mu_{X_i}(t)\} \phi_k(t) dt$  are uncorrelated random variables with mean zero and variances  $E\xi_{ik}^2 = \lambda_k$  subject to  $\sum_k \lambda_k < \infty$ .

For the purpose of characterizing the association between the dominant trends of the generalized longitudinal process and event process, assume that the covariance function  $G$  can be well approximated by the first few principal components, i.e., the eigenvalues  $\lambda_k$  tend to zero rapidly so that the variability is predominantly of large scale and low frequency. The individual trajectories can be approximately modelled by the first  $K$  leading principal components, i.e.,  $X_i(t) \approx \mu_{X_i}(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t)$ . One can adjust the truncation parameter  $K$  to capture more or less “wiggly” patterns. Then the FPC model for the underlying process  $X_i$  from which the generalized longitudinal data  $Y_{ij}$  are observed, is given by,

$$X_i(t) = \mu_X(t) + Z_i(t)^T \alpha + \sum_{k=1}^K \xi_{ik} \phi_k(t), \quad t \in [0, \tau]. \quad (8)$$

The overall mean function and covariance surface, and thus eigenfunctions are often assumed to be smooth. We model these functions using flexible basis functions, such as B-splines or regression splines. Let  $\bar{B}_p(t) = (\bar{B}_{p1}(t), \dots, \bar{B}_{pp}(t))^T$  be a set of basis functions on  $[0, \tau]$  for modelling the overall mean function  $\mu_X(t)$  with coefficients  $\beta = (\beta_1, \dots, \beta_p)^T$ , i.e.,  $\mu_X(t) = \bar{B}_p(t)^T \beta$ . Subject to the orthonormality of  $\{\phi_k\}_{k=1,\dots,K}$ , the eigenfunctions are represented by a set of orthonormal basis

functions  $B_q(t) = (B_{q1}(t), \dots, B_{qq}(t))^T$  with coefficients  $\theta_k = (\theta_{1k}, \dots, \theta_{qk})^T$  that are subject to

$$\int_0^\tau B_{q\kappa}(t)B_{q\ell}(t)dt = \delta_{\kappa\ell}, \quad \theta_k^T \theta_l = \delta_{kl}, \quad \kappa, \ell = 1, \dots, q, \quad k, l = 1, \dots, K, \quad (9)$$

where  $\delta_{ij} = 1$  if  $i = j$  and 0 otherwise.

Then model (8) becomes, letting  $\xi_i = (\xi_{i1}, \dots, \xi_{iK})^T$  and  $\Theta = (\theta_1, \dots, \theta_K)^T$ ,

$$X_i(t) = \bar{B}_p(t)^T \beta + Z_i(t)^T \alpha + B_q(t)^T \Theta \xi_i. \quad (10)$$

The joint modelling of  $\{Y_i, T_i, \Delta_i\}$  proceeds as described in Section 2.2 with the parametric model (1) replaced by the FPC model (8) for the latent process  $X_i$ . The implementation of the Monte Carlo EM algorithm is essentially the same except that the covariance  $\Sigma(\sigma_0^2)$  of the random effects  $u_i$  is replaced by  $\Theta \Lambda \Theta^T$  for the FPC scores  $\xi_i$ , where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$ . In the M-step, the estimation of  $\Lambda$  can be written in closed form sometimes, e.g., under Gaussian assumption of  $\xi_i$ . The estimation of  $\Theta$  can be obtained by another iterative procedure inside the iteratively weighted least squares, i.e., one estimates  $\theta_k$  at each time with the other columns fixed and iterates until no further change in  $\hat{\Theta}$ .

Note that the FPC model involves smoothing the longitudinal trajectories and choosing appropriate number of FPCs. We need to select tuning parameters so that the association between the generalized longitudinal and survival processes can be appropriately addressed to certain degree of smoothness. We use spline basis, such as B-splines or regression splines, and choose equi-quantile knots to avoid clustered observation times, particularly in the case of decreasing number of observations due to failure or censoring. For example, 25th, 50th and 75th percentiles of pooled observation times from all individuals will be used, if 3 inside knots are needed. Besides the choice of knots sequence, for FPC analysis, it is particularly important to identify the number of leading principal components that are needed to approximate the infinite-dimensional longitudinal process. Note that the number of eigenfunctions  $K$  and the dimensions of spline basis  $p$  and  $q$  are simultaneously related to the performance of the model. We adapt the Akaike information criterion (AIC) to the proposed model. The pseudo-Gaussian joint likelihood depending on  $K$ ,  $p$  and  $q$ , summing the contributions from all subjects and conditional on the estimated FPC scores  $\hat{\xi}_i$ , is given as follows,

$$\hat{l}(K, p, q) = \sum_{i=1}^n \left[ \log\{f(T_i, \Delta_i | \hat{\mu}_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\eta}, \hat{h}_0)\} + \log\{f(Y_i | \hat{\mu}_i, t_i, \hat{\alpha}, \hat{\beta}, \hat{\Theta}, \hat{\phi})\} \right], \quad (11)$$

where the densities of  $(T_i, \Delta_i)$  and  $Y_i$  are as in (5) and (6) with the estimated parameters  $\hat{\alpha}$ ,  $\hat{\beta}$ ,  $\hat{\Theta}$  defined in (10), and “ $\hat{\cdot}$ ” is the generic notation for the estimates obtained from the Monte Carlo EM algorithm. One has  $\hat{\mu}_i^H(T_i) = g^{-1}(\hat{X}_i^H(T_i))$  and  $\hat{\mu}_i = g^{-1}(\hat{X}_i)$ , where  $g^{-1}(\hat{X}_i^H(T_i)) =$

$\{g^{-1}(\widehat{X}_i(t)) : 0 \leq t < T_i\}$ ,  $g^{-1}(\widehat{X}_i) = (g^{-1}(\widehat{X}_i(t_{i1})), \dots, g^{-1}(\widehat{X}_i(t_{in_i})))^T$ , and  $\widehat{X}_i(t) = \bar{B}_p(t)^T \hat{\alpha} + Z_i(t)^T \hat{\beta} + B_q(t)^T \widehat{\Theta} \hat{\xi}_i$ . Then the AIC of the model involving  $K$ ,  $p$  and  $q$  is given by

$$AIC(K, p, q) = -2\hat{l}(K, p, q) + 2\{p + (K + 1)q + r + s + 1\}. \quad (12)$$

Minimization of AIC with respect to  $K$ ,  $p$  and  $q$  simultaneously requires intensive computation. Alternatively, we start with initial guesses for  $p$  and  $q$ , choose  $K$  using AIC (12), then choose  $p$  and  $q$  in turn using AIC, and then repeat until there is no further change. It has been observed in simulation studies and also in the data application that this iterative procedure usually converges fast (in 2 or 3 iterations) and is practically feasible.

