



Published in final edited form as:

Biometrics. 2008 June ; 64(2): 546–556. doi:10.1111/j.1541-0420.2007.00896.x.

Modeling Longitudinal Data with Nonparametric Multiplicative Random Effects Jointly with Survival Data

Jimin Ding

Mathematics Department, Washington University at St. Louis, Missouri 63130, U.S.A.
email:jmding@math.wustl.edu

Jane-Ling Wang

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

Summary

In clinical studies, longitudinal biomarkers are often used to monitor disease progression and failure time. Joint modeling of longitudinal and survival data has certain advantages and has emerged as an effective way to mutually enhance information. Typically, a parametric longitudinal model is assumed to facilitate the likelihood approach. However, the choice of a proper parametric model turns out to be more elusive than models for standard longitudinal studies in which no survival endpoint occurs. In this article, we propose a nonparametric multiplicative random effects model for the longitudinal process, which has many applications and leads to a flexible yet parsimonious nonparametric random effects model. A proportional hazards model is then used to link the biomarkers and event time. We use B-splines to represent the nonparametric longitudinal process, and select the number of knots and degrees based on a version of the Akaike information criterion (AIC). Unknown model parameters are estimated through maximizing the observed joint likelihood, which is iteratively maximized by the Monte Carlo Expectation Maximization (MCEM) algorithm. Due to the simplicity of the model structure, the proposed approach has good numerical stability and compares well with the competing parametric longitudinal approaches. The new approach is illustrated with primary biliary cirrhosis (PBC) data, aiming to capture nonlinear patterns of serum bilirubin time courses and their relationship with survival time of PBC patients.

Keywords

B-splines; EM algorithm; Functional data analysis; Missing data; Monte Carlo integration

1. Introduction

Often in medical studies both baseline and longitudinal covariates are collected for each subject, together with an event time of interest termed “survival time.” The goal is to model both the influence of covariates on survival time and the patterns of the longitudinal covariates. For instance, in AIDS studies the CD4 counts (a measure of immunologic and virologic status) of a patient are measured longitudinally and serve as biomarker for time-to-AIDS or time-to-death (Tsiatis, Degruittola, and Wulfsohn, 1995; Bycott and Taylor, 1998). Marginal approaches to separately model the longitudinal and survival components encounter difficulties

6. Supplementary Material

Web Appendixes referenced in Sections 1, 2.1, and 3 are available under the Paper Information link at the *Biometrics* web-site <http://www.biometrics.tibs.org>.

when (a) the longitudinal measurements are scattered and contain measurement errors, or (b) the longitudinal process is not observable after the event time. Whereas (a) induces bias only to the survival components, (b) will cause biases to both components. A solution to overcome both problems is to combine the survival and longitudinal components and to estimate them simultaneously. We refer to the review article in Tsiatis and Davidian (2004) for details and simulation studies, which together with those in Henderson, Diggle, and Dobson (2000) demonstrated the advantage of joint modeling. This approach not only corrects the aforementioned bias problems, but also permits the submodels for the longitudinal and the survival parts to mutually gain information from each other.

For simplicity and without loss of generality, we assume that there is one time-independent (or baseline) covariate Z_i , and one longitudinal covariate process (unobservable) $X_i(t)$, for the i th subject, whose survival time is T_i . The event times T_i may be subject to the usual random censoring, and then only the minimum of survival and censoring time, $V_i = \min(T_i, C_i)$, and the censoring indicator, $\Delta_i = 1[T_i \leq C_i]$, are observed. The longitudinal process is usually not fully observed, but rather is measured intermittently at times t_{ij} and with errors e_{ij} , which have zero mean and common variance σ_e^2 . Hence the observed longitudinal data is

$$W_{ij} = X_i(t_{ij}) + e_{ij}, \quad \text{where} \quad 0 \leq t_{ij} \leq V_i, \quad j=1, \dots, m_i, \quad (1)$$

where m_i is the number of longitudinal measurements for subject i .

For the survival component, the Cox proportional hazards model with baseline hazard function λ_0 is commonly adopted, with hazard function for the i th subject given by

$$\lambda(t|Z_i, \{X_i(s), 0 \leq s \leq t\}) = \lambda_0(t) \exp\{\beta_z Z_i + \beta_x X_i(t)\}. \quad (2)$$

So far, the joint modeling approaches have primarily relied on parametric mixed effects models for the longitudinal processes $X_i(t)$, where the unobserved random effects were treated as missing data and imputed by the expectation-maximization (EM) algorithm. The computations can be quite intensive as multidimensional numerical integration is involved in the E-steps, and the computational cost surges as the number of random effects increases. The EM-algorithm may also become unstable and might not converge. Computational efficiency thus often is the key to success for the joint modeling approaches. A parsimonious yet effective random effects model may provide a satisfactory solution and is the goal of this article.

Motivated by the primary biliary cirrhosis (PBC) data in Section 5, we propose a simple yet flexible longitudinal model using only one random effect to link the population mean function $\mu(t) = E[X_i(t)]$ to the subject-specific profile through

$$X_i(t) = b_i \mu(t) \quad \text{and} \quad E(b_i) = 1. \quad (3)$$

Consequently, each subject has a profile proportional to $\mu(t)$, with the multiplicative factor b_i describing the variation of each subject around the population mean. The unknown mean function $\mu(t)$ will be modeled and estimated nonparametrically through a set of basis functions in the *nonparametric multiplicative random effects model* (NMRE) given by (3).

Specifically, we approximate the mean function by B-spline basis functions $B_l(\cdot)$, $1 \leq l \leq L$, (using boldface for column vectors)

$$\mu(t) \approx \sum_{l=1}^L \gamma_l B_l(t) = \mathbf{B}^T(t) \boldsymbol{\gamma}. \quad (4)$$

The choice of the number of basis functions L will be discussed later. By allowing a sufficient number of B-spline functions we can approximate the overall mean function very well at little computational cost, as this involves estimating fixed effects. Meanwhile, we cut down the computational cost drastically as only one random effect is involved. A new estimating procedure for the mean function and variance components is described in Section 2. The proposed EM algorithm is very stable and converges quickly for the simulations in Section 3 and the data analysis in Section 4.

Note that our approach does not require any stationarity assumption and differs from the Bayesian approach in Brown, Ibrahim, and Degruittola (2005), where a more general nonparametric model is considered, but using the same number of random and fixed effects in an additive model. Their approach does not take advantage of the multiplicative structure of the longitudinal data that we propose. At first glance, the multiplicative random effects model (3) appears restrictive, but its applicability is much broader than anticipated. We note that longitudinal data are closely related to “functional data,” which are random functions on an interval I . A recent expository paper by Rice (2004) illuminates this by viewing longitudinal data as scattered realizations of functional data. The functional principal component (Karhunen—Loève) representation for $X(t)$,

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik} \phi_k(t), \quad t \in I, \quad (5)$$

is based on the orthogonal eigenfunctions $\phi_k(t)$ of the covariance function $G(s, t) = \text{cov}(X_i(s), X_i(t))$. The A_{ik} are uncorrelated random variables (known as the principal component scores) with mean 0 and variances λ_k , which are the eigenvalues of $G(s, t)$ in descending order. Applying functional principal components analysis (Rice and Silverman, 1991; Ramsay and Silverman, 2005) for many data sets, we observed that often the first eigenfunction $\phi_1(t)$ alone explains a large percentage (over 70%) of the variation of the data and has a similar shape as the mean function multiplied by a scale factor. This translates to $\phi_1(t) = c \mu(t)$ for some constant c , ignoring the remaining principal components A_{ik} for $k \geq 2$. Consequently, (5) is roughly equal to model (3) with $b_i = 1 + cA_{i1}$. Data sets with this feature are plentiful and include the frequently studied AIDS data in the joint modeling literature and the PBC data in Section 4. Specific examples are (i) the AIDS data in Figures 1 and 4 of Yao, Müller, and Wang (2005), (ii) the hip cycle data shown in Figure 2 of Rice and Wu (2001), and (iii) the time course of ^{14}C —*folate* in plasma of healthy adults (Yao et al., 2003). For these three data sets the first principal component explains, respectively, 76.9%, 71.2%, and 89% of the total variation, the mean functions mimic the corresponding first eigenfunctions, and the longitudinal trajectories from different subjects have different amplitude but similar overall shape, suggesting the appropriateness of the multiplicative random effects model (3) for these data.

