Statistica Sinica **17**(2007), 965-983

FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS FOR LONGITUDINAL AND SURVIVAL DATA

Fang Yao

University of Toronto

Abstract: This paper proposes a nonparametric approach for jointly modelling longitudinal and survival data using functional principal components. The proposed model is data-adaptive in the sense that it does not require pre-specified functional forms for longitudinal trajectories and it automatically detects characteristic patterns. The longitudinal trajectories observed with measurement error are represented by flexible basis functions, such as B-splines, and the model dimension is reduced by functional principal component analysis. The relationship between the longitudinal process and event history is assessed using Cox regression model. Although the proposed model inherits the flexibility of a nonparametric approach, the estimation procedure based on EM algorithm is intrinsically parametric, and thus is simple and easy to implement. The computation is more efficient by reducing the dimension of random coefficients, i.e., functional principal component scores. The reduction of dimension achieved from eigen-decomposition also makes the model particularly applicable for the sparse data often encountered in longitudinal studies. An iterative selection procedure based on the Akaike Information Criterion (AIC) is proposed to choose tuning parameters, such as the knots of spline basis and the number of principal components, so that an appropriate degree of smoothness can be assessed. The effectiveness of the proposed approach is illustrated through a simulation study, followed by an application to longitudinal CD4 counts and survival data collected in a clinical trial for comparing the efficacy and safety of two antiretroviral drugs.

Key words and phrases: Cox regression, EM algorithm, functional principal components, longitudinal data, smoothing, survival.

1. Introduction

Many scientific investigations generate longitudinal data with repeated measurements at a number of time points, and event history data that are possibly censored time-to-event, i.e.,"failure" or "survival", as well as additional covariate information. A typical example is that of HIV clinical trials, in which a biomarker such as CD4 lymphocyte count is measured intermittently and time to progression to AIDS or death is also recorded, with possible early dropout or failure to experience event by the end of study. It is important and necessary

to investigate the patterns of CD4 changes, and to characterize the relationship between CD4 features and time to progression or death (Pawitan and Self (1993), Tsiatis, Degruttola and Wulfsohn (1995), Wulfsohn and Tsiatis (1997), and others).

In practice latent longitudinal process is often unobservable due to measurement error and not available at necessary times, especially when failure occurs. It is well known that conventional partial likelihood approaches used for Cox model cannot avoid biased inference by using 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 (Bycott and Taylor (1998), Tsiatis et al. (1995)). This invoked the consideration of longitudinal and event processes simultaneously, i.e., the "so-called" joint modelling, that has attracted substantial recent interest. A standard approach of joint modelling is to characterize the longitudinal process by a parametric random effects model that focuses on smooth trends determined by a small number of random effects and that has been used to describe the "true" CD4 trajectories (Tsiatis et al. (1995), Wulfsohn and Tsiatis (1997), Bycott and Taylor (1998), Dafni and Tsiatis (1998)). Besides bias correction, joint modelling can also potentially improve the efficiency of parameter estimation because of simultaneous inference on both longitudinal and survival models, see Faucett and Thomas (1996); Slasor and Laird (2003), Hsieh, Tseng and Wang (2006) for more discussion on this issue.

Although the above-mentioned parametric models find features in the data which have been already incorporated *a priori* in the model, these models may not be adequate if the time courses are not well defined and do not fall into a preconceived class of functions. In such situations an analysis through nonparametric methods is advisable. There has been increasing interest in the nonparametric analysis of data that are in the form of sample of curves or trajectories, i.e., "functional data analysis", see Ramsay and Silverman (1997) for a summary. Functional principal component analysis (FPCA) attempts to find the dominant modes of variation around overall trend functions, and is thus a key technique in functional data analysis (Berkey and Kent (1983), Besse and Ramsay (1986), Castro, Lawton and Sylvestre (1986), Rice and Silverman (1991), Silverman (1996), James, Hastie and Sugar (2000), Yao et al. (2003, 2005), Yao and Lee (2006), and many others).

The proposed approach is motivated by an analysis of longitudinal CD4 counts and survival data collected in a clinical trial for comparing the efficacy and safety of two antiretroviral drugs. From the data plotted in Figure 1 the time courses are not well defined, and one would be reluctant to fit a pre-specified parametric model such as a linear random effects model to characterize the longitudinal CD4 trajectories. In this paper, a nonparametric approach is proposed to

model the individual trajectories using flexible basis functions, such as B-splines, with the covariance structure modelled by a set of orthogonal eigenfunctions and random coefficients that are called functional principal component (FPC) scores. This produces a low rank, low frequency approximation to the covariance structure of a longitudinal process. We can also adjust the tuning parameters in the model, such as the knots of the spline functions and the number of principal components, to capture some important "wiggles" or fluctuations. The model is data-adaptive and will automatically capture important features of individual trajectories. Moreover, a typical feature of longitudinal data often encountered in practice is that only a few repeated and irregularly spaced measurements are available per subject. The proposed model is particularly applicable for handling such irregular and sparse longitudinal data through dimension reduction, using FPC analysis. This inherits the merits of the reduced rank model proposed by James, Hastie and Sugar (2000) that does not treat censoring information.

In contrast, a closely related model proposed by Rice and Wu (2000) did not consider dimension reduction and might not be applicable when data are sparse (see James et al. (2000) for a comparison of these two approaches). This makes the difference between the proposed model and that in Brown et al. (2005) explicit. Another advantage of the proposed joint model with FPCs is the computational efficiency achieved by the dimension reduction using FPCs with a diagonal covariance matrix, while the joint model in Brown et al. (2005) with B-splines usually contains more random coefficients with an unstructured covariance matrix. Appropriate interpretation of the orthogonal eigenfunctions and FPC scores often provides more insight than the B-spline model. Alternatively Wang and Taylor (2001) incorporated an integrated Ornstein-Uhlenbeck (IOU) stochastic process to model non-specified longitudinal trajectories in a joint model context, in similar spirit to smoothing splines. In particular, the IOU process presents a family of covariance structures with a random effects model and Brownian motion as special cases. Other related work that incorporated zero-mean processes to model individual fluctuations includes Henderson, Diggle and Dobson (2000), Xu and Zeger (2001b).