## 4 Simulation

An extensive literature has shown that, when the longitudinal model is correctly specified during estimation procedure, the joint modelling approaches for continuous longitudinal and survival data improve the parameter estimation upon some sort of imputation of the latent longitudinal process, such as last-value-carried-forward method (Prentice, 1982). In this section, we examined the behavior of the joint modelling of generalized longitudinal observations and event data. More specifically binary outcomes, a typical type of generalized data, were used. To illustrate the flexibility and adaptiveness of the FPC model, we did not use any information about the true longitudinal model during estimation procedure. In addition, the simulation study also demonstrates the notable ‘‘robustness’’ of the likelihood approach with the random effects or FPC scores taken to be non-normal random variables.

We constructed a scenario where the assumptions on censoring and timing of measurements were satisfied. Comparisons to the ‘‘naive’’ approach, i.e., last-value-carried-forward method, as well as the ideal case where the true longitudinal trajectories are completely known, were provided. Both 400 normal and 400 non-normal samples consisting of  $n = 100$  individuals were considered to demonstrate the robustness of the procedure to the Gaussian assumption. Assume that  $\eta = 0$ ,  $\gamma = -1.0$  in the survival model (3) with Weibull baseline  $h_0(t) = 3t^2$  for  $t \geq 0$ . Censoring times  $C_i$  generated independently of all other variables are i.i.d. Weibull random variables with 10% dropouts at  $t = 0.6$ , 40% dropouts at  $t = 0.8$ , and a final truncation time of  $\tau = 1$ .

The longitudinal process has a mean function  $\mu_X(t) = 3 \sin(3\pi t/2)$  with  $\alpha = 0_r$  in model (7), and a covariance function derived from one eigenfunction  $\phi_1(t) = -\sqrt{2} \cos(\pi t)$ ,  $0 \leq t \leq 1$ . We chose  $\lambda_1 = 4$  and  $\lambda_k = 0$  for  $k \geq 2$  as eigenvalues. Note that the longitudinal process can not be expressed by

a simple polynomial. For the normal samples, the FPC scores  $\xi_{ik}$  were generated from  $\mathcal{N}(0, \lambda_k)$ , while the  $\xi_{ik}$  for the non-normal samples were generated from a mixture of two normals,  $\mathcal{N}(\sqrt{\lambda_k/2}, \lambda_k/2)$  with probability 1/2 and  $\mathcal{N}(-\sqrt{\lambda_k/2}, \lambda_k/2)$  with probability 1/2. Then the outcomes  $Y_{ij}$  were generated from a binary distribution with the probability  $\mu_i(t_{ij}) = g^{-1}(X_i(t_{ij}))$ , where the canonical link function,  $g(p) = \log\{p/(1-p)\}$  for  $0 < p < 1$ . For an equally spaced grid  $\{c_0, \dots, c_{50}\}$  on  $[0, 1]$  with  $c_0 = 0$ ,  $c_{50} = 1$ , let  $s_i = c_i + e_i$ , where  $e_i$  are i.i.d. from  $\mathcal{N}(0, 0.004^2)$ ,  $s_i = 0$  if  $s_i < 0$  and  $s_i = 1$  if  $s_i > 1$ , allowing for non-equidistant “jittered” designs. Each curve was sampled at a random number of points, chosen from a discrete uniform distribution on  $\{15, \dots, 20\}$ , and the locations of the measurements were randomly chosen from  $\{c_1, \dots, c_{50}\}$  without replacement.

For each normal and mixture sample,  $\gamma$  was estimated in three ways:

- using the “ideal” approach, where  $\mu_i(t) = g^{-1}(X_i(t))$  is known for all  $0 \leq t \leq 1$  and  $\gamma$  was estimated by conventional partial likelihood method (Cox, 1975), denoted by IDEAL;
- using the last-value-carried-forward approach, and  $\gamma$  was estimated by conventional partial likelihood method (Cox, 1975), denoted by LVCF;
- using the the proposed joint modelling approach with FPC model, denoted by JFPC, where the number of eigenfunctions  $K$  and the inside equi-quantile knots  $p$  and  $q$  were chosen objectively by iteratively inspecting the AIC (12). Note that we do not specify any particular form for the longitudinal model, but let the FPC model itself characterize the relationship.

From the results summarized in Table 1, one can see that, the simple imputation technique, last-value-carried-forward method, led to biased parameter estimation. Such bias is more serious than that observed in joint models for continuous longitudinal data, as the binary outcomes play a more significant role in masking the latent probability. The proposed joint model with FPCs yields approximately unbiased estimates in both normal and mixture scenarios. These results are comparable to those obtained from the IDEAL case where the true trajectories  $X_i(t)$  are used in estimation. This suggests that the proposed approach is applicable for generalized (especially categorical) longitudinal outcomes and survival data. Moreover, without the knowledge of the true longitudinal process, the proposed model automatically detects the underlying relationship and provides satisfactory approximation to the true functional form due to its flexibility. It is notable that the Gaussian assumption does not compromise the accuracy of the estimation of  $\gamma$  under the mixture scenario, which is similar to those observed in the joint models for continuous longitudinal data (e.g. Tsiatis and Davidian, 2004). For comparison, we also included the case with  $\lambda_1 = 9$  and  $\lambda_k = 0$  for  $k \geq 2$  for both normal and non-normal situations in Table 1. Note that the noise of the binomial model is determined

by the longitudinal process through  $\mu_i(t)\{1 - \mu_i(t)\}$ , where  $\mu_i(t) = \text{logit}\{X_i(t)\}$ . The relationship between noise and signal here is more complex than that in continuous joint models. Nevertheless one can see that the approximate unbiasedness and the robustness to Gaussian assumption still hold for different levels of noise. Regarding the selection of the tuning parameters, such as the numbers of eigenfunctions and inside equi-quantile knots, i.e.,  $K$ ,  $p$  and  $q$ , we used the information criterion AIC calculated from (12), and  $K = 1$  was correctly chosen for most (more than 95%) of the simulated datasets. This provides the empirical support for the proposed selection procedure.