To our knowledge, although more sophisticated random effect models have appeared in the literature, the simple random effect model in (3) has rarely been studied and certainly not in the joint modeling setting. We thus provide in the web-based supplementary material a simple algorithm to marginally estimate such longitudinal data as this could be of independent interest. The marginal estimates also serve as initial estimates for the EM algorithm in Section 2. The

new approach is applied to the PBC data in Section 4, where about 17% of the subjects have no more than two longitudinal measurements and on the average five measurements per subject have been recorded. The measurement errors add further complications if one attempted to find a suitable parametric form for the trajectories of the longitudinal covariates. A nonparametric approach such as the multiplicative model in (3) is indeed worth exploring.

2. The Joint Likelihood Approach

Among the various approaches to jointly model survival and longitudinal data, the most promising one is perhaps the joint semiparametric likelihood approach, first proposed by Wulfsohn and Tsiatis (1997). We adopt such an efficient joint likelihood approach by assuming normal measurement errors e_{ij} , which are independent and identically distributed (i.i.d.) for all i and j . The random effects are independent of the measurement errors and are independent across subjects with some known distribution.

Let $\theta = \{\beta_z, \beta_x, \lambda_0(\cdot), \gamma, \theta_b, \sigma_e^2\}$ denote the vector of all unknown parameters defined in (1)–(4) and $\lambda_0(\cdot)$ is a function of time, with the addition of θ_b to represent the parameter that determines the distribution of the random effects. Likewise, we denote the longitudinal covariate process $X_i(t)$ as $X_i(\cdot)$. The vector notations, $\mathbf{X}_i = (X_{i1}, \dots, X_{imi})^T$ and $\mathbf{W} = (w_{i1}, \dots, w_{imi})^T$, denote, respectively, the unobserved covariates and observed measurements at the scheduled visits, where $X_{ij} = X_i(t_{ij})$. With these notations the observed data for the i th subject are $(V_i, \Delta_i, Z_i, \mathbf{W}_i)$, and $i = 1, \dots, n$. Once the random effects b_i are specified, the longitudinal and survival components are independent of each other, and the joint likelihood based on the observed data is easily found to be

$$L = \prod_{i=1}^n \int f_s(V_i, \Delta_i; Z_i, X_i(\cdot), \beta_z, \beta_x, \lambda_0(\cdot)) f_l(\mathbf{W}_i; \mathbf{X}_i, \sigma_e^2) f_b(b_i; \theta_b) db_i = \prod_{i=1}^n \int f_{si} f_{li} f_{bi} db_i, \tag{6}$$

where $f_b(b_i; \theta_b)$ denotes the density function of the random effects and the other two density functions corresponding to the survival and longitudinal components are defined as follows:

$$f_{si} = [\lambda_0(V_i) \exp\{\beta_z Z_i + \beta_x X_i(V_i)\}]^{\Delta_i} \times \exp\left[-\int_0^{V_i} \lambda_0(u) \exp\{\beta_z Z_i + \beta_x X_i(u)\} du\right],$$

$$f_{li} = \frac{1}{(2\pi\sigma_e^2)^{m_i/2}} \exp\left[-\frac{1}{2\sigma_e^2} \sum_{j=1}^{m_i} (W_{ij} - X_{ij})^2\right]. \tag{7}$$

Because the maximum of the likelihood in (6) is infinity if λ_0 is completely unrestricted, we consider the nonparametric maximum likelihood (NPMLE) approach instead. Similar to the arguments in Johansen (1983) one can show that the NPMLE for λ_0 is a point mass function with positive masses λ_k assigned to only those V_k whose Δ_k is equal to 1, where $k = 1, \dots, D$, with D being the total number of uncensored observations. We can thus parameterize it by $\lambda_0(\cdot) = \{\lambda_1, \dots, \lambda_D$ and replace the baseline hazard function in (6) and (7) by this NPMLE. This results in a high-dimensional parameter whose dimension is of the same order as the sample size n . The likelihood in (6) involves unobserved random effects and integrations, so EM algorithm is typically employed to maximize it. However, the resulting EM algorithm might be unstable, due to the need of numerical integration in the E-step, especially if many random effects are included in the model. This numerical challenge is faced by all joint modeling approaches, but the simplicity of the proposed model (3) provides improvement and

computational stability, as only one-dimensional numerical integration is involved for each covariate.

2.1 EM Algorithm

The likelihood for the complete data in the EM algorithm is simply the product $\prod_{i=1}^n f_{si} f_{li} f_{bi}$. Below we describe the EM algorithm and use the superscript k in parentheses to indicate the position of the iteration, where all functions are evaluated at the current values of parameters.

For example, $E_i^{(k)} [g(b_i)] = E [g(b_i) | V_i, \Delta_i, \mathbf{W}_i, Z_i, \theta^{(k)}]$ is the conditional expectation of some function g , given the observed data of the i th subject and the estimate $\theta^{(k)}$ at the k th iteration.

Step 0 (initialization): We propose a two-stage procedure, which does not require numerical integrations and is computationally fast, for the starting ($k = 0$) parameter estimates $\theta^{(0)} = (\beta_z^{(0)}, \beta_x^{(0)}, \lambda_0^{(0)}(\cdot), \gamma^{(0)}, \theta_b^{(0)}, \sigma_e^{(0)})$ and $b_i^{(0)}$. The first stage is to estimate the longitudinal parametric components and then the resulting random effects for all subjects. This allows us to impute the longitudinal process $\widehat{X}_i(t) = b_i^{(0)} \widehat{\mu}(t)$ with the complete history for each subject. Details of the procedure are described in the supplementary material. In the second stage, we calculate a partial likelihood estimate for the survival regression parameters based on the imputed longitudinal process and then obtain the corresponding Breslow estimate for the baseline hazard function. This completes the two-stage procedure.

Step 1 (E-step): Due to the involvement of the survival density f_{si} , numerical integration is needed to evaluate the conditional expectations, $E_i^{(k)} [g(b_i)]$, for several g -functions in the M-step based on the current estimate $\theta^{(k)}$. We adopt the Monte Carlo method to generate a random sample, $\{b^1_i, \dots, b^M_i\}$, of size M for b_i from the partial posterior density $f(b_i | \mathbf{W}_i, \theta^{(k)})$ and approximate

$$E_i^{(k)} [g(b_i)] = \frac{\int_{-\infty}^{+\infty} g(b_i) f_{si}(b_i) f(b_i | \mathbf{W}_i, \widehat{\theta}) db_i}{\int_{-\infty}^{+\infty} f_{si}(b_i) f(b_i | \mathbf{W}_i, \widehat{\theta}) db_i} \quad \text{by} \\ \times \frac{\sum_{j=1}^M g(b^j_i) f_{si}(b^j_i)}{\sum_{j=1}^M f_{si}(b^j_i)}.$$

The computational advantages of the simple multiplicative model are now evident, as the above Monte Carlo EM (MCEM) procedure involves only one random effect for each longitudinal covariate.