For model selection, we suggest an iterative procedure to choose the tuning parameters simultaneously, such as the number of knots and the number of principal components, using the Akaike Information Criterion (AIC) that considers the joint likelihood of longitudinal and survival models. Another advantage of the proposed method is that, although it is a data-driven approach with a nonparametric feature, the implementation is intrinsically parametric, and thus is computationally efficient and easily implemented in standard statistical software packages. Extensions of the proposed model to multiple longitudinal processes or a sequence of survival times (recurrent events) are also possible, see Xu and Zeger (2001a), Song et al. (2002a) and Henderson et al. (2000) for modifications.

The remainder of the paper is organized as follow. In Section 2 we present the general methodology of the proposed joint model, including functional principal component and survival models for longitudinal and event processes. Simulation results that illustrate the effectiveness of the methodology are reported in Section 3. An application of the proposed model to longitudinal CD4 counts and survival data collected in a clinical trial to compare the efficacy and safety of two antiretroviral drugs, is provided in Section 4. Technical details are deferred to the Appendix.

2. Methodology

Assumptions and notations

For a sample of subjects, one observes longitudinal covariates and time-to-event data. Without loss of generality, we assume a single longitudinal process $X_i(t)$ for the ith individual, $i = 1, \ldots, n$. The survival time S_i is subject to independent right censoring by C_i, then one observes $T_i = \min(S_i, C_i)$ and the failure indicator Δ_i which is 1 if failure is observed, $S_i \leq C_i$, and is 0 otherwise. The observed covariate $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ for the ith individual is often assumed to be sampled from the latent longitudinal process $X_i(t)$, measured intermittently at $t_i = (t_{i1}, \ldots, t_{in_i})^T$, and terminated at the endpoint T_i, subject to measurement error $\epsilon_i = (e_{i1}, \ldots, e_{in_i})^T$. Note that $t_{in_i} \leq T_i \leq t_{in_i+1}$, and no longitudinal measurements are available after T_i. Then one has $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, where $t_{ij} \in [0, T_i]$, $\max\{T_i : i = 1 \ldots, n\} \leq \tau$, and τ is the length of study follow-up.

Besides the longitudinal and time-to-event data, there might be other covariates that possibly have significant effects on longitudinal or survival processes. Let $Z_i(t) = \{Z_{i1}(t), \ldots, Z_{ir}(t)\}^T$ denote the vector of covariates valued at time $t \leq T_i$ associated with the longitudinal covariate, and $V_i(t) = \{V_{i1}(t), \ldots, V_{is}(t)\}^T$ the vector of covariates at $t \leq T_i$ having effects on the survival time. 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 ith subject can be exactly observed at any time $t \leq T_i$ and, in particular, the $V_i(T_i)$'s are available for all subjects. In contrast, the longitudinal covariate Y_i is assumed to be subject to measurement error and only available at t_i. Denote the $n_i \times r$ design matrix formed by the covariate $Z_i(t)$ at t_i by $Z_i = \{Z_i(t_{i1}), \ldots, Z_i(t_{in_i})\}^T$.

To validate the specification of our proposed method, the assumed independent right censoring is in the sense of Kalbfleisch and Prentice (2002): the hazard at time t conditional on the whole history only depends on the survival of that individual to time t. Most analyses, including those based on likelihoods, are valid under this assumption. One also needs to require 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, see Tsiatis and Davidian (2004) for detail. 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 and covariates $Z_i(t)$ and $V_i(t)$, and the data from different subjects are generated by independent processes. We suppose that the longitudinal process has significant influence on the failure time, and that the occurrence of time-to-event may also introduce informative censoring for the longitudinal process, which is of primary interest in joint modelling problems.

2.2. Joint modelling and implementation for longitudinal and survival data using functional principal components

We extend the principal component model proposed by James et al. (2000) to joint modelling approaches. Recall that $X_i(t)$ is the realization of the latent longitudinal process for the ith individual, let $\mu(t)$ be the overall mean function without considering the vector of covariates $Z_i(t)$. If the effects of $Z_i(t)$ are also taken into account, then

$$\mu_i(t) = \mu(t|Z_i) = \mu(t) + \beta^T Z_i(t), \qquad t \in [0, \tau],$$

where $\beta = (\beta_1, \ldots, \beta_r)^T$ and $Z_i(t) = \{Z_{i1}(t), \ldots, Z_{ir}(t)\}^T$. The covariance structure of $X_i(t)$ might also depend on the components of $Z_i(t)$, e.g., if $Z_i(t)$ contains a treatment indicator. For convenience, we suppose that the covariance structure of $Z_i(t)$ is the same covariance for all i, with $G(s,t) = \mathrm{cov}(X_i(s), X_i(t))$. In terms of orthogonal eigenfunctions $\{\phi_k\}_{k=1,2,\ldots}$ and non-increasing eigenvalues $\{\lambda_k\}_{k=1,2,\ldots}$, let $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$, $s, t \in [0,\tau]$. The Karhunen-Loéve representation in classical functional principal component analysis implies that the individual trajectories can be expressed as $X_i(t) = \mu_i(t) + \sum_k \xi_{ik}\phi_k(t)$, where $\mu_i(t)$ is the mean function for the ith subject, the coefficients $\xi_{ik} = \int_0 \{X_i(t) - \mu_i(t)\}\phi_k(t)dt$ are uncorrelated random variables with mean zero and variances $E\xi_{ik}^2 = \lambda_k$, subject to $\sum_k \lambda_k < \infty$.