For practical implementation, the Monte Carlo sample size  $N$  is increased with predetermined number of iterations through an *ad hoc* method as employed in McCulloch (1997). The Monte Carlo estimates usually reach the neighborhood of the exact estimates fast, but continue to show random variation due to Monte Carlo error. From our experience, the Monte Carlo size required for stochastic estimates to converge with three or four decimal accuracy (relative to the estimates themselves) is often very large and timely prohibitive. For time saving in simulation, we used  $N = 50$  for iterations 1-10,  $N = 200$  for iterations 11-30 and  $N = 500$  for iterations 31-50, which results in about two-decimal relative accuracy. More discussion on the choices of Monte Carlo sample sizes and stopping rules are given in Section 6. The rejection sampling scheme is the same as that in standard GLMM implemented by Monte Carlo EM algorithm. So are the rejection rates.

## 5 Application to longitudinal CD4 count and survival data

In a recent clinical trial both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs in treating patients that were intolerant of or failed zidovudine (AZT) therapy. Totally 467 HIV-infected patients who met entry conditions (either an AIDS diagnosis or two CD4 counts of 300 or less, and fulfilling specific criteria for AZT intolerance or failure) were enrolled in this trial and randomly assigned to receive either zalcitabine (ddC) or didanosine (ddI) treatment. CD4 counts were recorded at study entry and again at the 2-, 6-, 12- and 18-month visits. The time to death was recorded. For full details regarding the conduct of the trial and data description, see Abrams *et al.* (1994), Goldman *et al.* (1996) and Guo and Carlin (2004).

To demonstrate the proposed method, we focus on investigating the association among CD4 counts of two drug groups (ddC and ddI) and survival time, including the 160 patients that had no previous infection (AIDS diagnosis) at study entry in the following analysis. We apply the proposed joint model that treats the CD4 count as categorical data. In literature CD4 counts were often transformed by logarithm to achieve homogeneity of within-subject variance so that

linear mixed effects models can be applied. However, it is more natural to model the original CD4 counts using Poisson model with a logarithm link function as specified in (2), where now  $g(\mu_i(t)) = \log(\mu_i(t)) = X_i(t)$ ,  $\mu_i(t)$  are the longitudinal trajectories that generate observed CD4 counts  $Y_{ij}$  at  $t_{ij}$ , and  $X_i(t)$  are the underlying process which corresponds to the log-transformed true CD4 trajectories. From Figure 1 which displays the CD4 counts of 16 randomly selected patients on log scale, one can see that the CD4 counts are enormously noisy and fluctuate dramatically within subject. It is not easy to find appropriate pre-specified parametric forms for the mean and variation of CD4 trajectories. Therefore we incorporate FPCs, where the mean CD4 curves of two groups are modelled separately using B-spline basis, and a common covariance structure is used for the two groups due to similar variation patterns. The model for  $X_i(t)$  becomes

$$X_i(t_{ij}) = \mu_{g_i}(t_{ij}) + \sum_{k=1}^K \xi_{ik} \phi_k(t_{ij}) = \bar{B}_p(t_{ij})^T (\alpha + g_i \beta) + B_q(t_{ij})^T \Theta \xi_i, \quad (13)$$

where  $g_i = 0$  for the ddC group and  $g_i = 1$  for the ddI group,  $K$  is the number of principal components that would be chosen by the iterative procedure together with  $p$  and  $q$  based on AIC (12). The vector of coefficients  $\beta = (\beta_1, \dots, \beta_p)^T$  is to model the difference between two drug groups, the other notations are the same as in model (8) and (10).

For comparison, we also look at the joint model that treats the log-transformed CD4 count as continuous data. More specifically, the log-transformed CD4 counts are incorporated into the FPC model (13) with  $X_i(t_{ij})$  substituted by  $\log(Y_{ij})$ , random errors  $\epsilon_{ij}$  are i.i.d. from  $N(0, \sigma^2)$ , and the link becomes an identity function, i.e.,

$$\log(Y_{ij}) = X_i(t_{ij}) + \epsilon_{ij} = \bar{B}_p(t_{ij})^T (\alpha + g_i \beta) + B_q(t_{ij})^T \Theta \xi_i + \epsilon_{ij}. \quad (14)$$

One can see that this is similar to the comparison between standard and generalized linear models. Here we model  $X_i(t)$  as the covariate process in the Cox regression model,

$$h_i(t) = h_0(t) \exp\{\gamma X_i(t) + \eta g_i\}, \quad t \in [0, \tau], \quad (15)$$

where the duration of the study is 21.4 weeks ( $\tau = 21.4$ ), and the other notations are as in (3). Let the average prediction errors  $APE = (1/n) \sum_{i=1}^n (1/n_i) \sum_{j=1}^{n_i} \{\log(Y_{ij}) - \widehat{X}_i(t_{ij})\}^2$ , where  $\widehat{X}_i(t_{ij})$  are obtained by fitting the joint model with Poisson FPC model (13) or linear FPC model (14). Note that  $\widehat{X}_i(t_{ij})$  are on log scale in both models. The Poisson model yields  $APE=0.0827$  that reduces  $APE=0.0953$  obtained from the linear FPC model by around 15%. Although cross-validation may be employed for such a comparison to take model complexity into account, for computational convenience, we did not obtain cross-validated comparison here. Figure 2 displays the fitted values  $\widehat{X}_i(t_{ij})$  from the joint Poisson FPC model that reasonably agree with the observed values on log

scale, which also validates the use of the Poisson FPC model (13). Here we used the Monte Carlo sizes  $N = 50$  for iterations 1-10,  $N = 200$  for iterations 11-30,  $N = 500$  for iterations 31-50 and  $N = 1000$  for iterations 51-80 so that the resulting Monte Carlo estimates converge with two decimal relative accuracy.