Step 2 (M-step): For the B-spline basis, let $\mathbf{B}(t) = (B_1(t), \dots, B_L(t))^T$ be a column vector, and $\mathbf{B}_i = (\mathbf{B}(t_{i1}), \dots, \mathbf{B}(t_{imi}))$ be an $L \times m_i$ matrix. To indicate the at-risk status of the i th subject \times at time t , we define $R_i(t) = 1$ if $V_i \geq t$, and $R_i(t) = 0$ otherwise. Direct maximization of the conditional expectation log-likelihood $\sum_{i=1}^n E_i^{(k)} [\log(f_{si}) + \log(f_{li}) + \log(f_{bi})]$ leads to closed-form estimates for σ_e^2 and the point mass baseline function $\lambda_0(t)$, yielding updated estimates

$$\widehat{\sigma}_e^{2(k+1)} = \frac{1}{N} \sum_{i=1}^n E_i^{(k)} \left[\sum_{j=1}^{m_i} (W_{ij} - X_{ij})^2 \right] \\ = \frac{1}{N} \left[\sum_{i=1}^n E_i^{(k)} \left[(\mathbf{W}_i - \mathbf{B}_i \gamma^{(k)} b_i)' (\mathbf{W}_i - \mathbf{B}_i \gamma^{(k)} b_i) \right] \right], \tag{8}$$

$$\widehat{\lambda}_0^{(k+1)}(t) = \frac{\sum_{i=1}^n \prod_{[V_i=t]} \Delta_i}{\sum_{i=1}^n E_i^{(k)} \left[\exp \left\{ \beta_Z^{(k)} Z_i + \beta_X^{(k)} X_i(t) \right\} R_i(t) \right]}, \tag{9}$$

where $N = \sum_{i=1}^n m_i$ is the total number of longitudinal measurements for all subjects, and I is the indicator function. The prior parameter θ_b for the random effects may have a closed-form solution, such as for Gaussian random effects it is

$$\widehat{\theta}_b^{(k+1)} = \widehat{\sigma}_b^{2(k+1)} = \frac{1}{n} \sum_{i=1}^n E_i^{(k)} \left[(b_i - 1)^2 \right].$$

Otherwise, it can be estimated similar to β_Z, β_X , and γ , all of which have no closed-form solution. To see this, define for $h = 0, 1$, and 2 ,

$$M_U(i, h) = \sum_{j=1}^n E_j^{(k)} \left[U_j^h(V_i) \exp \left\{ \beta_Z Z_j + \beta_X X_j(V_i) \right\} R_j(V_i) \right], \tag{10}$$

where U is a generic notation and will be replaced by b, Z , and X later. The profile score function for the coefficients from the mean curve is

$$\begin{aligned} S_\gamma = & \sum_{i=1}^n \Delta_i E_i^{(k)}(b_i) \beta_X \mathbf{B}(V_i) - \sum_{i=1}^n \Delta_i \beta_X \mathbf{B}(V_i) \frac{M_b(i, 1)}{M_b(i, 0)} \\ & + \frac{\sum_{i=1}^n E_i^{(k)}(b_i) \mathbf{B}'_i \mathbf{W}_i - \sum_{i=1}^n E_i^{(k)}(b_i^2) \mathbf{B}'_i \mathbf{B}_i \gamma}{\sum_{i=1}^n E_i^{(k)} [(\mathbf{W}_i - \mathbf{B}_i \gamma b_i)' (\mathbf{W}_i - \mathbf{B}_i \gamma b_i)] / N}, \end{aligned} \tag{11}$$

and the score functions for the regression coefficients are

$$\begin{aligned} S_{\beta_Z} &= \sum_{i=1}^n \Delta_i Z_i - \sum_{i=1}^n \Delta_i \frac{M_Z(i, 1)}{M_Z(i, 0)} \\ S_{\beta_X} &= \sum_{i=1}^n \Delta_i E_i^{(k)} [b_i] \mathbf{B}'(V_i) \gamma - \sum_{i=1}^n \Delta_i \frac{M_X(i, 1)}{M_X(i, 0)}. \end{aligned} \tag{12}$$

The score functions (11) and (12) are nonlinear functions of the current γ, β_Z , and β_X . Ideally, at each EM iteration we would solve these nonlinear systems together, which could be high dimensional especially if the number L of basis functions is large. However, the EM algorithm is already an iterative procedure, so it may not be necessary to accurately update these parameters at each M-step as long as the likelihood increases steadily during the iterations (Caffo, Jank, and Jones, 2005). We thus employ the one-step Newton—Raphson approach to get the updated $\lambda^{(k+1)}, \beta_Z^{(k+1)}$, and $\beta_X^{(k+1)}$ as in Henderson et al. (2000) and Wulfsohn and Tsiatis (1997). Simulation results in Section 4 confirm the satisfactory performance of this shortcut.

Step 3 (normalizing $E(b_i) = 1$): We set $\gamma^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \times E_i^{(k+1)}(b_i) \tilde{\gamma}^{(k+1)}$ and normalize $E_i^{(k+1)}(b_i)$ by dividing $\frac{1}{n} \sum_{i=1}^n E_i^{(k+1)}(b_i)$ to ensure $E(b_i)=1$ in model (3).

Step 4: Iterate among steps 1–3 until the algorithm numerically converges.

Remarks:

1. Conceptually, it is not necessary to normalize \hat{b}_i because the approach will automatically yield estimates close to this constraint. However, we have found that employing step 3 helps to improve the estimation of the mean function when the parameters are still biased during earlier iterations, especially at the beginning of the iteration.
2. We did not control the convergence criterion for $\lambda_0(\cdot)$ in step 4 because it is not so crucial to accurately estimate λ_0 , as long as the cumulative baseline hazard function can be estimated precisely, which is often the case when all the other parameters have converged.
3. The choice of L for the number of basis functions is a model selection problem and will depend on the number of knots and the order of the B-spline functions. We used the Akaike information criterion (AIC) in the simulation and data application (Sections 5 and 6), based on the asymptotic optimality results of Shibata (1981), applicable when the number of basis functions in the true model is infinite or increases with the sample size, as is the case here.
4. The two-stage procedure in step 0 is simple and works well to generate the initial estimates. We propose two ways to estimate the random effects in the supplementary material. Tsiatis et al. (1995) had suggested a risk-set regression calibration approach to reduce the biases in the two-stage procedure when the longitudinal process is a linear mixed effects model. We could follow a similar scheme to reduce the biases in our multiplicative random effects model. However, it is not necessary to go through such an elaborated scheme just for the purpose of initial estimation.
5. There is some computational advantages to use normal random effects as the partial posterior density $f(b_i | \mathbf{W}_i, \theta^{(k)})$ needed to generate the Monte Carlo sample in the k th E-step would have a normal distribution. In non-Gaussian situations, this partial posterior density may not be trackable and Gibbs sampling may be needed to generate the Monte Carlo sample. This would induce additional computation cost. We note here that the joint likelihood approach is fairly robust against the assumption of normal random effects as observed in simulation studies in Song, Davidian, and Tsiatis (2002) and Tsiatis and Davidian (2004) and further explained from theoretical aspect in Hsieh, Tseng, and Wang (2006).