Since interest is in the relationship between the dominant trends of the longitudinal process and event times, we suppose that the covariance function G can be well-approximated by the first few terms in the eigen-decomposition, i.e., that the eigenvalues λ_k tend to zero rapidly enough that the variability is predominantly of large scale and low frequency, and that the individual trajectories can be modelled by using the first K leading principal components, $X_i(t) = \mu_i(t) + \sum_{k=1} \xi_{ik}\phi_k(t)$. The truncation parameter K can be adjusted. To realistically model the noisy observations Y_{ij} of $X_i(t)$ at time t_{ij}, we incorporate

uncorrelated measurement errors ϵ_{ij} which have mean zero, constant variance σ^2, and are independent of ξ_{ik}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, $k = 1, 2, \ldots$. Then the sub-model for the longitudinal covariate Y_{ij} is

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + Z_i(t_{ij})^T\beta + \sum_{k=1} \xi_{ik}\phi_k(t_{ij}) + \epsilon_{ij}, \quad t \square [0, \tau]. \quad (1)$$

The overall mean function and covariance surface, and thus the eigenfunctions, are assumed to be smooth, we model them using expansions of a set of smooth basis functions, such as B-splines or regression splines. Let $\bar{B}_p(t) = (\bar{B}_{p1}(t), \ldots, \bar{B}_{pp}(t))^T$ be a set of basis functions on $[0, \tau]$ used to model the overall mean function $\mu(t)$, with coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)^T$, i.e., $\mu(t) = \bar{B}_p(t)^T\alpha$. Due to the orthonormality of $\{\phi_k\}_{k=1}$, the eigenfunctions are modelled by using a set of orthonormal basis functions $B_q(t) = (B_{q1}(t), \ldots, B_{qq}(t))^T$ with coefficients $\theta_k = (\theta_{1k}, \ldots, \theta_{qk})^T$ that are subject to

$$\int_0 B_q\ (t)B_q\ (t)dt = \delta\ \ , \quad \theta_k^T\theta\ = \delta_k\ , \quad \kappa, \ell = 1, \ldots, q, \quad k, l = 1, \ldots, K, \quad (2)$$

which implies the orthonormal constraints on $\{\phi_k\}_{k=1}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Note that $\theta_k^T\theta\ = \delta_k$ in (2) resolves the identifiability between Θ and ξ_i, see the Appendix for detail that guarantees $\theta_k^T\theta\ = \delta_k$, while the basis functions $B_q(t)$ for the covariance are orthonormalized.

Letting $\xi_i = (\xi_{i1}, \ldots, \xi_i\ \ )^T$ and $\Theta = (\theta_1, \ldots, \theta\ \ )^T$, (1) becomes

$$Y_{ij} = \bar{B}_p(t_{ij})^T\alpha + Z_i(t_{ij})^T\beta + B_q(t_{ij})^T\Theta\xi_i + \epsilon_{ij}. \quad (3)$$

The covariance between observed values of the longitudinal process is, letting $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda\ \ )$,

$$\mathrm{cov}(Y_{ij}, Y_i\ ) = B_q(t_{ij})^T\Theta\Lambda\Theta^T B_q(t_i\ ) + \sigma^2\delta_j\ . \quad (4)$$

This can be viewed as an approximation, where low frequency components of the covariance kernel are captured in the first term and the remainder is approximated by the second term. Let $\bar{B}_i = (\bar{B}_p(t_{i1}), \ldots, \bar{B}_p(t_{in_i}))^T$, $B_i = (B_q(t_{i1}), \ldots, B_q(t_{in_i}))^T$, and recall that $Z_i = (Z_i(t_{i1}), \ldots, Z_i(t_{in_i}))^T$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$. The model (3) can be written in matrix from as

$$Y_i = \bar{B}_i\alpha + Z_i\beta + B_i\Theta\xi_i + \epsilon_i, \qquad i = 1, \ldots, n, \quad (5)$$

under constraints given by (2).

Note that the dominant modes of variation of a longitudinal process can often be captured by the first few principal components in practical applications,

i.e., the trajectories can be well-approximated by truncated eigen-decomposition with a few eigenfunctions. Therefore the dimension of random coefficients (FPC scores), K, is usually small in the forgoing model, leading to fast and efficient implementation. This is the difference between (5) and the B-spline model of Brown et al. (2005).

We model failure through the proportional hazards model. Following Cox (1972), Cox (1975), and under the conditions discussed by Kalbfleisch and Prentice (2002), we use the original Cox model formulation, 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. The framework for characterizing associations among the longitudinal and survival processes, as well as other covariates, is then given by

$$\begin{aligned} h_i(t) &= \lim_{dt\to 0} P\{t \le T_i < t + dt | T_i \ge t, X_i^H(t), V_i(t)\}/dt \\ &= h_0(t)\exp\{\gamma X_i(t) + V_i(t)^T\zeta\}, \end{aligned} \tag{6}$$

where the coefficients γ and $\zeta = (\zeta_1, \ldots, \zeta_s)^T$ reflect the association of interest, and $X_i^H(t) = \{X_i(u) : 0 \le u < t\}$ is the history of the longitudinal process X_i up to time t. One notes that implementation is complicated by the fact that the longitudinal covariate Y_i is subject to measurement error, and only available intermittently for each subject at $t_i = (t_{i1}, \ldots, t_{in_i})^T$.

