Biostatistics (2014), **0**, 0, *pp.* 1–24
doi:10.1093/biostatistics/MainManuscript

Standard error estimation in the EM algorithm when joint modeling of survival and longitudinal data

Cong Xu, Paul D. Baines and Jane-Ling Wang*

Department of Statistics, University of California, Davis, California 95616, U.S.A

jlwang.ucdavis@gmail.com

Summary

Joint modeling of survival and longitudinal data has been studied extensively in recent literature. The likelihood approach is one of the most popular estimation methods employed within the joint modeling framework. Typically the parameters are estimated using maximum likelihood, with computation performed by the EM algorithm. However, one drawback of this approach is that standard error (SE) estimates are not automatically produced when using the EM algorithm. Many different procedures have been proposed to obtain the asymptotic variance-covariance matrix for the parameters when the number of parameters is typically small. In the joint modeling context, however, there may be an infinite dimensional parameter, the baseline hazard function, which greatly complicates the problem so that the existing methods cannot be readily applied. The profile likelihood and the bootstrap methods overcome the difficulty to some extent, however, they can be computationally intensive. In this paper, we propose two new methods for SE estimation using the EM algorithm that allow for more efficient computation of the SE of a subset of

*To whom correspondence should be addressed.

parametric components in a semi-parametric or high-dimensional parametric model. The precision and computation time are evaluated through a thorough simulation study comparing our new methods with the profile likelihood and bootstrap methods. We conclude with an application of our SE estimation method to analyze an HIV clinical trial dataset.

Key words: EM algorithm; HIV clinical trial; Numerical differentiation; Observed information matrix; Profile likelihood; Semiparametric joint modeling.

1. Introduction

In biomedical studies it has become increasingly common to record key longitudinal measurements up to a possibly censored time-to-event (or survival time) along with additional relevant covariates. The classical example in this context is an HIV clinical trial with the CD4 counts being the key longitudinal measurements. Researchers' interests are usually twofold: (1) to model the pattern of change of the longitudinal process, and, (2) to characterize the relationship between the survival process, the longitudinal process and any additional covariates. Unfortunately, difficulties arise if traditional methods are employed. The longitudinal process is subject to informative missingness due to the occurrence of the survival time; moreover, the longitudinal responses are only collected intermittently and may involve measurement error. Joint modeling approaches that model the event time and longitudinal process jointly can be readily applied to overcome the difficulties and have been studied extensively in the recent literature. **?** and **?** provide an overview of the joint modeling literature in this context. As detailed in Section 2, the presence of a high- or infinite-dimensional hazard function makes standard error estimation for joint models challenging. To overcome these difficulties we propose a novel profiling approach for SE estimation using the EM algorithm. While we focus on the joint modeling of survival and longitudinal data, the proposed profiling procedure is very general and can be applied to any context where SE estimates are required in the presence of a high-dimensional nuisance parameter or any joint modeling setting, where two, possibly connected, models for different targets, are modeled simultaneously together. An example is the joint modeling of mean and covariance structure, which could be but has not been illustrated in this paper.

In early joint modeling literature, the survival times were modeled parametrically, i.e. the baseline hazard was assumed to be the hazard function of some standard survival distribution. These joint models lead to tractable computation and can be valuable in many contexts where the assumptions can be verified. However, due to the strong model assumptions more flexible modeling is needed both for comparative purposes and to handle applications where parametric modeling assumptions are violated. Some later papers suggested approximating the baseline hazard by piecewise constant functions or spline-based methods while others adopt the Cox proportional hazards model, where the baseline hazard is left completely unspecified. Both the piecewise constant and the spline-based approaches, as implemented in the R package "`JM`" (**?**), are examples of the method of sieves, where the sieve spaces are seemingly parametric after a proper sieve dimension has been selected. **?** discusses properties of the method of sieves approach in the context of joint modeling and the issue of sieves biases as the true baseline hazard function may not belong to the sieve spaces. Typically, a moderate or large number of parameters is needed to model the sieve space so the sieve bias can be contained. In this regard, the profiling method proposed in Section 3.2 and 3.3 can be useful for standard error estimation under such flexible hazard models where the number of parameters is large.