2.2 Variance Estimation

Variance estimation in joint modeling is an intriguing issue, as there is “missing information” involved in any EM algorithm. Louis (1982) suggested a correct variance formula, which essentially involves the calculation of the observed Fisher information matrix for the entire θ . However, this is a huge matrix in the joint modeling setting, due to the high dimensionality of the baseline hazard function. There is no shortcut to invert it in any joint modeling approach as explained in Hsieh et al. (2006). We thus revert to the bootstrap procedure to estimate the variance (or covariance) of θ . The exact bootstrap procedure (Efron, 1994) is similar to the one described in Section 3.5 of Tseng, Hsieh, and Wang (2005) and was employed to the PBC data. The resulting bootstrapped variance estimates seem quite reliable.

3. Simulation Study

Two simulations were performed: one has a linear longitudinal process with mean curve $\mu(t) = 3 - 0.7t$ and a simple constant baseline $\lambda_0(t) = 1$, the other has a nonlinear longitudinal trajectory with $\mu(t) = 4\log(t+1) - 2t + 3$ and a Weibull baseline function $\lambda(t) = 6t^{1/2}$. In both settings, the random effects b_i are normally distributed with mean 1 and variance 0.4. The measurement errors are first sampled from $N(0, 0.25)$, independently of the random effects, but in the second nonlinear setting they are enlarged to have variance 0.4, to demonstrate the stability of our procedure when the signal-to-noise ratio is small. The censoring times are from exponential distributions. To mimic clinical studies, all subjects alive by the end of the study (time 5) are assumed to be censored. Overall, the censoring rates are 15.3% and 21.5% in the two settings. For each subject, 1–10 longitudinal measurements are randomly located in the time interval $[0, 5]$, and these measurements are further truncated by V_i . On average, there are 4 and 3 longitudinal measurements per subject in the two simulations, respectively. We use $\beta_X = -1$ as the regression coefficient for the survival component, and 100 Monte Carlo samples with sample size 200.

Table 1 presents the bias, Monte Carlo standard deviation (SD), and mean square error (MSE) for estimates of β_X and the variance components in the longitudinal model. The PLME columns stand for the parametric linear mixed effects approach in Wulfsohn and Tsiatis (1997), which serves as benchmark, because this method utilizes the true parametric form of the longitudinal data. We report two types of two-stage (marginal) procedures. The first uses only longitudinal data before the observed event times V_i , and is designed to demonstrate the bias incurred by the informative dropout on the longitudinal data and the successful bias removal by the joint modeling approach. The second (adjacent in the table) uses all longitudinal data, including those beyond the event-times that are unobservable, so that the marginal approach in the web-based supplementary material should produce consistent estimates for the longitudinal component. The proposed joint modeling approach, which uses the first two-stage estimates as the initial estimates in the EM algorithm, is reported under the NMRE column. Our experience suggests that convergence of the PLME approach hinges upon a good choice of initial values, which should be selected with some care. The two-stage procedure worked well in our simulations and yielded less than 3% divergence rates for the PLME and none for our NMRE procedure. Moreover, the NMRE procedure performed well in terms of MSEs and nearly as well as the PLME. This is rather encouraging, as the NMRE utilizes no knowledge of the longitudinal mean shape. This excellent performance is probably due to the fact that the PLME involves two to three random effects, whereas NMRE involves only one and it demonstrates the computational accuracy gained by using fewer random effects. Moreover, the PLME approach could be grossly biased when the wrong parametric form was employed. We illustrate this in the second column (marked by ^a) of the PLME under the nonlinear simulation setting, in which the parametric longitudinal model is misspecified as being linear. In such a situation, the wrong PLME is associated with serious bias and inflated variance.

As expected, marginal estimates from the informative missing longitudinal profiles have serious bias. The marginal approach using longitudinal measurements past event time leads to much smaller biases for β with almost no biases for the variance components. Its precision is comparable with that of the joint modeling approaches, indicating the feasibility of the proposed marginal approach (described in the web-based supplementary material) when there is no informative dropout involved in the study. Because this approach uses more longitudinal measurements per subject, but yields comparable MSEs as the NMRE, we have thereby also demonstrated the advantage of utilizing the survival information to improve the longitudinal estimation in the joint modeling approach.

In Figure 1, we present the nonparametric estimates of the longitudinal mean function (left panel) and the cumulative hazard function estimates (right panel) with their point-wise 95% confidence intervals, calculated from Monte Carlo quantiles. It is interesting to note that, when the survival regression coefficient β_X is negative, the subjects with larger longitudinal measurements will have smaller risks and hence we observe more patients with lower risk or, equivalently, larger longitudinal measurements. This explains why the initial two-stage estimates are always noticeably biased upward. When β_X is positive, the situation would be reversed, and this occurs for the PBC data in Section 4. Different knots choices were explored in the simulations and provided similar results, so only the cases with knot choices $[0, 2, 5]$ and cubic splines are reported here. We also randomly selected eight individual trajectories and show the individual fitting results in Figure 2 under the nonlinear setting. The top left plot has a genuinely different shape compared to others (see the solid curve), and our estimate reflects such a variant shape well, as the fitted curve (dashed line) overlaps well with the true trajectory. The linear setting produced good fits too, which are not shown here to save space.

4. Analysis of PBC Data

We apply the proposed approach to the PBC data (Murtaugh et al., 1994) collected by the Mayo Clinic from 1974 to 1984. PBC is a chronic, fatal, but rare liver disease characterized by inflammatory destruction of the small bile ducts within the liver, which eventually leads to cirrhosis of the liver. Patients often present abnormalities in their blood tests, such as elevated and gradually increased serum bilirubin. Characterizing the patterns of time courses of bilirubin levels and their association with death due to PBC is of medical interest. In this randomized clinical trial, 158 out of 312 patients took the drug D-penicillamine, whereas the other patients were assigned to a control group. Some baseline covariates, such as age and gender, were recorded at the beginning of the study. Lab test results, such as serum bilirubin and albumin, were measured at the time of recruitment and at irregularly occurring follow-up visits, recorded until death or censoring. The observed event time ranges from 41 to 5225 days and 140 out of 312 patients had died by the end of study.

Fleming and Harrington (1991) studied the PBC data using only baseline covariates, and concluded that the drug D-penicillamine is not effective and five baseline covariates, including bilirubin, are significant. In our analysis, we take into account the subsequent longitudinal measurements during the follow-up period (Therneau and Grambsch, 2000). For simplicity, we only include the most significant biomarker, serum bilirubin, and one time-independent covariate, the treatment type. Model (2) will then allow us to (i) examine the treatment effect on survival after adjusting the longitudinal bilirubin levels, and (ii) check whether bilirubin still carries information on survival during the follow-up period after accounting for treatment assignment. If, however, we are interested to learn the effect of the treatment on bilirubin time courses itself and whether there is additional information in bilirubin on survival, after adjusting for any such potential treatment effect, we would need to modify both models (2) and (3) along the lines suggested in Aalen et al. (2004) and Fosen et al. (2006). Results not presented here indicate no treatment effects to bilirubin, so models (2) and (3) seem appropriate.