We now model the longitudinal covariate and survival processes jointly. Note that observed values Y_{ij} of the longitudinal process and the failure information (T_i, D_i) are conditionally independent given the latent process $X_i(t)$ and covariates $Z_i(t)$ and $V_i(t)$. The observed data for each individual is denoted by $O_i = (T_i, D_i, Y_i, Z_i, V_i, t_i)$, and we cannot observe the latent longitudinal process $X_i(t)$ or the FPC scores ξ_i. Let $\tilde{X}_i = \{X_i(t_{i1}), \ldots, X_i(t_{in_i})\}^T$. One can see that the random components of trajectories $X_i(t)$ are determined by ξ_i. Therefore the observed data likelihood for the full set of parameters of interest $\Omega = \{\gamma, \zeta, h_0(\cdot), \alpha, \beta, \Theta, \Lambda, \sigma^2\}$ is given by

$$L_o = \prod_{i=1}^{n} \left\{ \int f(T_i, \Delta_i | X_i^H(T_i), V_i(T_i), \gamma, \zeta, h_0) f(Y_i | \tilde{X}_i, \sigma^2, t_i) f(\xi_i | \Lambda) d\xi_i \right\}, \tag{7}$$

where $X_i^H(T_i) = \{X_i(t) : 0 \le t < T_i\}$, and

$$\begin{aligned} f(T_i, \Delta_i | X_i^H(T_i), V_i(T_i), \gamma, \zeta, h_0) &= [h_0(T_i)\exp\{\gamma X_i(T_i) + V_i(T_i)^T\zeta\}]^{\Delta_i} \\ &\quad \times \exp\Big[- \int_0^{T_i} h_0(u)\exp\{\gamma X_i(u) + V_i(u)^T\zeta\} du \Big]. \end{aligned} \tag{8}$$

For the distributions of the FPC scores ξ_i and the measurement error ϵ_i, assume

$$\xi_i \sim N(0, \Lambda), \quad \epsilon_i \sim N(0_{n_i}, \sigma^2 I_{n_i}), \quad \xi_i \perp \epsilon_i, \qquad i = 1, \ldots, n, \tag{9}$$

where "$\perp$" stands for statistical independence and 0 is a $K \times 1$ vector of 0's. The normality assumption (9) can be relaxed, assuming only that the ξ_i have "smooth" densities in certain well-defined class for instance, see Song, Davidian and Tsiatis (2002b). However, from simulation studies reported in Section 3 and also in Tsiatis and Davidian (2004), the procedure resulting from the Gaussian assumption works well more generally. The "robustness" of the Gaussian assumption has also been observed in Yao et al. (2005), and a theoretical justification was recently given by Hsieh et al. (2006). For convenience we use (9) in the joint likelihood. Then the densities in (7) are given by

$$f(Y_i|\tilde{X}_i, \sigma^2, t_i) = (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp\{-\frac{1}{2\sigma^2}(Y_i - \tilde{X}_i)^T(Y_i - \tilde{X}_i)\}, \tag{10}$$

$$f(\xi_i|\Lambda) = (2\pi|\Lambda|)^{-\frac{1}{2}} \exp(-\frac{1}{2}\xi_i^T \Lambda^{-1} \xi_i), \tag{11}$$

where $\tilde{X}_i = \bar{B}_i\alpha + Z_i\beta + B_i\xi_i$, and $|\Lambda|$ denotes the determinant of $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{\quad})$.

The EM algorithm described by Wulfsohn and Tsiatis (1997) can be easily extended to the proposed model. The computation time of the EM algorithm is determined mostly by the dimension of random coefficients that are treated as missing values, since the numerical approximations to the expectations of functions of these random coefficients are needed in the E-step. When the Gaussian-Hermite quadrature becomes time-consuming in large dimensions, an alternative method is to approximate the numerical integrals by a Monte Carlo integration with antithetic simulation for variance reduction, see Henderson et al. (2000) for more details. We implement our approach using Monte Carlo integration, rather than Gaussian-Hermite quadrature, within the EM algorithm. Details are in the Appendix.

2.3. Iterative selection procedure for choosing tuning parameters

Due to the nonparametric feature of the proposed joint model involving smoothing and functional principal components, we need to choose tuning parameters in the model so that the association between appropriate degrees of the smooth trend of the longitudinal process and survival times can be assessed.

In practical implementation, spline bases, such as B-spline or regression splines, is often used to model the mean and covariance functions. For such a spline basis the degree of smoothness is determined by the sequence of knots (number and locations). For the locations of the knots, we choose equi-quantile

knots rather than equidistant knots to avoid clustered observation times, particularly when observations become fewer due to failure or censoring. For example, 25th, 50th and 75th percentiles of pooled observation times from all individuals are used if three inside knots are needed. It is obvious that choosing the number of knots is equivalent to choosing the dimensions of spline basis, i.e., p and q, for fixed and random components in model (3).

Besides the choice of knots, for functional principal component analysis it is particularly important to identify the number of terms needed in the approximation to 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 behavior of the model. We need a selection procedure to choose K ($K \leq q$), p and q together. Note that the estimation procedure of the proposed model is intrinsically parametric, though the FPC model inherits the flexibility of the nonparametric approach. We thus propose to adapt the Akaike information criterion (AIC) to the joint model. By analogy to Brown et al. (2005), a pseudo-Gaussian joint log-likelihood depending on K, p and q, summing the contributions from all subjects, conditional on the estimated FPC scores $\hat{\xi}_i$, is

$$\widehat{l}_c(K,p,q) = \sum_{i=1}^{n} \Big[\log\{f(T_i, \Delta_i | \widehat{X}_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0)\} + \log\{f(Y_i | \widehat{\widetilde{X}}_i, \hat{\sigma}^2, t_i)\}\Big], \tag{12}$$

where the densities of (T_i, Δ_i) and Y_i are as in (8) and (10), " $\hat{}$ " is the generic notation for the estimates obtained from the EM algorithm, $\widehat{X}_i^H(T_i) = \{\widehat{X}_i(t) : 0 \leq t \leq T_i\}$, $\widehat{X}_i(t) = \bar{B}_p(t)^T\hat{\alpha} + Z_i(t)^T\hat{\beta} + B_q(t)^T\widehat{\Theta}\hat{\xi}_i$, and $\widehat{\widetilde{X}}_i = (\widehat{X}_i(t_{i1}), \ldots, \widehat{X}_i(t_{in_i}))^T$. Then the AIC of the model involving K, p and q is

$$AIC(K,p,q) = -2\widehat{l}_c(K,p,q) + 2\{p + (K+1)q + r + s + 1\}. \tag{13}$$

Minimization of AIC with respect to K, p and q requires extensive computation. Alternatively, we start with initial guesses for p and q, choose K using (13), choose p and q in turn based on (13), and then repeat until there is no further change. It has been observed that this iterative procedure usually converges quickly (in two or three iterations) and is practically feasible. This is demonstrated empirically in Section 3.