In the rest of the paper we focus on the more challenging Cox proportional hazards model and the semiparametric likelihood approach first proposed by **?**, where a nonparametric maximum likelihood method was employed and an EM algorithm was derived for parameter estimation. **?** and **?** proved consistency and asymptotic normality of the MLE but no explicit form for the asymptotic variance-covariance matrix is available. This raises the question of how to estimate the standard error, i.e. the standard deviation, of the parameter estimates. Several approaches, for instance using the bootstrap (**?**) or the profile likelihood (**?**) approach, have been proposed in the literature but their performance has not been examined systematically for joint modeling. The goal of this paper is to address the important issue of standard error (SE) estimation in the semiparametric joint modeling setting and, where necessary, to provide new solutions.

Two key factors contribute to the difficulty of standard error estimation for the semiparametric joint modeling. The first is that the likelihood function typically involves integrals that cannot be computed analytically. The second is the presence of the nonparametric baseline hazard function employed in the Cox model. As a result, direct maximization of the likelihood function is unstable and the EM algorithm is utilized to provide computational stability. Since first appearing in the statistical literature in **?**, the EM algorithm has become a popular tool for computing maximum likelihood estimates for multi-level and missing data models. The celebrated property of monotone convergence in the observed data log-likelihood endows the algorithm with a high degree of numerical stability. However, one drawback of the EM algorithm is that the SE estimates of the parameters are not automatically produced, thus requiring additional procedures to enable practical inference.

The first major contribution to SE estimation using the EM algorithm came from **?**. The approach therein uses a formula for computing the observed information matrix in terms of the complete and missing information matrices. However, the missing information requires calculating the conditional expectation of the outer product of the complete-data score vector, an inherently problem-specific task that can require much computational effort as discussed in **?**. To address these deficiencies, many alternative EM-based procedures have been proposed to estimate SEs. Key references include **?**, **?** and **?**. The overwhelming majority of applications utilizing these methods have been restricted to parametric settings where the number of parameters is typically small, and, to our best knowledge, none explicitly address SE estimation for semi-parametric models. The baseline hazard function (the nuisance parameter) under the semiparametric joint modeling setting complicates the problem even though our interest is typically only in the SE estimates for the finite dimensional parameter. The SEM algorithm proposed by **?** turns out to be poorly suited to joint modeling applications since it requires very accurate MLEs for all parameters. Since the E-step in the joint modeling context cannot be computed in closed form, the method can be numerically unstable, as also addressed in **?**.

In this paper, we build from the core ideas of the FDM/REM/FDS/RES methods introduced in **?**. The basic approach of the FDM/REM algorithms is to numerically differentiate the EM update operator using either forward difference (FDM) or Richardson extrapolation (REM). The FDS/RES methods instead proceed by numerically differentiate the Fisher score vector, again using either forward differencing (FDS) or Richardson extrapolation (RES). Applying these algorithms directly in our joint modeling context is typically infeasible due to the dimensionality of the baseline hazard function. Therefore, we propose a novel profile technique to profile out the nuisance parameter so that the methods can be comfortably implemented in our semiparametric joint modeling setting.

In general semiparametric settings two main techniques have been studied for SE estimation, namely the bootstrap (**?**) and the profile likelihood approach (**?**). Numerical performance of the methods in the joint modeling setting has been verified in **?** and **?** respectively. However, both of these two approaches have limitations. The bootstrap approach is computationally intensive and its theoretical properties in non-iid contexts remain unclear. The profile likelihood approach is based on an approximation to the second derivative of the profile likelihood function (see (3.1.2)), which in practice has been found to be sensitive to the accuracy when evaluating the function and to the increment chosen when performing the numerical differentiation. Hence, a second goal in this paper is to determine which SE estimation methods provide the best trade-off between reliably providing precise SE estimates and computational efficiency.

The remainder of the article is structured as follows. The basic semiparametric joint modeling framework and notation are introduced in Section 2. In Section 3, we explain the idea of each SE estimation method and the corresponding implementation details for the semiparametric setting. A simulation study is provided in Section 4 to facilitate the comparison of all the aforementioned methods and a substantive data analysis (HIV clinical trial data) follows. Finally, in Section 5, we present conclusions, recommendations and possible extensions. Some technical details of the algorithms are deferred to the Appendix of the Supplementary Materials.