Smooth estimates of the mean CD4 trajectories on log scale,  $\mu_{g_i}(t)$ , of two drug groups obtained from the joint Poisson FPC model are shown in the left panel of Figure 3, which presents similar patterns for the two groups with slightly different changing rates. The 2-week flat period at the beginning of both groups may correspond to better health conditions of patients when they entered the study. It is noticed that the mean curves of the both groups decrease till middle of the study, then increase to a peak and then drop rapidly. One needs to interpret these estimates with caution, as bootstrap confidence bands of the mean functions (not reported here) support the shape of the mean functions but no significant difference between the mean trends (consistent with the results in Guo and Carlin, 2004). However, since the emphasis is to study the influence of CD4 levels on the survival of two groups, we still model two groups respectively in the joint model. Three eigenfunctions shown in the right panel of Figure 3 are used to approximate the infinite-dimensional process. The choices  $K = 3$ ,  $p = 6$  and  $q = 6$  are suggested by the iterative selection procedure based on AIC (12). Here  $p = 6$  and  $q = 6$  imply that 4 knots are selected for both B-spline basis, and that the inside knots are the 33th and 66th percentiles of the pooled observation times. The first two eigenfunctions are somewhat similar to the mean functions, and the third one indicates a slight contrast between early and very late times, accounting for about 68%, 28% and 2% of the total variation respectively. The fitted longitudinal CD4 trajectories on log scale obtained by joint Poisson FPC model for 8 randomly selected patients from each group,  $\hat{X}_i(t) = \log(\hat{\mu}_i(t)) = \hat{\mu}_{g_i}(t_{ij}) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t_{ij})$ , are shown in Figure 1. The fitted curves are seen to be reasonably close to the observations. Even for individuals with nonlinear patterns, we can still effectively recover their trajectories using 3 leading FPCs determined by the data. It may be difficult to anticipate such features regarding population- and subject-specific patterns from traditional parametric models, e.g., linear or quadratic random effects models.

Regarding the survival in the two drug groups, the empirical survival curves (Nelson-Aalen estimates) are shown in Figure 4, indicating that the cumulative hazards of the ddI group is increasingly higher than that of the ddC group with time elapsing. The coefficient  $\gamma$  that describes the strength of relationship between CD4 on log scale and survival is estimated as  $-0.416$  by jointly maximizing the likelihoods of Poisson FPC and Cox models. The 95% confidence interval  $(-0.660, -0.171)$  indicates the significance of the relationship. This means that, for a fixed time, a reduction of CD4 count by 1 on log scale will result in the risk of death increased by 52% with the 95% confidence interval

(19%, 93%). By comparison, the joint linear FPC model yields  $\hat{\gamma} = -0.297$  with 95% confidence interval  $(-0.487, -0.107)$ . This provides evidence that the joint Poisson FPC model tends to characterize the relationship between CD4 counts and survival more significantly than the joint linear FPC model. The membership of the drug group also plays a significant role in determining the survival, where the estimate  $\hat{\eta} = 1.098$  with 95% confidence interval  $(0.273, 1.923)$ . This suggests that, if two patients have the same CD4 counts, the risk of death for the patient of the ddI group is about 3 times compared to that of the ddC group [confidence interval  $(1.31, 6.84)$ ]. The estimated average cumulative hazards, given by

$$\hat{H}_{g_i}(t) = \int_0^t \hat{h}_0(s) \exp\{\hat{\gamma}\mu_{g_i}(s) + \hat{\eta}g_i\} ds, \quad (16)$$

for the patients of ddC and ddI groups are shown in Figure 4. This suggests that the average cumulative hazard of the ddC group is increasingly lower than that of the ddI group. The estimates obtained from the proposed model reasonably agree with the empirical estimates (Nelson-Aalen estimates) and are well covered by the 95% empirical confidence bands, showing evidence that the proposed model provides a reasonable and adequate fit to the data.

## 6 Discussion

In this paper we propose a flexible approach for jointly modelling generalized longitudinal and survival data using spline-based GLMM and Cox models. In particular, we exploit FPCs that capture the dominant modes of variation in longitudinal trajectories. This data-adaptive approach does not require pre-specified functional form for longitudinal trajectories, and automatically detects important patterns. Although the proposed model possesses nonparametric flexibility, the estimation procedure based on Monte Carlo EM algorithm is intrinsically parametric, and thus is straightforward to implement. The dimension reduction achieved by FPC analysis also reduces computational cost. An iterative selection procedure based on AIC is proposed to choose tuning parameters, such as the knots of the spline functions and the number of principal components. For implementation, an *ad hoc* method is used to increase the Monte Carlo sample size as in McCulloch (1997) due to the complexity of the proposed model. Booth and Hobert (1999) discussed an automated Monte Carlo EM algorithm for GLMM that chooses Monte Carlo sample size and stopping rules by taking Monte Carlo error into account. Developing such automated procedures for the proposed joint modelling approach is a challenging task and deserves further investigation.

## ACKNOWLEDGMENT

The author thanks the editor, an associate editor and two anonymous referees for their constructive remarks. He also acknowledges Professor Radu V. Craiu for his helpful comments that improved the presentation. The research is partially supported by an individual Discovery grant from the Natural Sciences and Engineering Research Council of Canada and a Connaught grant from the University of Toronto.

## APPENDIX

### *Monte Carlo EM Algorithm*

In this section we provide details of the Monte Carlo EM algorithm for jointly modelling the generalized longitudinal and survival data described in Section 2.2., as well as the model with FPCs introduced in Section 3.

1. *E-step.* (i) Consider the random coefficients,  $u_i = (u_{i1}, \dots, u_{iq})^T$  as missing data. The conditional expectation of some function  $l(\cdot)$  of the random coefficients is denoted by  $E\{l(u_i)|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega}\}$ , where  $\widehat{\Omega}$  is the set of current estimates of parameters  $\Omega = \{\gamma, \zeta, h_0(\cdot), \beta, \phi, \sigma_0^2\}$ . This expectation is taken with respect to the conditional density  $f(u_i|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega})$  and can be written as that in Wulfsohn and Tsiatis (1997). Under the conditional independence of  $(T_i, \Delta_i)$  and  $Y_i$  given  $\mu_i(\cdot)$  and other covariates  $Z_i$ , this conditional expectation is given by

$$\frac{\int l(u_i) f(T_i, \Delta_i | \mu_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0) f(u_i | Y_i, t_i, \hat{\beta}, \hat{\phi}, \hat{\sigma}_0^2) du_i}{\int f(T_i, \Delta_i | \mu_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0) f(u_i | Y_i, t_i, \hat{\beta}, \hat{\phi}, \hat{\sigma}_0^2) du_i}, \quad (17)$$

where  $f(T_i, \Delta_i | \mu_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0)$  is as in (5). Outside the Gaussian mixed model setting, the conditional density of  $u_i$  given  $Y_i$  is typically a non-standard multivariate density depending on unknown (normalizing) constraint,

$$f(u_i | Y_i, t_i, \beta, \phi, \sigma_0^2) = c f(u_i | \sigma_0^2) \prod_i^{n_i} f(Y_{ij} | X_i(t_{ij}); \beta, \phi), \quad (18)$$

where  $f(Y_{ij} | X_i(t_{ij}); \beta, \phi)$  is as in (2), and  $f(u_i | \sigma_0^2)$  is assumed to be multivariate Gaussian density with mean zero and covariance matrix  $\Sigma = \Sigma(\sigma_0^2)$ , and  $c$  is the unknown normalizing constant depending on the data and parameters. Following Booth and Hobert (1999), we use rejection sampling to obtain a random sample of  $u_i$  from the exact distribution (18) using  $f(u_i | \sigma_0^2)$  as a candidate.