In a preprocessing step, we follow the clinical literature and take the logarithmic transformation of bilirubin. In the upper left panel of Figure 3, a three-dimensional plot shows all 312 bilirubin profiles, ordered by their event time V_i . Among them, 30 random samples are selected and shown in the upper right panel. Noticeably, the observed bilirubin data are noisy and vary a lot among patients, so it is not easy to pin down a suitable parametric model. We thus turn for the moment to the nonparametric functional principal component analysis approach described in Yao et al. (2005), which well accommodates sparse longitudinal data like these. The estimated mean function, $\mu(t)$, and the first eigenfunction, $\varphi_1(t)$, are shown in the lower panel of Figure 3 and referred as the “naive” mean function and eigenfunction estimates, because we

have not dealt with the informative drop-outs yet. The first eigenfunction alone explained 79.6% of the variation and apart from scale has a similar shape as the naive mean function except in the right boundary region where the data are sparse and where both the mean and eigenfunctions are subject to heavier bias due to the informative drop-outs. This suggests that the proposed multiplicative random effects model (3) is suitable for the bilirubin data. The results are presented in Figure 4a and b, for four internal knots as selected by AIC and the quadratic B-spline basis. The resulting longitudinal plots are shown in Figure 4a with the 95% pointwise bootstrap confidence intervals based on 50 bootstrap samples. The fitted mean curve is nearly the same as the bootstrap mean curve, but the bias of the naive mean curve, replotted from Figure 3 is visually striking. The direction of the bias is as expected, because bilirubin is expected to rise as the disease progresses, and patients with high bilirubin usually die earlier. The naive mean curve underestimates the seriousness of the disease.

We also randomly select six subjects and plot their expected bilirubin trajectories together with 95% confidence bands in Figure 4a next to the mean curves. The top three plots are from the control group and the bottom three from the treatment group. The estimated cumulative hazard and baseline hazard functions are displayed in Figure 4b. For the baseline hazard estimate we use the kernel hazard estimates in Müller and Wang (1994), which employed Epanechnikov kernel with boundary adaptation and local optimal bandwidth that minimizes the estimated local MSE. The baseline hazard function has a slight curvature but is not significantly different from a constant function. This matches the cumulative hazard function, which is very close to a linear function, and suggests that exponential distribution might be a good model for the baseline hazard function if one wishes to adopt a parametric approach for the survival data.

The parametric estimates for survival regression parameter and the variance components of the longitudinal data, as well as their bootstrap means and SDs, are presented in Table 2. The coefficient estimated from the baseline bilirubin in Fleming and Harrington's book is 0.8, significantly lower than 1.07, the estimate from NRME. This suggests that using only the baseline information on bilirubin may underestimate the risk for a patient. The estimate for the treatment coefficient is $\hat{\beta}_z = 0.07$ with a bootstrap standard error of 0.17, indicating an insignificant drug effect. This is consistent with the findings in the literature based on baseline bilirubin levels. We thus conclude that D-penicillamine is ineffective to treat PBC, whether one controls for bilirubin levels or not. We also combined the treatment and control groups in a second analysis to see how bilirubin relates to the patient's risk. The results were similar to those above and are not reported in detail.

5. Conclusions and Discussion

We have presented a parsimonious nonparametric approach to fit the longitudinal profiles simultaneously with the survival data. Through expansions on B-splines and the joint modeling approach we recover the population- and subject-specific trajectories of the longitudinal covariates, and characterize the association between covariates and event time. The multiplicative random effects model is computationally attractive and allows us to incorporate multiple longitudinal covariates and time-independent covariates.

To check the applicability of the multiplicative assumption, one may plot the logarithm of the observed or presmoothed individual profiles to see whether they are parallel, when the longitudinal measurements are abundant per subject. In the case of sparse longitudinal data, the shape of individual profiles will be difficult to detect and model checking may have to be carried out through some more elaborate scheme. For instance, one could follow the motivation provided in Section 1 to perform a functional principal component analysis (Yao et al., 2005) for the longitudinal component and then check whether (1) the first eigenfunction has a similar shape as the mean function, and (2) the first principal component explains a large proportion

of the total variation. It is a judgement call to determine which fraction of variation is deemed sufficient to explain most of the variation of the longitudinal data. In our experience, we find the 70% threshold adequate, but the threshold will generally depend on the data. An unbiased functional principal component analysis in the joint modeling setting is currently being investigated. An ad hoc approach at this point would be to first carry out the Bayesian B-spline approaches in Brown et al. (2005) and derive from this a preliminary estimate of the covariance structure for the longitudinal data. Once we have a reliable covariance estimate, the approach in Yao et al. (2005) could be employed and will not lead to biases. Formal model diagnostics about the multiplicative model assumption can be done by including additional eigenfunctions (Chiou, Müller, and Wang, 2004) in the joint modeling setting and checking the adequacy of the first term.

We have discussed briefly in Section 4 how to add treatment effects to the multiplicative longitudinal model following the framework in Aalen et al. (2004) and Fosen et al. (2006). Other covariates can also be added to the longitudinal process by allowing the random effects to depend on the covariates. For instance, if there is only one covariate Z , the regression function under the multiplicative longitudinal model (3) would be $E(X(t)|Z)=E(b|Z) \mu(t)$, and one could model $E(b|Z)$ either parametrically or / nonparametrically. When there are multiple covariates $Z = (Z_1, \dots, Z_p)$, a fully nonparametric approach will not be practical and some dimension reduction approaches will be needed to model the covariate effects. One approach is to assume

additive effects for each covariates, namely, $E(b|Z) = \sum_{i=1}^p g_i(Z)$ with the functions g_i unspecified. Another approach is to assume that a few indices summarize the information in all covariates, for instance, $E(b|Z) = g(\alpha^T Z)$ for some unknown function g and p -dim vector α . Note that this would be parametric if g is a known function. None of the above approaches would affect the estimate for $\mu(t)$ because it is still the overall mean function.

We have allowed for negative b_i and negative $X_i(t)$ in the simulations and data analysis. This is because some subjects have negative values for their logarithm of bilirubin or their log (bilirubin) have an overall decreasing trend. The possibility of negative b_i is also supported by the SD estimate of 1.31 in Table 2 as the mean is 1. In other applications, it may be impossible or undesirable to have negative b_i . To impose such a constraint we could use truncated normal distribution for the random effects or other non-Gaussian distribution, such as lognormal, Weibull, and Gamma, on the random effects.

Theoretical arguments on the legitimacy of AIC in the joint modeling setting are not available yet and are worth pursuing. Other model selection procedures, such as BIC, HQ, and cross-validation (or generalized cross-validation) could be employed as well, but they would all be ad hoc at this point, as there are no theoretical justifications available for any model selection procedure under the joint modeling setting. Obviously, there is much work to be done, including model diagnostics and formal inference procedures based on theoretical advances.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The research of Jane-Ling Wang was supported in part by the U.S. National Science Foundation (DMS04-06430). The authors wish to thank the reviewers, especially the associate editor, for insightful suggestions, which led to a much improved version of the article.