2. Semiparametric Joint modeling of survival and longitudinal processes

In the semiparametric joint modeling framework, the continuous longitudinal outcomes are commonly modeled with a linear mixed effects model and the survival times are assumed to follow a proportional hazards model (**?**). The model proposed by **?** assumes that the fixed and random effects in the linear mixed effects model share the same set of covariates and the entire longitudinal trajectory is involved in the survival model. Here, we adopt a more flexible model described below.

Let $\mathbf{X}_i(t)$, $\mathbf{Z}_i(t)$ be vectors of the observed covariates which are assumed to be either time-independent or time-dependent. Typically $\mathbf{Z}_i(t)$ is a subvector of $\mathbf{X}_i(t)$ but this is not required. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})$ denote the longitudinal process which is modeled as

$$Y_{ij} = Y_i(t_{ij}) = m_i(t_{ij}) + \varepsilon_{ij} = \mathbf{X}_i^\top(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i^\top(t_{ij})\mathbf{b}_i + \varepsilon_{ij} = \mathbf{X}_{ij}^\top\boldsymbol{\beta} + \mathbf{Z}_{ij}^\top\mathbf{b}_i + \varepsilon_{ij}, \tag{2.1}$$

with $\mathbf{b}_i \sim \mathcal{N}(0, \Sigma_b)$ and $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$. The survival time T_i may be subject to right censoring by the censoring time R_i. The observed data are $V_i = \min(T_i, R_i)$ and the censoring indicator $\Delta_i = I(T_i \leqslant R_i)$. The hazard function of the survival time can be modeled as

$$\lambda(t|\mathbf{b}_i, m_i(t), \mathbf{W}_i(t)) = \lambda(t) \exp\left\{\mathbf{W}_i^\top(t)\boldsymbol{\gamma} + \alpha m_i(t)\right\}, \tag{2.2}$$

or,

$$\lambda(t|\mathbf{b}_i, \mathbf{Z}_i(t), \mathbf{W}_i(t)) = \lambda(t) \exp\left\{\mathbf{W}_i^\top(t)\tilde{\boldsymbol{\gamma}} + \tilde{\alpha}\mathbf{Z}_i^\top(t)\mathbf{b}_i\right\} \tag{2.3}$$

where $\mathbf{W}_i(t)$ is also a vector of the observed covariates. $\lambda(t)$ is the baseline hazard function and is left completely unspecified. In (2.2), the error-free longitudinal process serves as a covariate in the survival model. Model (2.3) is a reparameterization of (2.2) where up to a scale factor, the survival model only shares the same random effects with the longitudinal trajectory but possibly different fixed effects. Model (2.3) is computationally simpler since $\boldsymbol{\beta}$ is not involved in the survival model while (2.2) has more explanatory power for characterizing the effect of the longitudinal process on the survival time. For the following derivations we focus on model (2.2), although the corresponding derivations based on model (2.3) follow similarly.

The observed and complete data are denoted by $\mathbf{O}_i = (\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, V_i, \Delta_i)$ and $\mathbf{C}_i = (\mathbf{O}_i, \mathbf{b}_i)$, respectively. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_e^2, \Sigma_b, \boldsymbol{\gamma}, \alpha)$ (a vector of length p) denote the finite dimensional parameter, Λ the cumulative baseline hazard function and $\boldsymbol{\eta} = (\boldsymbol{\theta}, \Lambda)$. Generally, $\boldsymbol{\theta}$ is the parameter of interest and Λ is considered a nuisance parameter. The observed- and complete-data likelihood functions are

$$f_n(\mathbf{O}|\boldsymbol{\eta}) = \mathcal{L}_n(\boldsymbol{\eta}|\mathbf{O}) = \prod_{i=1}^{n} \int f(\mathbf{Y}_i|\mathbf{b}_i; \boldsymbol{\theta}) f(V_i, \Delta_i|\mathbf{b}_i; \boldsymbol{\eta}) f(\mathbf{b}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i, \tag{2.4}$$

$$f_n(\mathbf{C}|\boldsymbol{\eta}) = \mathcal{L}_n(\boldsymbol{\eta}|\mathbf{C}) = \prod_{i=1}^{n} f(\mathbf{Y}_i|\mathbf{b}_i; \boldsymbol{\theta}) f(V_i, \Delta_i|\mathbf{b}_i; \boldsymbol{\eta}) f(\mathbf{b}_i; \boldsymbol{\theta}), \tag{2.5}$$