- (ii) while FPC model (10) is used to characterize the underlying process of the generalized longitudinal observations, the FPC scores  $\xi_i = (\xi_{i1}, \dots, \xi_{iK})^T$  are treated as missing data with the set of parameters  $\Omega = \{\gamma, \zeta, h_0(\cdot), \alpha, \beta, \Theta, \Lambda\}$ .

2. *M-step.* (i) Denote the estimate of the conditional expectation  $E\{l(u_i)|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega}\}$  by  $E_i\{l(u_i)\}$  for convenience. Since the generalized longitudinal observations  $Y_i$  and the event data  $(T_i, \Delta_i)$  are conditionally independent given the latent process  $X_i$ , the observed likelihood is given by (4) with separate parameters for  $f(T_i, \Delta_i|\mu_i^H(T_i), V_i(T_i), \gamma, \zeta, h_0)$ ,  $f(Y_i|\tilde{\mu}_i, t_i, \beta, \phi)$  and  $f(u_i|\sigma_0^2)$ .

Applying iteratively weighted least squares conditional on Monte Carlo estimates of some functionals of  $u_i$ , the estimation of  $\beta$ ,  $\phi$  and  $\sigma_0^2$  are obtained as in the GLMM context (McCulloch, 1997). More specifically, denote the canonical parameters  $\theta_i^* = (\theta_{i1}^*, \dots, \theta_{in_i}^*)^T$ , let the design matrices  $X_i^* = (x_{i,t_{i1}}, \dots, x_{i,t_{in_i}})^T$ ,  $Z_i^* = (z_{i,t_{i1}}, \dots, z_{i,t_{in_i}})^T$  and  $v(\cdot)$  be the variance function of  $Y_{ij}$ , i.e.,  $\text{var}(Y_{ij}|X_i(t_{ij})) = \phi v(\mu_i(t_{ij}))$ . Let  $W_i^{-1}(\theta_i^*, u_i) = W_i^{-1} = \text{diag}\{[g^{(1)}(\mu_i(t_{ij}))]^2 v(\mu_i(t_{ij}))\}_{j=1, \dots, n_i}$  and  $d_i(\theta_i^*, u_i) = d_i = (d_{i1}, \dots, d_{in_i})^T$ , where  $d_{ij} = d_{ij}(\theta_i^*, u_i) = X_i(t_{ij}) + (Y_{ij} - \mu_{ij})g^{(1)}(\mu_i(t_{ij}))$ . Then one has the solution to the iteratively weighted least squares, at the  $l$ th iteration,

$$\hat{\beta}^{(l)} = \left\{ \sum_{i=1}^n X_i^{*T} E_i^{(l-1)}(W_i) X_i^* \right\}^{-1} \left[ \sum_{i=1}^n X_i^{*T} E_i^{(l-1)}\{W_i(d_i - Z_i^* u_i)\} \right], \quad (19)$$

where  $E_i^{(l-1)}(W_i)$  and  $E_i^{(l-1)}\{W_i(d_i - Z_i^* u_i)\}$  are the Monte Carlo estimates of the nonlinear functionals  $W_i$  and  $W_i(d_i - Z_i^* u_i)$  with respect to  $u_i$  at the  $(l-1)$ th iteration. The parameter  $\phi$  can be estimated through a scoring equation by solving  $\partial E_i\{\log f(Y_i|\tilde{\mu}_i, t_i, \beta, \phi)\}/\partial \phi = 0$  or a Fisher scoring equation. The estimation of  $\sigma_0^2$  only involves the distribution of  $f(u_i|\sigma_0^2)$  and is often fairly easy to solve, for example, when  $f(u_i|\sigma_0^2)$  is a multivariate normal density.

The parameter of interest in the Cox model  $\zeta = (\gamma, \eta^T)^T$  is estimated by a third iterative procedure, Newton-Raphson algorithm, i.e., at the  $l$ th iteration,

$$\hat{\zeta}^{(l)} = \hat{\zeta}^{(l-1)} + I_{\hat{\zeta}^{(l-1)}}^{-1} S_{\hat{\zeta}^{(l-1)}}, \quad (20)$$

where  $S_{\hat{\zeta}^{(l-1)}}$  and  $I_{\hat{\zeta}^{(l-1)}}$  are the score and the observed information of the partial likelihood of  $(T_i, \Delta_i)$  valued at the  $(l-1)$ th iteration by plugging in  $\hat{\zeta}^{(l-1)}$ , see Wulfsohn and Tsiatis (1997) for explicit expressions. The baseline hazard  $h_0(t)$  can then be estimated by

$$\hat{h}_0(t) = \sum_{i=1}^n \frac{\Delta_i I(T_i = t)}{\sum_{j=1}^n E_j[\exp\{\gamma \mu_j(t) + V_j(t)^T \eta\}] R_j(t)}, \quad (21)$$

where  $R_j(t)$  is a risk indicator that is equal to  $I(T_j \geq t)$ , and  $I(\cdot)$  is the indicator function.