3. Simulation Studies

From extensive simulations in the literature, it has been found that the joint modelling approach improves parameter estimation over "naive" or "two-stage" approaches (Prentice (1982), Self and Pawitan (1992), Tsiatis et al. (1995),

Song et al. (2002b), Tsiatis and Davidian (2004), and others), while Henderson et al. (2000) examined the effect of ignoring latent association between longitudinal and survival processes. However, these simulation studies were constructed based on using parametric models, such as linear models, during the estimation procedure. It is not clear how the joint modeling approaches behave when the underlying parametric forms are not used in the estimation. Our simulation is designed to demonstrate the empirical performance of the proposed joint model that does not require knowledge of the underlying longitudinal sub-model.

We assumed sufficient assumptions on censoring and timing of measurements were satisfied. A comparison was provided to the parametric joint model with the true longitudinal sub-model incorporated. Both 200 normal and 200 non-normal samples consisting of $n = 200$ independently and identically distributed individuals were considered, to demonstrate the robustness of the procedure regarding the sensitivity to the Gaussian assumption (9). For simplicity, we took $\eta = 0$, $\gamma = -1.0$ in the survival model (6), with Weibull baseline $h_0(t) = t^2/100$ for $t \geq 0$. Censoring times C_i were generated independently of all other variables as Weibull random variables with 20% dropouts at $t = 6$ and 70% dropouts at $t = 9$, and a final truncation time of $\tau = 10$ was used.

The longitudinal process had mean function $\mu(t) = \sin(3\pi t/40)/3$ with $\beta = 0_r$ in model (1), and a covariance function derived from two eigenfunctions $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, and $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, $0 \leq t \leq 10$. We chose $\lambda_1 = 10$, $\lambda_2 = 1$ and $\lambda_k = 0$, $k \geq 3$, as eigenvalues, and $\sigma^2 = 0.1$ as variance of the additional measurement errors ϵ_{ij} in (1), assumed to be normal with mean 0. For the 200 normal samples, the FPC scores ξ_{ik} were generated from $\mathbf{N}(0, \lambda_k)$, while the ξ_{ik} for the non-normal samples were generated from a mixture of two normals, $\mathbf{N}(\sqrt{\lambda_k/2}, \lambda_k/2)$ with probability 1/2 and $\mathbf{N}(-\sqrt{\lambda_k/2}, \lambda_k/2)$ with probability 1/2. For an equally spaced grid $\{c_1, \ldots, c_9\}$ on $[0, 10]$ with $c_i = i$, $s_i = c_i + e_i$, where e_i were i.i.d. with $\mathbf{N}(0, 0.1)$, $s_i = 0$ if $s_i < 0$ and $s_i = 10$ if $s_i > 10$, allowing for non-equidistant "jittered" designs. Measurements at any of these times were missing with probability 0.5 and would be terminated by the observed event time T_i, i.e., $t_{in_i} \leq T_i$, demonstrating sparseness of the simulated data, as the expected number of available observations would be seriously downward-biased by failure or censoring.

For each normal and mixture sample, γ was estimated in three ways: (i) using the joint model with the true functional relation for the longitudinal sub-model, denoted by TRUE; (ii) using the the proposed joint model with functional principal component sub-model, denoted by FPCA, where the number of eigenfunction K and the dimensions p and q were chosen objectively at (13); (iii) using the "ideal" approach, where $X_i(t)$ is known for all $0 \leq t \leq 10$, and γ was estimated by conventional partial likelihood approach (Cox (1975)), denoted by IDEAL.

One can see that the results summarized in Table 1 are consistent with the results obtained from previous work, such as in Tsiatis and Davidian (2004). If the true functional relationship is specified in the longitudinal sub-model (TRUE) during estimation, then unbiased inference was achieved. More importantly, the proposed model (FPCA) that did not incorporate any prior knowledge of the true relation also yields approximately unbiased estimate in both normal and mixture scenarios. In particular, these results are comparable with those obtained from the TRUE case where the underlying form is used in the longitudinal model and the IDEAL case where the true trajectories $X_i(t)$ are used in estimation. This suggests that, without the knowledge of the true longitudinal process, the proposed model can automatically detect the underlying relationship and provide satisfactory approximation to the true functional form due to its nonparametric flexibility. It is notable that the Gaussian assumption does not compromise accuracy of the estimation of γ under the mixture scenario. This is similar to what was found in Tsiatis and Davidian (2004), while a theoretical justification of the robustness phenomenon was recently given by Hsieh et al. (2006).

Table 1. Simulation results obtained from 200 normal and 200 mixture Monte Carlo datasets for the estimation of $\gamma = -1.0$ by the joint model, incorporating the true (TRUE) longitudinal sub-model, the proposed model using functional principal components (FPCA), and the "ideal" approach, where $X_i(t)$ is known for all $t \in [0, 10]$ (IDEAL), see Section 3 for details. Shown are the Monte Carlo average of estimates (Mean), the Monte Carlo standard deviation (SD) of estimates as well as the Monte Carlo average of estimated standard errors (SE).