where

$$f(\mathbf{Y}_i|\mathbf{b}_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{-\frac{(Y_{ij} - \mathbf{X}_{ij}^\top\boldsymbol{\beta} - \mathbf{Z}_{ij}^\top\mathbf{b}_i)^2}{2\sigma_e^2}\right\},$$

$$f(V_i, \Delta_i|\mathbf{b}_i; \boldsymbol{\eta}) = \left[\lambda(V_i) \exp\{\mathbf{W}_i^\top(V_i)\boldsymbol{\gamma} + \alpha m_i(V_i)\}\right]^{\Delta_i} \exp\left\{-\int_0^{V_i} \exp\{\mathbf{W}_i^\top(t)\boldsymbol{\gamma} + \alpha m_i(t)\} \mathrm{d}\Lambda(t)\right\},$$

$$f(\mathbf{b}_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\Sigma_b|}} \exp\left\{-\frac{1}{2}\mathbf{b}_i^\top \Sigma_b^{-1} \mathbf{b}_i\right\}.$$

In light of the possibly multidimensional integral, direct maximization of (2.4) is quite difficult. Fortunately, the EM algorithm is ideally suited to this context and is thus employed to obtain the

MLE. Numerical integration techniques such as Laplace approximation, Gaussian quadrature and Monte Carlo methods can be applied in the E-step to evaluate the conditional expectations. In our setting, Gauss-Hermite quadrature is preferred due to its precision and computational speed. The likelihood approach results in a nonparametric maximum likelihood estimate (NPMLE) (**?**) $\hat{\Lambda}$ for the cumulative baseline hazard function. $\hat{\Lambda}$ is a function with positive jumps only at the observed survival times so that the dimension of $\hat{\Lambda}$ equals n_u, the number of unique uncensored event times. Details of the EM algorithm and the asymptotic properties of the MLEs are provided in **?** so we omit them and focus on the SE estimation methods. We further introduce the following notation for better illustration. During the E-step of the EM algorithm, we calculate

$$Q(\boldsymbol{\eta}', \boldsymbol{\eta}) = E\left[\log\{f_n(\mathbf{C}|\boldsymbol{\eta}')\}|\mathbf{O}, \boldsymbol{\eta}\right], \tag{2.6}$$

and in the M-step, $Q(\boldsymbol{\eta}', \boldsymbol{\eta})$ is maximized as a function of $\boldsymbol{\eta}'$ given $\boldsymbol{\eta}$. The EM update operator $M(\boldsymbol{\eta})$ can be expressed as

$$M(\boldsymbol{\eta}) = \underset{\boldsymbol{\eta}'}{\operatorname{argmax}}\, Q(\boldsymbol{\eta}', \boldsymbol{\eta}). \tag{2.7}$$

3. Standard error estimation

3.1 *Existing Methods: the Profile Likelihood and the Bootstrap*

The profile likelihood method proposed by **?** and **?** obtains the variance estimate for $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$ (the MLE for $\boldsymbol{\theta}$) by taking the second derivative of the profile likelihood function:

$$pl_n(\boldsymbol{\theta}) = \max_{\Lambda}\left\{\frac{1}{n}\log \mathcal{L}_n(\boldsymbol{\theta}, \Lambda|\mathbf{O})\right\}. \tag{3.1.1}$$

Let I_o be the observed-data information matrix for $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$, then:

$$I_o(i,j) \approx -\frac{pl_n(\boldsymbol{\theta}^* + h\mathbf{e}_i + h\mathbf{e}_j) - pl_n(\boldsymbol{\theta}^* + h\mathbf{e}_i) - pl_n(\boldsymbol{\theta}^* + h\mathbf{e}_j) + pl_n(\boldsymbol{\theta}^*)}{h^2}, \tag{3.1.2}$$

where $\mathbf{e}_i$ is the ith coordinate vector and $h > 0$ is the increment used to obtain a second-order numerical differentiation. It is important to note that the choice of h is subjective, with **?**

suggesting $h = O(\sqrt{1/n})$.