References

- Aalan OO, Fosen J, Weedon-Fekjær H, Borgan Ø, Husebye E. Dynamic analysis of multivariate failure time data. *Biometrics* 2004;60:764–773. [PubMed: 15339300]
- Brown ER, Ibrahim JG, Degruetola V. A flexible {B-spline} model for multiple longitudinal biomarkers and survival. *Biometrics* 2005;61:64–73. [PubMed: 15737079]
- Bycott P, Taylor J. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* 1998;17:2061–2077. [PubMed: 9789914]
- Caffo BS, Jank W, Jones G. Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society, Series B* 2005;67:235–251.
- Chiou JM, Müller HG, Wang JL. Functional response models. *Statistica Sinica* 2004;14:675–693.
- Efron B. Missing data, imputation and bootstrap (with discussion). *Journal of the American Statistical Association* 1994;89:463–479.
- Fleming, TR.; Harrington, DP. *Counting Process and Survival Analysis*. Wiley; New York: 1991.
- Fosen J, Borgan Ø, Weedon-Fekjær H, Aalan OO. Dynamic analysis of recurrent event data using the additive hazard model. *Biometrical Journal* 2006;48:381–398. [PubMed: 16845903]
- Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000;4:465–480. [PubMed: 12933568]
- Hsieh F, Tseng YK, Wang JL. Joint modelling of survival and longitudinal data likelihood approach revisited. *Biometrics* 2006;62:1037–1043. [PubMed: 17156277]
- Johansen S. An extension of Cox's regression model. *International Statistics Review* 1983;51:165–174.
- Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1982;44:226–233.
- Müller HG, Wang JL. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* 1994;50:61–76. [PubMed: 8086616]
- Murtaugh PA, Dickson ER, Van Dam GM, Malincho M, Grambsch PM, Langworthy AL, Gips CH. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology* 1994;20:126–134. [PubMed: 8020881]
- Rice J. Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica* 2004;14:631–647.
- Rice J, Silverman B. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* 1991;53:233–243.
- Rice J, Wu C. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 2001;57:253–259. [PubMed: 11252607]
- Ramsay, J.; Silverman, B. *Functional Data Analysis*. Vol. 2nd. Springer; New York: 2005.
- Shibata R. An optimal selection of regression variables. *Biometrika* 1981;68:45–54.
- Song X, Davidian M, Tsiatis AA. A semi-parametric likelihood approach to joint modelling of longitudinal and time-to-event data. *Biometrics* 2002;58:742–753. [PubMed: 12495128]
- Therneau, TM.; Grambsch, PM. *Modeling Survival Data: Extending the Cox Model*. Springer; New York: 2000.
- Tseng YK, Hsieh F, Wang JL. Joint modelling of accelerated failure time and longitudinal data. *Biometrika* 2005;92:587–603.
- Tsiatis AA, Davidian M. Joint modelling of longitudinal and time-to-event data: An overview. *Statistica Sinica* 2004;14:809–834.
- Tsiatis AA, Degruetola V, Wulfsohn MS. 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* 1995;90:27–37.
- Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997;53:330–339. [PubMed: 9147598]
- Yao F, Müller HG, Clifford AJ, Dueker SR, Follett J, Lin Y, Buchholz BA, Vogel JS. Shrinkage estimation for functional principle component scores with application to the population kinetics of plasma folate. *Biometrics* 2003;59:676–685. [PubMed: 14601769]

Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal analysis. *Journal of the American Statistical Association* 2005;100:577–590.

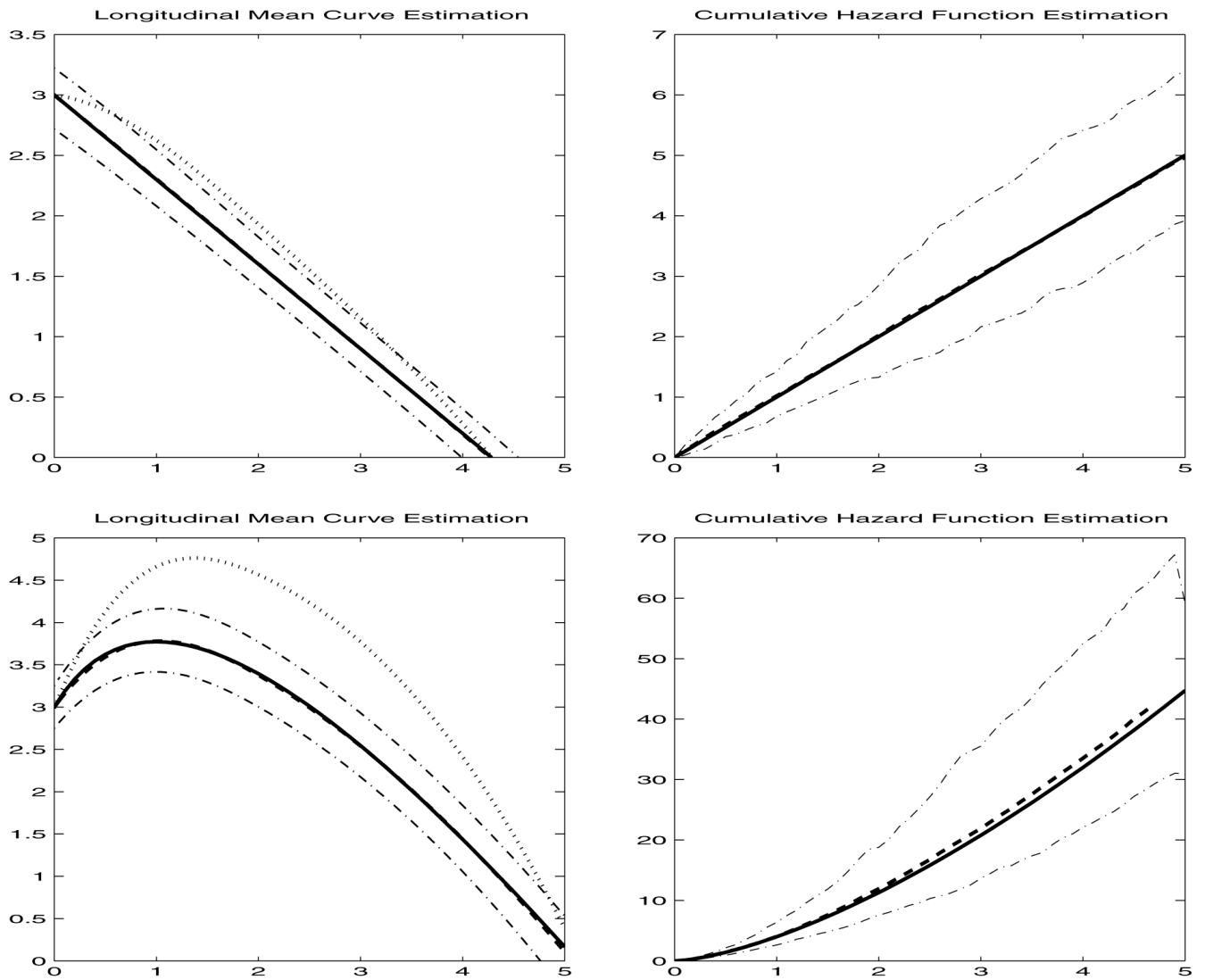


Figure 1. Simulation results: estimated mean function (top: $\mu(t) = 3 - 0.7t$, bottom: $\mu(t) = 4 \log(t + 1) - 2t + 3$) and cumulative hazard function with 95% Monte Carlo confidence intervals (dash-dot lines). The solid lines are the true function, dotted lines are the two-stage estimates, and the dashed lines are the estimates based on the procedure in Sections 2 and 3.