	Normal			Mixture		
	Mean	SD	SE	Mean	SD	SE
TRUE	-0.998	0.114	0.109	-1.03	0.117	0.118
FPCA	-1.05	0.112	0.121	-.997	0.115	0.119
IDEAL	-1.02	0.119	0.113	-1.02	0.108	0.107

Regarding the selection of the tuning parameters, the numbers of eigenfunctions and inside equi-quantile knots, we used (13), and $K = 2$ was correctly chosen for most (more than 95%) of the simulated datasets. This provides empirical justification of the effectiveness of the proposed model selection procedure and distinguishes it from the previously studied parametric joint models.

4. Application to Longitudinal CD4 Counts and Survival Data

In this clinical trial both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs in treating patients

that failed or were intolerant of zidovudine (AZT) therapy. There were 467 HIV-infected patients who met entry conditions (either an AIDS diagnosis or two CD4 counts of 300 or fewer, 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 ($n_i \leq 5$). The time to death was also 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 opportunistic infection (AIDS diagnosis) at study entry. As is customary, CD4 counts were transformed by a fourth-root power to achieve homogeneity of within-subject variance, i.e., Y_{ij} and $X_i(t)$ represent $(\mathrm{CD4}+\delta)^{1\ 4}$ (Taylor et al. (1991)), where $\delta = 1$ is to ensure the fourth-root is positive. The observed CD4 counts of eight patients from each group are displayed in Figure 1. The sample size at the five time points are $(79, 62, 62, 58, 11)$ for the ddC group, and $(81, 71, 60, 51, 10)$ for the ddI group. There is a sharply increasing amount of missing data over time, due to deaths or dropouts that are usually caused by inadequate CD4 counts. Regarding the

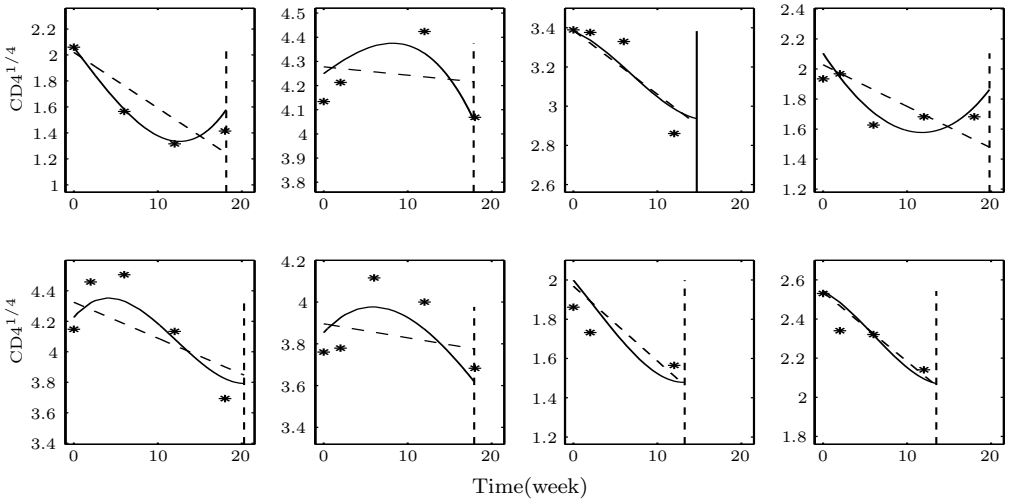


Figure 1. Observed (asterisks) CD4 counts in fourth-root scale and fitted trajectories (solid) obtained from the proposed joint model using FPCs for four randomly selected patients in the ddC group (top row), and four patients in the ddI group (bottom row), compared to the fitted trajectories obtained from the joint model with linear random effects relation (dashed). The vertical lines represent the censoring (dashed) or event (solid) time.

survival in the two drug groups, the empirical survival curves (Nelson-Aalen estimates) are shown in Figure 2, indicating that the survival rate in the ddC group looks increasingly better than that in the ddI group in time.

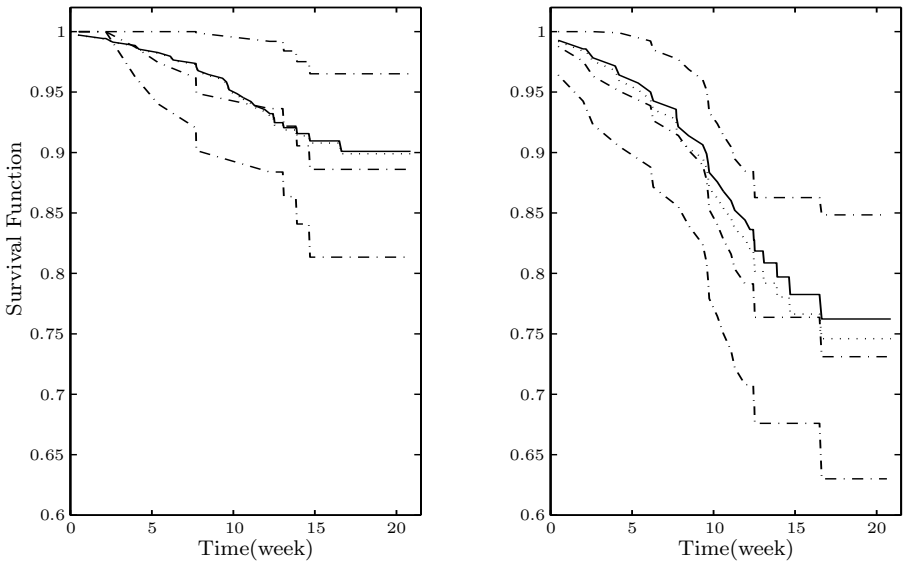


Figure 2. Estimated average survival rates in (16), obtained from the proposed model (solid lines) and the joint model with linear random effects (dotted lines) for the ddC group (left panel) and the ddI group (right panel), compared with empirical survival rates (middle dash-dotted lines) obtained from Nelson-Aalen estimates, as well as corresponding 95% empirical confidence bands (lower and upper dash-dotted lines).