In contrast, the bootstrap SE estimation method is based on the idea of resampling full observational units. The observed data is denoted by $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_n)$ and the number of bootstrap samples by B. For $i = 1, 2, \cdots, B$, sample with replacement from $\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_n$ to form a new observed dataset $\mathbf{O}^{(i)} = (\mathbf{O}_{i1}, \mathbf{O}_{i2}, \cdots, \mathbf{O}_{in})$ and obtain the corresponding parameter estimate $\boldsymbol{\theta}^{(i)*}$ through the EM algorithm. The full covariance matrix and elementwise SE estimates of the parameters are then given by the analogous sample quantities for $\boldsymbol{\theta}^{(1)*}, \boldsymbol{\theta}^{(2)*}, \cdots, \boldsymbol{\theta}^{(B)*}$.

3.2 *PFDM and PREM: The FDM and REM Algorithms with a Profile Technique*

The FDM and REM algorithms introduced by **?** are built on ideas first presented in the SEM algorithm of **?**. These methods are all based on differentiation of the EM update operator (2.7) and seek to relate it to the asymptotic variance-covariance matrix. The FDM and REM algorithms avoid the outer layer of iterations required by the SEM algorithm by directly calculating the first derivative of the EM operator using two different numerical differentiation techniques. Each type of differentiation method, forward difference (FD) and Richardson extrapolation (RE), leads to its own corresponding SE estimation algorithm (FDM, REM).

If the FDM and REM algorithms are applied directly to the entire MLE vector $\boldsymbol{\eta}$ at $\boldsymbol{\eta}^* = (\boldsymbol{\theta}^*, \Lambda^*)$, then the resulting derivative of the EM operator $DM_{\boldsymbol{\eta}^*}$ would be a $(p + n_u) \times (p + n_u)$ matrix with the ith row calculated using either forward difference

$$DM_{\boldsymbol{\eta}^*}(i,) = \frac{M(\boldsymbol{\eta}^* + h\mathbf{e}_i) - M(\boldsymbol{\eta}^*)}{h} = \frac{M(\boldsymbol{\eta}^* + h\mathbf{e}_i) - \boldsymbol{\eta}^*}{h}, \tag{3.2.1}$$

or Richardson extrapolation

$$DM_{\boldsymbol{\eta}^*}(i,) = \frac{M(\boldsymbol{\eta}^* - 2h\mathbf{e}_i) - 8M(\boldsymbol{\eta}^* - h\mathbf{e}_i) + 8M(\boldsymbol{\eta}^* + h\mathbf{e}_i) - M(\boldsymbol{\eta}^* + 2h\mathbf{e}_i)}{12h}. \tag{3.2.2}$$

Since n_u (number of unique uncensored event times) is usually large, this makes the computation of $DM_{\boldsymbol{\eta}^*}$ slow despite the fact that our interest is only in the finite dimensional parameter $\boldsymbol{\theta}$. Moreover, although computing the derivative with respect to Λ is numerically feasible due to discretization, differentiation with respect to an infinite dimensional parameter is not suitably defined. Therefore, we now propose a new profile modification to the standard FDM and REM algorithms. In place of (3.2.1) and (3.2.2), our method instead evaluates $DM_{\boldsymbol{\theta}^*}$ (a $p \times p$ matrix), the derivative of the EM update operator with respect to $\boldsymbol{\theta}$ only. Let

$$\hat{\Lambda}(\boldsymbol{\theta}) = \underset{\Lambda}{\operatorname{argmax}} \log \mathcal{L}_n(\boldsymbol{\theta}, \Lambda | \mathbf{O}) \tag{3.2.3}$$

be the estimate of the nuisance parameter Λ given the parameter of interest $\boldsymbol{\theta}$. The derivative of the EM update operator at the MLE, $DM_{\boldsymbol{\theta}^*}$, can be obtained through the following algorithm. For illustration purposes, we present the algorithm using Richardson extrapolation (PREM). The corresponding PFDM algorithm can be derived similarly.