- (ii) When the FPC model (10) is used, denoting the Monte Carlo estimates of the conditional expectation  $E\{l(\xi_i)|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega}\}$  by  $E_i\{l(\xi_i)\}$ , the estimation of  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_K\}$  only involves  $f(\xi_i|\Lambda)$  and is often achieved by  $\hat{\lambda}_k = \sum_{i=1}^n E_i(\xi_{ik}^2)/n$  under Gaussian assumption of

$\xi_i$ . The estimation of  $\zeta$  and  $h_0$  are obtained by (20) and (21). Comparing (1) and (10), let  $\bar{B}_i = (\bar{B}_p(t_{i1}), \dots, \bar{B}_p(t_{in_i}))^T$ ,  $Z_i^* = (Z_i(t_{i1}), \dots, Z_i(t_{in_i}))^T$ ,  $B_i = (B_q(t_{i1}), \dots, B_q(t_{in_i}))^T$  and  $A_i = (\bar{B}_i, Z_i^*)$ . One can modify the iteratively weighted least squares for estimating  $\beta$  and  $\alpha$  as follows,

$$\begin{pmatrix} \hat{\beta}^{(l)} \\ \hat{\alpha}^{(l)} \end{pmatrix} = \left[ \sum_{i=1}^n A_i^T E_i^{(l-1)} (W_i) A_i \right]^{-1} \left[ \sum_{i=1}^n A_i^T E_i^{(l-1)} \{W_i (d_i - B_i \hat{\Theta} \xi_i)\} \right]. \quad (22)$$

Given the eigen-decomposition of the FPC model (10), one has to estimate an unknown matrix  $\Theta = (\theta_1, \dots, \theta_K)$ . Estimating  $\Theta$  involves an inside iterative procedure, where each column  $\theta_k$  is estimated separately holding all other columns fixed (James *et al.*, 2001). Applying iteratively weighted least squares, one can estimate  $\theta_k$  by

$$\hat{\theta}_k = \left\{ \sum_{i=1}^n B_i^T E_i^{(l-1)} (\xi_{ik}^2 W_i) B_i \right\}^{-1} \left[ \sum_{i=1}^n B_i^T E_i^{(l-1)} \left\{ \xi_{ik} W_i (d_i - \bar{B}_i \hat{\beta} - Z_i^* \hat{\alpha} - \sum_{\ell \neq k} \xi_{i\ell} B_i \hat{\theta}_\ell) \right\} \right]. \quad (23)$$

This procedure is repeated for each column of  $\Theta$  and iterate until there is no further change in  $\hat{\Theta}$ . Due to the orthonormal constraints (9), letting  $\hat{\Gamma} = \hat{\Theta} \hat{\Lambda} \hat{\Theta}^T$  and setting the final estimate  $\hat{\Theta}$  equal to the first  $K$  eigenvectors of  $\hat{\Gamma}$ , the estimate  $\hat{\Lambda}$  becomes the diagonal matrix consisting of the first  $K$  eigenvalues of  $\hat{\Gamma}$ .

From our experience the inner iterative procedure for estimating  $\beta$ ,  $\alpha$ ,  $\Theta$  and  $\eta$  usually converges very fast. The computation time required for the whole algorithm is mainly determined by the dimension of random effects  $u_i$  or FPC scores  $\xi_i$ , and the Monte Carlo sample sizes that are used to approximate conditional expectations in E-step.

## References

- Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., C. D. J., Haris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J. H., Thompson, M., Deyton, L. and the Terry Beinr Comminuty Programs for Clinical Research on AIDS (1994) Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant or have failed zidovudine therapy. *New England Journal of Medicine*, **330**, 657–662.
- Berkey, C. S. and Kent, R. L. J. (1983) Longitudinal principal components and non-linear regression models of early childhood growth. *Annals of Human Biology*, **10**, 523–536.
- Besse, P. and Ramsay, J. O. (1986) Principal components analysis of sampled functions. *Psychometrika*, **51**, 285–311.
- Booth, J. G. and Hobert, J. P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 265–285.
- Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.
- Breslow, N. E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
- Bycott, P. and Taylor, J. (1998) A comparison of smoothing techniques for CD4 data measured with error in a time-dependent cox proportional hazard model. *Statistics in Medicine*, **17**, 2061–2077.
- Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986) Principal modes of variation for processes with continuous sample curves. *Technometrics*, **28**, 329–337.
- Cox, D. R. (1972) Regression models and lifetables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–200.
- (1975) Partial likelihood. *Biometrika*, **62**, 269–276.
- Dafni, U. G. and Tsiatis, A. A. (1998) Evaluating surrogate markers if clinical outcomes measured with error. *Biometrics*, **54**, 1445–1462.
- Diggle, P., Hergerty, P., Liang, K. Y. and Zeger, S. (2002) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

- Faucett, C. L., Schenker, N. and Elashoff, R. M. (1998) Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association*, **93**, 427–437.
- Faucett, C. L. and Thomas, D. C. (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.
- Goldman, A. I., Carlin, B. P., Crane, L. R., Launer, C., K. J. A., Deyton, L. and Abrams, D. I. (1996) Response of CD4+ and clinical consequences to treatment using ddI in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **11**, 161–169.
- Goldstein, H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, **78**, 45–51.
- Guo, X. and Carlin, B. P. (2004) Separate and joint modelling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 1–9.
- Henderson, R., Diggle, P. J. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **4**, 465–480.
- James, G., Hastie, T. G. and Sugar, C. A. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.
- James, L. F., Priebe, C. E. and Marchette, D. J. (2001) Consistent estimation of mixture complexity. *The Annals of Statistics*, **29**, 1281–1296.
- Larsen, K. (2005) The Cox proportional hazards model with a continuous latent variable measured by binary indicators. *Biometrics*, **61**, 1049–1055.
- Lin, X. and Breslow, N. E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. London: Chapman and Hall, 2nd edn.
- McCulloch, C. E. (1994) Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330–335.
- (1997) Maximum likelihood algorithm for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

- Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, **84**, 33–44.
- Prentice, R. (1982) Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika*, **69**, 331–342.
- Raboud, J., Reid, N., Coates, R. A. and Farewell, V. T. (1993) Estimating risks of progressing to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society, Series A*, **156**, 396–406.
- Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer.
- (2002) *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.
- Rice, J. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53**, 233–243.
- Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.
- Silverman, B. W. (1996) Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**, 1–24.
- Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994) A stochastic model for analysis of longitudinal data. *Journal of the American Statistical Association*, **89**, 727–776.
- Taylor, J. M. G., Tan, S. J., Detels, R. and Giorgi, J. V. (1991) Applications of computer simulation model of the natural history of CD4 T-cell number in HIV-infected individuals. *AIDS*, **5**, 159–167.
- Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.
- Tsiatis, A. A., Degruittola, 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 modelling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, **96**, 895–905.
- Wolfinger, R. and O’Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

- Wulfsohn, M. S. and Tsiatis, A. A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.
- Xu, J. and Zeger, S. L. (2001) Joint analysis of longitudinal data comprising repeated measures and times to event. *Applied Statistics*, **50**, 375–387.
- Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., L., Y., Buchholz, B. A. and Vogel, J. S. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, **59**, 676–685.
- Yao, F., Müller, H. G. and Wang, J. L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.
- Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Table 1: Simulation results obtained from 400 normal and 400 mixture simulated datasets for the estimation of  $\gamma$  with the true value ( $-1$ ) by the “ideal” approach where  $X_i(t)$  is known for all  $t \in [0, 10]$  (IDEAL), last-value-carried-forward (LVCF), and the proposed joint model using FPCs (JPFC), see Section 4 for details. Shown are Monte Carlo averages of estimates and standard errors (in parentheses) for two choices of eigenvalues  $\lambda_1 = 4$  and  $\lambda_1 = 9$ .