The primary interest is to evaluate the relationship between CD4 trajectories of the two drug groups and the survival time. Since the CD4 counts are 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 apply the proposed joint model incorporating functional principal components to the data, where the mean CD4 curves of the two groups are modelled separately and nonparametrically using a B-spline basis. A common covariance structure is used for two groups. Then the longitudinal model is

$$\begin{aligned} Y_{ij} &= \mu_{g_i}(t_{ij}) + \sum_{k=1} \xi_{ik}\phi_k(t_{ij}) + \epsilon_{ij}, \\ &= \bar{B}_p(t_{ij})^T(\alpha + g_i\beta) + B_q(t_{ij})^T\Theta\xi_i + \epsilon_{ij}, \end{aligned} \tag{14}$$

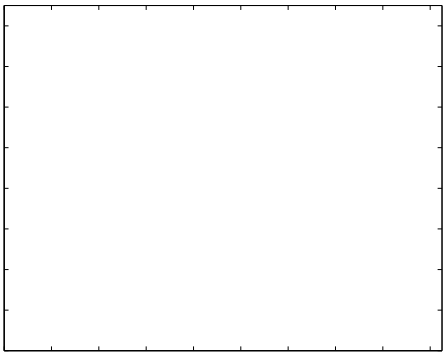
where $g_i = 0$ for the ddC group and $g_i = 1$ for the ddI group, K is the appropriate number of principal components that would be chosen together with p and q at

(13) , $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of coefficients for modelling the difference between two drug groups, and the other notations are the same as in (1) and (3). Regarding the Cox model, the effects of the underlying CD4 processes and drug groups on survival time are considered, with

$$h(t|X_i^H(t), g_i) = h_0(t) \exp\{\gamma X_i(t) + \zeta g_i\}, \qquad t \in [0, \tau], \tag{15}$$

where the duration of the study is 21.4 weeks ($\tau = 21.4$), and the other notations are as in (6).

Smooth estimates of the mean CD4 trajectories of the two drug groups obtained from the joint model combining (14) and (15) are shown in the left panel of Figure 3, which presents similar patterns for the two groups, with slightly different shapes. The large CD4 counts at the beginning of both groups may correspond to the better health conditions of patients when they entered the study. The overall trends of both groups decrease over time, while the short flat period at the end might not be scientifically important due to possible boundary effects and limited information at the longer follow-up times. Although $K = 3$, $p = 5$ and $q = 5$ are selected by the iterative procedure based on (13), the resulting population- and subject-specific curves vary considerably and it is questionable whether they represent the true CD4 counts. We instead chose to use $K = 2$, $p = 4$ and $q = 4$, noting that the AIC only increased by around 1% and the results were more realistic and interpretable, where $p = 4$ and $q = 4$ mean that 2 knots are selected for both B-spline bases. The two eigenfunctions shown in the right panel of Figure 3 are used to approximate the infinite-dimensional longitudinal process. The first eigenfunction represents an overall shift, the second corresponds to a contrast between early and late times, similar to the mean trend of, especially, the ddI group. These eigenfunctions account for about 96% and 3% of the total variation respectively. The fitted longitudinal CD4 trajectories obtained from the FPC model for four randomly selected patients from each group, $\widehat{X}_i(t) = \hat{\mu}_{g_i}(t_{ij}) + \sum_{k=1} \hat{\xi}_{ik}\hat{\phi}_k(t_{ij})$, are shown in Figure 1. The fitted curves are seen to be reasonably close to the observations. For comparison, we also fit a joint model with a linear random effects model. The group mean trends with linear patterns (left panel of Figure 3), and the estimated survival curves (Figure 2) obtained from the linear joint model are not far from those obtained from the proposed model, while the linear joint model results in a slightly larger AIC (around 3 %). Moreover one can see that the linear model fails to characterize the subjects with nonlinear patterns, while the proposed model can effectively recover the random trajectories based on just two leading FPCs determined by the data.



Acknowledgement

The author is most grateful to an associate editor and the reviewers for their constructive comments. He also thanks Thomas C. M. Lee for helpful suggestions on the presentation.

Appendix. Estimation procedures

In this section we provide details of the EM estimation procedure of the joint model combined with functional principal component analysis.

1. *E-step.* Consider the FPC scores, $\xi_i = (\xi_{i1}, \ldots, \xi_{i\ })^T$ as missing data. The conditional expectation of any function $g(\cdot)$ of ξ_i is denoted by $E\{g(\xi_i)|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega}\}$, where $\widehat{\Omega}$ is the set of current estimates of parameters $\Omega = \{\gamma, \zeta, h_0(\cdot), \alpha, \beta, \Theta, \Lambda, \sigma^2\}$. This expectation is taken with respect to the conditional density $f(\xi_i|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega})$, and can be written as (Wulfsohn and Tsiatis (1997)), under the conditional independence of (T_i, Δ_i) and Y_i given $X_i(\cdot)$ and other covariates,

$$\frac{\int g(\xi_i) f(T_i, \Delta_i|X_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0) f(\xi_i|Y_i, t_i, \hat{\alpha}, \hat{\beta}, \widehat{\Theta}, \widehat{\Lambda}, \hat{\sigma}^2) d\xi_i}{\int f(T_i, \Delta_i|X_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0) f(\xi_i|Y_i, t_i, \hat{\alpha}, \hat{\beta}, \widehat{\Theta}, \widehat{\Lambda}, \hat{\sigma}^2) d\xi_i}, \qquad (17)$$

where $f(T_i, \Delta_i|X_i^H(T_i), V_i(T_i), \hat{\gamma}, \hat{\zeta}, \hat{h}_0)$ is as in (8), and $f(\xi_i|Y_i, t_i, \hat{\alpha}, \hat{\beta}, \widehat{\Theta}, \widehat{\Lambda}, \hat{\sigma}^2)$ is the Gaussian conditional likelihood of the random coefficients given the data and current parameter estimates. The integral can be approximated by Gaussian-Hermite quadrature if K is small. To accelerate the algorithm, we use Monte Carlo integration with antithetic simulation for variance reduction (Henderson et al. (2000)).