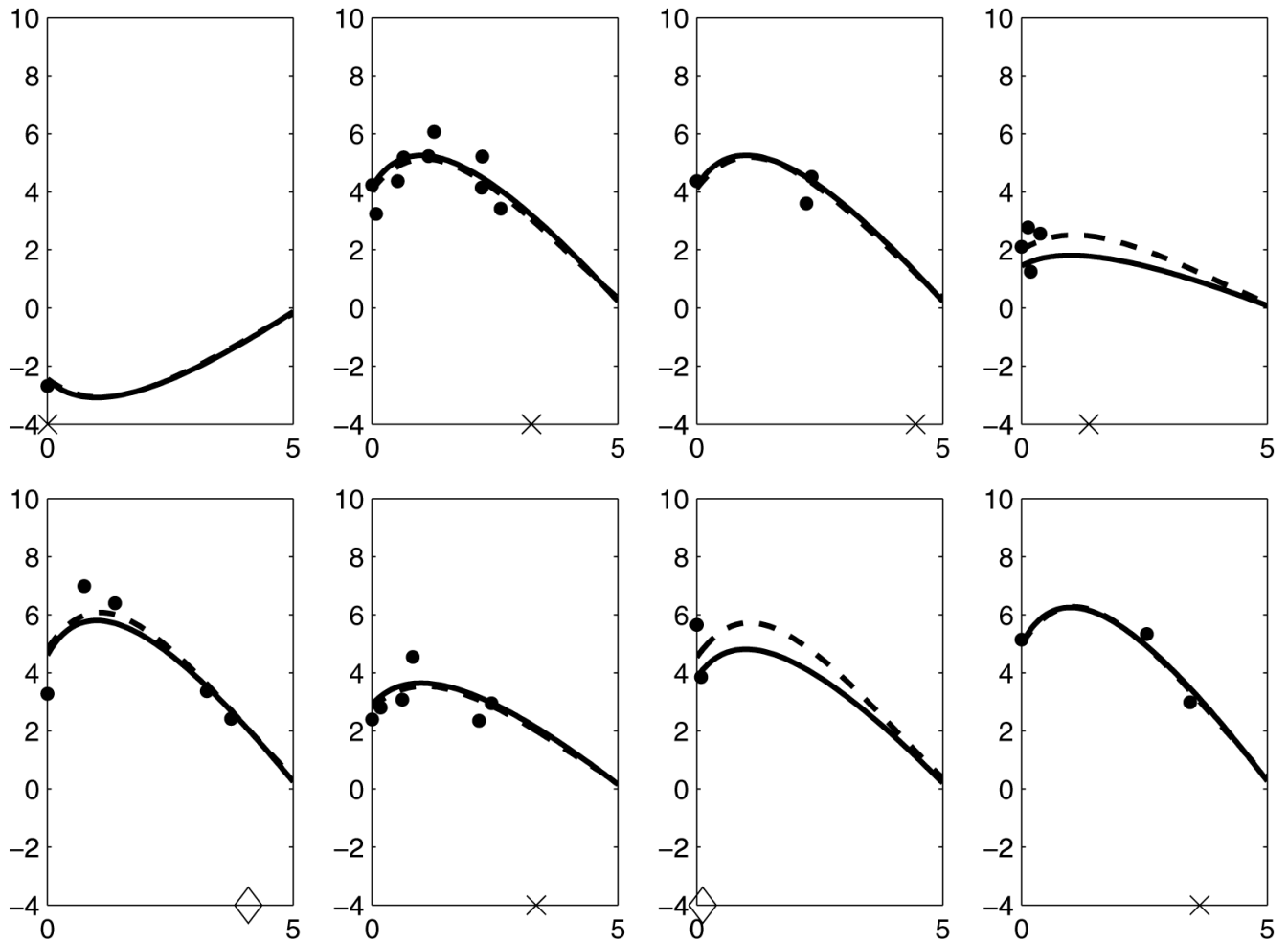


Figure 2. Individual longitudinal profiles: the true (solid line) and fitted (dashed line) profiles of eight randomly selected individuals from a random Monte Carlo sample. The observed longitudinal data are also plotted around the true curve, and the event time for each individual is marked on the x -axis with a cross indicating death and a diamond for censored observation.

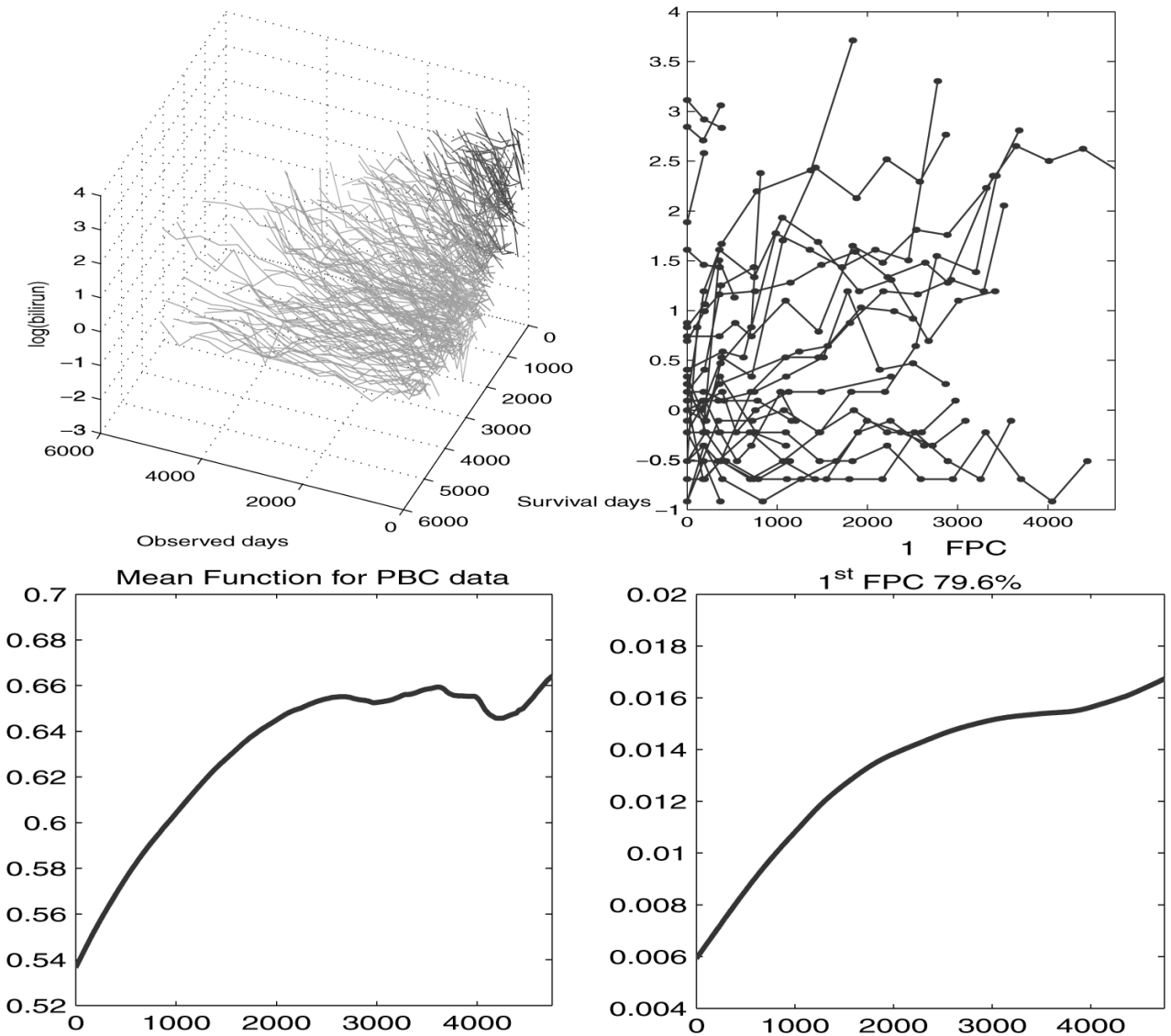


Figure 3. Upper left panel: 3D observed logarithm serum bilirubin ordered by event time for all patients; Upper right panel: the observed logarithm serum bilirubin trajectories for 30 randomly selected subjects, where the observed longitudinal measurements are denoted by dots; Lower left panel: the naive mean estimate for serum bilirubin; Lower right panel: the first functional principal component that explains 79.6% of the variation of the data. The mean curve and covariance smoothing bandwidth used for the lower panel figures are 1547 days and 1095 days, respectively, chosen by cross-validation in each procedure.