	Normal	Mixture	Normal	Mixture
	$\lambda_1 = 4$	$\lambda_1 = 4$	$\lambda_1 = 9$	$\lambda_1 = 9$
IDEAL	0.979 (0.116)	0.963 (0.121)	0.971 (0.119)	1.044 (0.115)
LVCF	0.375 (0.121)	0.349 (0.113)	0.357 (0.122)	0.384 (0.124)
JPFC	1.049 (0.118)	0.956 (0.117)	1.063 (0.126)	0.969 (0.123)

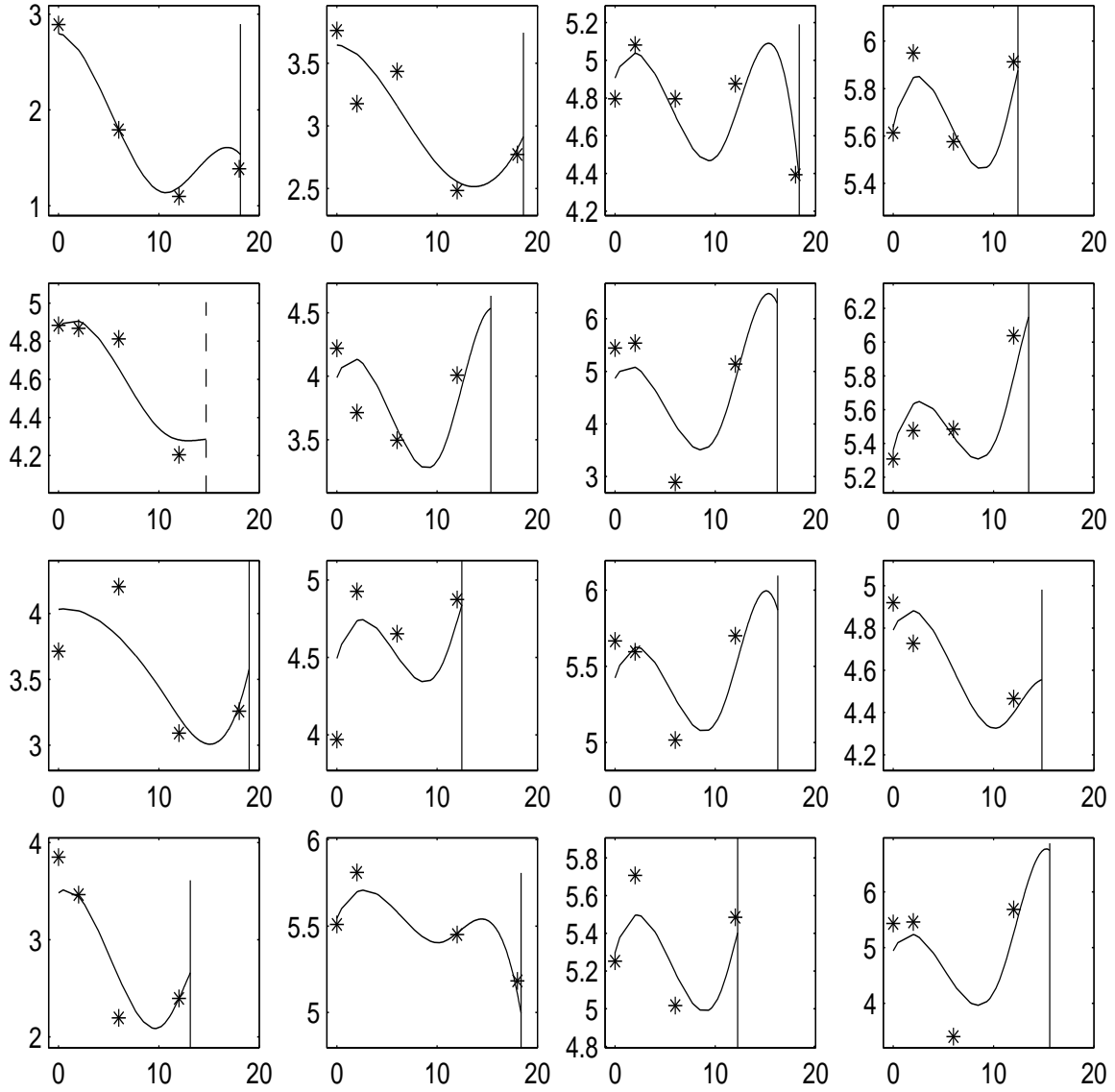


Figure 1: Observed (asterisks) CD4 counts on log scale and fitted trajectories obtained from the proposed joint Poisson FPC model for 8 randomly selected patients from the ddC group (two top rows) and 8 patients from the ddI group (two bottom rows). The vertical lines represent the censoring (solid) or event (dashed) time.