2. *M-step.* Denote the estimate of the conditional expectation $E\{g(\xi_i)|T_i, \Delta_i, Y_i, t_i, \widehat{\Omega}\}$ by $E_i\{g(\xi_i)\}$ for convenience. The closed-from maximum likelihood estimates are

$$\hat{\lambda}_k = \frac{1}{n}\sum_{i=1}^{n} E_i(\xi_{ik}^2), \quad \hat{\sigma}^2 = \frac{1}{\sum_i n_i}\sum_{i=1}^{n} E_i\{\|Y_i - \bar{B}_i\hat{\alpha} - Z_i\hat{\beta} - B_i\widehat{\Theta}\xi_i\|^2\}, \qquad (18)$$

where [illegible]

(James et al. (2000)). Note that

$$\sum_{i=1}^{n} \|Y_i - \bar{B}_i\alpha - Z_i\beta - B_i\Theta\xi_i\|^2 = \sum_{i=1}^{n} \|(Y_i - \bar{B}_i\alpha - Z_i\beta - \sum_{\neq k} \xi_i\ B_i\theta\ ) - \xi_{ik}B_i\theta_k\|^2.$$

Then the estimate of θ_k is given by

$$\hat{\theta}_k = \Big\{ \sum_{i=1}^{n} E_i(\xi_{ik}^2) B_i^T B_i \Big\}^{-1} \sum_{i=1}^{n} B_i^T \Big\{ E_i(\xi_{ik})(Y_i - \bar{B}_i\hat{\alpha} - Z_i\hat{\beta}) - \sum_{\neq k} E_i(\xi_{ik}\xi_i\ ) B_i\hat{\theta}\ \Big\}. \quad (20)$$

This procedure is repeated for each column of Θ and iterated until no further change occurs.

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

$$\hat{\eta}^{(\ )} = \hat{\eta}^{(\ -1)} + I^{-1}_{\hat{}(l-1)} S_{\hat{}(l-1)}, \quad (21)$$

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

$$\hat{h}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i I(T_i = t)}{\sum_{j=1}^{n} E_j[\exp\{\gamma X_j(t) + V_j(t)^T\zeta\}]R_j(t)}, \quad (22)$$

where $R_j(t)$ is an at-risk indicator that is equal to $I(T_j \geq t)$, and $I(\cdot)$ is the indicator function. In our experience, the two inner iterative procedures for estimating Θ and η converge very quickly, and the computation time required for the whole algorithm is mainly determined by the dimension of the random coefficients, i.e., the number of principal components K, and the number of antithetical pairs that are used to approximate conditional expectations in the E-step while using the Monte Carlo integration method.

3. Since the matrix produced by this procedure will not be orthonormal, we need to orthonormalize it by letting $\widehat{\Gamma} = \widehat{\Theta}\widehat{\Lambda}\widehat{\Theta}^T$ and setting the final estimate $\widehat{\Theta}$ equal to the first K eigenvectors of $\widehat{\Gamma}$, while $\widehat{\Lambda}$ is the diagonal matrix consisting of the first K eigenvalues of $\widehat{\Gamma}$. The estimated FPC scores $\hat{\xi}_i$ are then obtained by computing $E_i(\xi_i)$ once more as described in the 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 Beirn 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 J. Medicine* , 657-662.

Berkey, C. S. and Kent, R. L. J. (1983). Longitudinal principal components and non-linear regression models of early childhood growth. *Ann. Human Biology* , 523-536.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* , 285-311.

Brown, E. R., Ibrahim, J. G. and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* , 64-73.

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. *Statist. Medicine* **7**, 2061–2077.

Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* , 329-337.

Cox, D. R. (1972). Regression models and lifetables (with discussion). *J. Roy. Statist. Soc. Ser. B* , 187-200.

Cox, D. R. (1975). Partial likelihood. *Biometrika* , 269-276.

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers if clinical outcomes measured with error. *Biometrics* , 1445-1462.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statist. Medicine* , 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. *J. Acquired Immune Deficiency Syndromes and Human Retrovirology* , 161-169.

Guo, X. and Carlin, B. P. (2004). Separate and joint modelling of longitudinal and event time data using standard computer packages. *Amer. Statist.* , 1-9.

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

Hsieh, F., Tseng, Y. K. and Wang, J. L. (2006). Joint modelling of survival and longitudinal data: likelihood approach revisited. *Biometrics*. To appear.

James, G., Hastie, T. G. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **7**, 587-602.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd edition. John Wiley, New York.

Pawitan, Y. and Self, S. (1993). Modelling disease marker processes in AIDS. *J. Amer. Statist. Assoc.* , 719-726.

Prentice, R. (1982). Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika* **9**, 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. *J. Roy. Statist. Soc. Ser. A* , 396-406.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York.

Rice, J. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* , 233-243.

Rice, J. and Wu, C. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **7**, 253-259.

Self, S. and Pawitan, Y. (1992). Modelling a marker of disease progression and onset of disease. In *AIDS Epidemiology: Methodological Issures* (Edited by N. P. Jewell, K. Dietz and V. T. Farewell). Birkhäuser, Boston.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* , 1-24.

Slasor, P. and Laird, N. (2003). Joint models for efficient estimation in proportional hazards regression models. *Statist. Medicine* , 2137-2148.

Song, X., Davidian, M. and Tsiatis, A. A. (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics* , 511-528.

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

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.