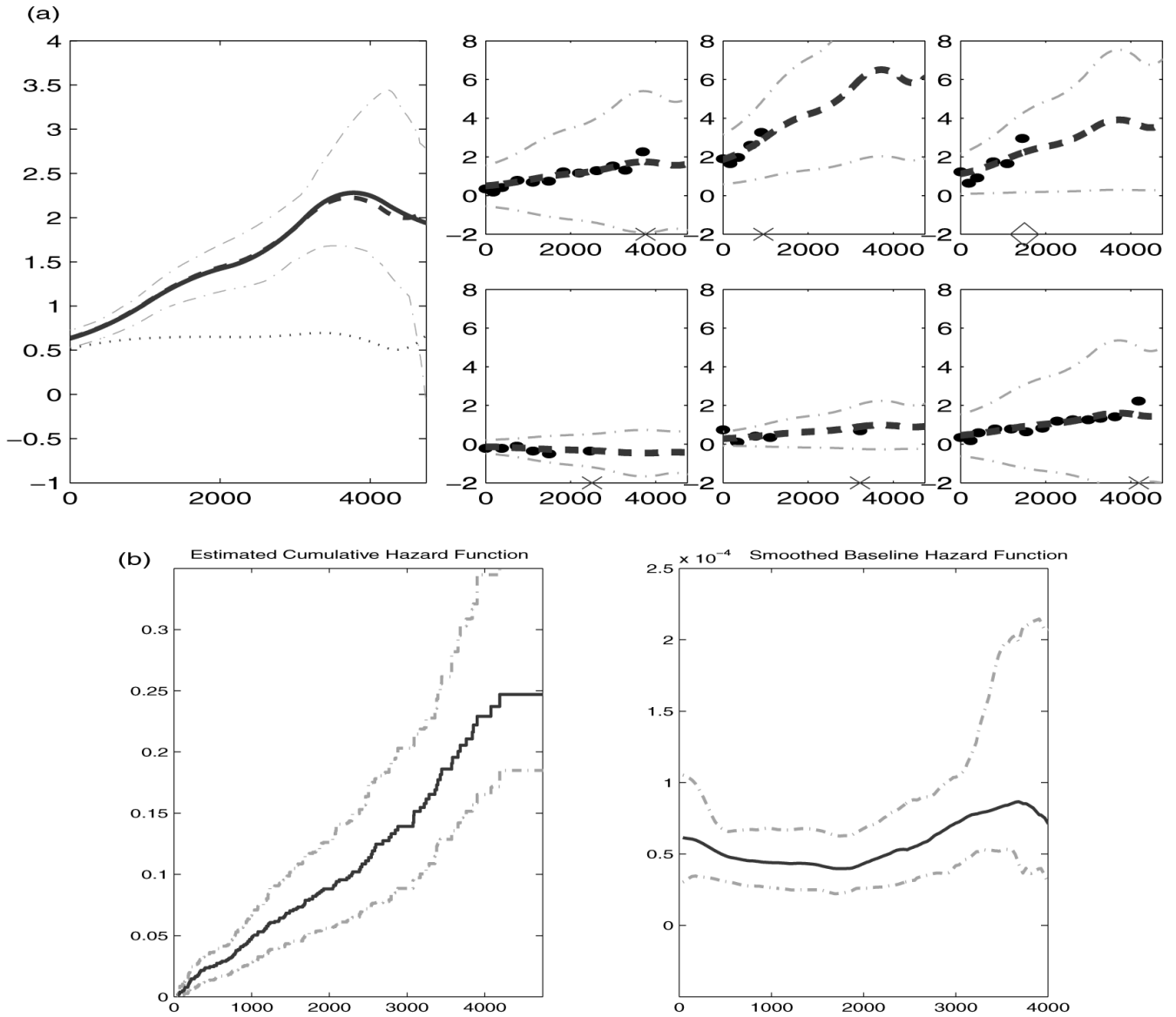


Figure 4. Part (a) shows mean curve estimation and individual bilirubin profiles. Left panel is fitted longitudinal mean (dash line for NMRE and dotted line for TwoStage) with 95% bootstrap confidence interval (dash-dot lines) and mean of bootstrap estimation (solid line); right panel contains longitudinal bilirubin process for six randomly selected individuals (dots: observed bilirubin; crosses on x -axis: event time; diamonds on x -axis: censoring time; dashed lines: fitted longitudinal bilirubin; dash-dot lines: 95% prediction intervals). Part (b) shows the estimated baseline cumulative hazard (in the left panel) and smoothed hazard functions (in the right panel), with their 95% bootstrap confidence intervals (dash-dot lines).

Table 1

Simulation results: estimation of β and the variance components, σ_b^2, σ_e^2 in the two settings. Four methods are compared: joint likelihood based on parametric linear mixed effects (PLME) model, two types of two-stage procedures (Twostage) with/without informative dropouts, and the joint likelihood approach of Section 2 with nonparametric multiplicative random effects (NMRE) model.

Linear longitudinal mean: $\mu(t) = 3 - 0.7t$									
β									
Method	PLME			Twostage			NMRE		
	PLME	Twostage	NMRE	PLME	Twostage	NMRE	PLME	Twostage	NMRE
σ_e^2									
σ_b^2									
Bias	-0.007	0.081/.019	0.006	-0.004	0.190/.014	-0.0004	-0.124/-0.009	0.001	0.001
SD	0.110	0.098/.098	0.110	0.014	0.100/.063	0.014	0.043/.071	0.067	0.067
MSE	0.012	0.016/.010	0.012	0.0002	0.046/.004	0.0002	0.017/.005	0.004	0.004
Nonlinear longitudinal mean: $\mu(t) = 4 \log(t+1) - 2t + 3$									
β									
Method	PLME ^a			Twostage			NMRE		
	PLME ^a	Twostage	NMRE	PLME	Twostage	NMRE	PLME	Twostage	NMRE
σ_e^2									
σ_b^2									
Bias	0.0004/-0.202	0.122/.040	-0.002	-0.013	1.016/.002	0.0001	-0.262/.023	0.008	0.008
SD	0.095/.176	0.104/.083	0.094	0.029	0.338/.063	0.029	0.030/.149	0.056	0.056
MSE	0.009/.072	0.026/.009	0.009	0.001	1.146/.004	0.001	0.069/.023	0.003	0.003

^aMisspecified PLME model with only slope and intercept.

Table 2

NMRE results for PBC data

	β_z	β_x	σ_e^2	σ_b^2
Estimates	0.07	1.07	0.20	1.72
Bootstrap mean	0.02	1.07	0.21	1.76
Bootstrap SD	0.17	0.11	0.01	0.23