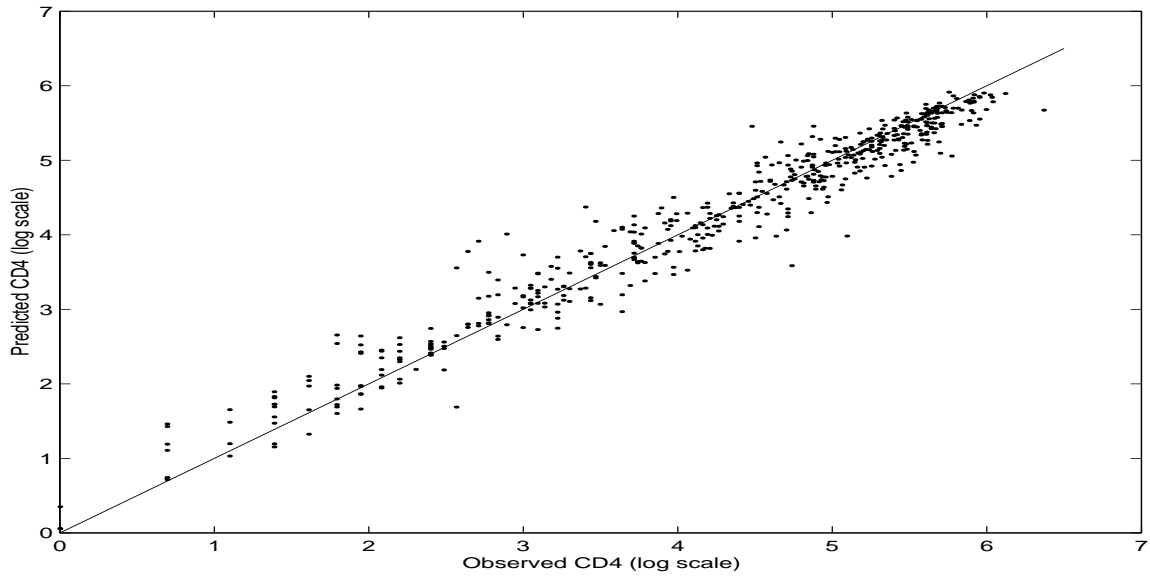


Figure 2: Fitted values obtained from the proposed joint Poisson FPC model versus the observed CD4 counts on log scale, along with the  $45^\circ$  straight line.

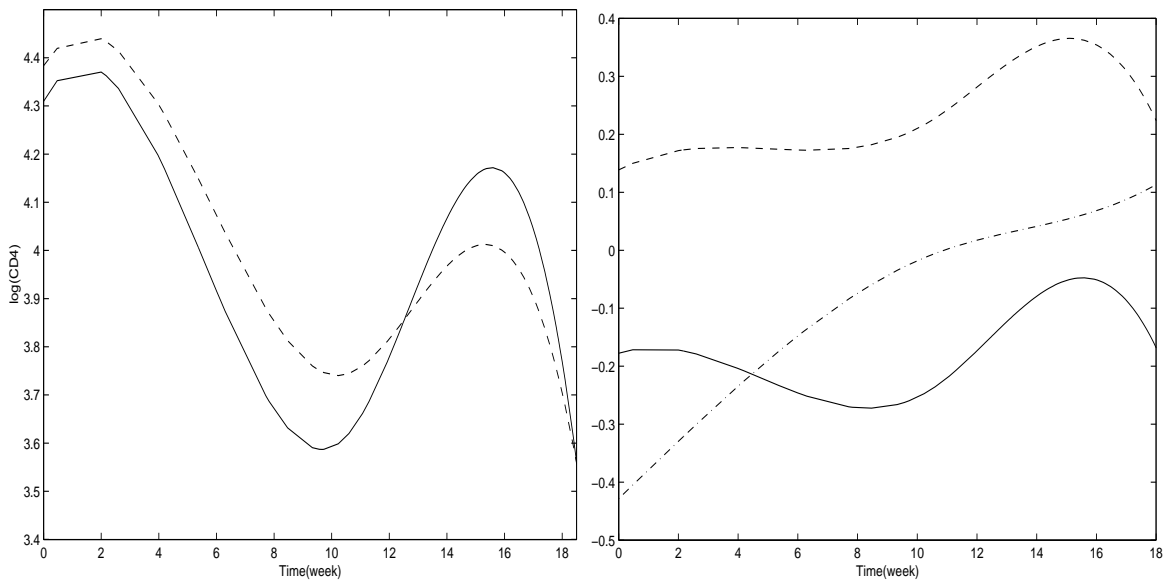


Figure 3: Left: Smooth estimates of the mean functions for the ddC (solid) group and the ddI (dashed) group. Right: Smooth estimates of the first (solid), second (dashed) and third (dash-dotted) eigenfunctions.

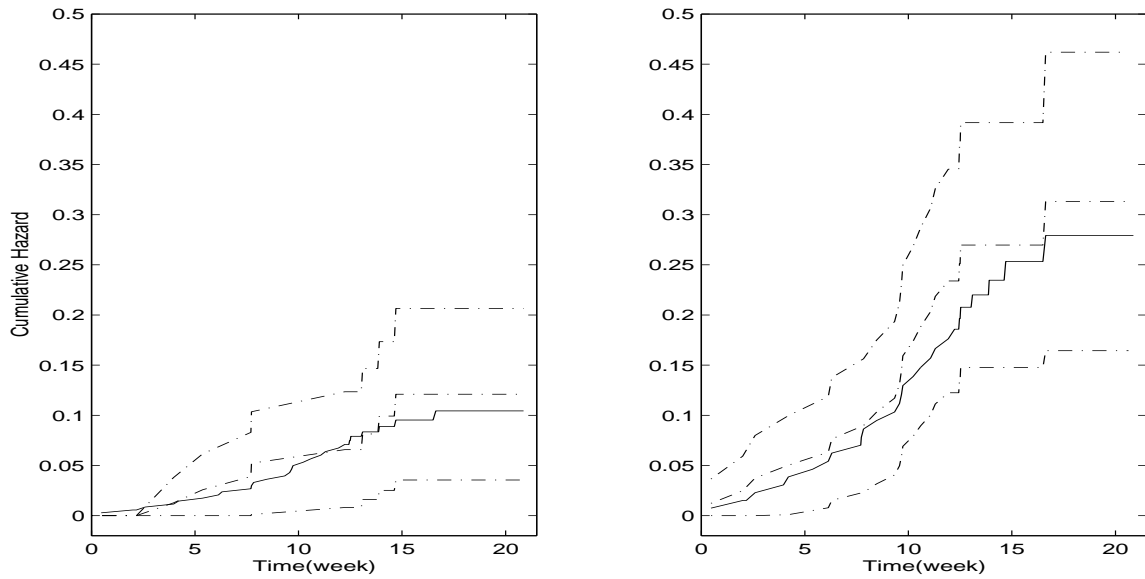


Figure 4: Estimated average cumulative hazards (solid lines) as in (16) obtained from proposed joint Poisson FPC model for the ddC group (left panel) and the ddI group (right panel), compared with empirical cumulative hazards (middle dash-dotted lines) obtained from Nelson-Aalen estimates as well as corresponding 95% empirical confidence bands (lower and upper dash-dotted lines).