1. Let $\hat{\boldsymbol{\theta}}_1(i) = \boldsymbol{\theta}^* - 2h\mathbf{e}_i$, $\hat{\boldsymbol{\theta}}_2(i) = \boldsymbol{\theta}^* - h\mathbf{e}_i$, $\hat{\boldsymbol{\theta}}_3(i) = \boldsymbol{\theta}^* + h\mathbf{e}_i$ and $\hat{\boldsymbol{\theta}}_4(i) = \boldsymbol{\theta}^* + 2h\mathbf{e}_i$, obtain $\hat{\Lambda}(\hat{\boldsymbol{\theta}}_1(i))$, $\hat{\Lambda}(\hat{\boldsymbol{\theta}}_2(i))$, $\hat{\Lambda}(\hat{\boldsymbol{\theta}}_3(i))$ and $\hat{\Lambda}(\hat{\boldsymbol{\theta}}_4(i))$;

2. For $k = 1, 2, 3, 4$, treat $\hat{\boldsymbol{\eta}}_k(i) = \left(\hat{\boldsymbol{\theta}}_k(i), \hat{\Lambda}(\hat{\boldsymbol{\theta}}_k(i))\right)$ as the current estimate and run one iteration of the EM algorithm to obtain the updated estimate $\tilde{\boldsymbol{\theta}}_k(i)$ for $\boldsymbol{\theta}$;

3. Calculate the ith row of $DM_{\boldsymbol{\theta}^*}$:
$$DM_{\boldsymbol{\theta}^*}(i,) = \frac{\tilde{\boldsymbol{\theta}}_1(i) - 8\tilde{\boldsymbol{\theta}}_2(i) + 8\tilde{\boldsymbol{\theta}}_3(i) - \tilde{\boldsymbol{\theta}}_4(i)}{12h};$$

4. The asymptotic variance-covariance matrix can be obtained using the identity (**?**):
$$V_* = I_{oc}^{-1} + I_{oc}^{-1} DM_{\boldsymbol{\theta}^*}(I - DM_{\boldsymbol{\theta}^*})^{-1}, \tag{3.2.4}$$

5. If V_* is symmetric, set $V = V_*$. Otherwise, symmetrize the matrix by setting:
$$V = \frac{1}{2}\left(V_* + V_*^T\right).$$

Note that in step 4, I_{oc} is the conditional expectation of the complete-data information matrix given the observed-data evaluated at the MLE $\boldsymbol{\theta}^*$. Refer to Appendix A.1 of the Supplementary Materials for technical details about the computation of I_{oc}.

3.3 *PFDS and PRES: The FDS and RES Algorithms with a Profile Technique*

The central idea of the FDS and RES algorithms of **?** was first noted in **?**. The key idea is that the observed-data information matrix can be approximated by numerical differentiation of the Fisher score vector defined by (3.3.1). Let $D^{10}Q(\boldsymbol{\eta}', \boldsymbol{\eta})$ denote the first derivative of $Q(\boldsymbol{\eta}', \boldsymbol{\eta})$, defined by equation (2.6), as a function of its first argument $\boldsymbol{\eta}'$, i.e. $D^{10}Q(\boldsymbol{\eta}', \boldsymbol{\eta}) = \partial Q(\boldsymbol{\eta}', \boldsymbol{\eta})/\partial \boldsymbol{\eta}'$. The Fisher score vector is then defined as:

$$S(\boldsymbol{\eta}) = D^{10}Q(\boldsymbol{\eta}', \boldsymbol{\eta})|_{\boldsymbol{\eta}'=\boldsymbol{\eta}}, \tag{3.3.1}$$

and the observed-data information matrix I_o can be obtained by numerically differentiating $S(\boldsymbol{\eta})$ using forward difference (FDS) or Richardson extrapolation (RES). Again, numerically differentiating the entire parameter vector is both unnecessary and undesirable so a profile technique can be used to speed up the algorithm. The profile version of the $D^{10}Q$ function, denoted $D^{10}Q_{\boldsymbol{\theta}}(\boldsymbol{\eta}', \boldsymbol{\eta})$, and the corresponding profile Fisher score vector $S_{\boldsymbol{\theta}}(\boldsymbol{\eta})$ are given in Appendix A.2 of the Supplementary Materials. The detailed steps for the PRES algorithm are stated below and the PFDS algorithm is similar.

1. Same as the the PREM algorithm;

2. For $k = 1, 2, 3, 4$, let $\hat{\boldsymbol{\eta}}_k(i) = \left(\hat{\boldsymbol{\theta}}_k(i), \hat{\Lambda}(\hat{\boldsymbol{\theta}}_k(i))\right)$ and evaluate $S_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\eta}}_k(i)\right) = D^{10}Q\left(\hat{\boldsymbol{\eta}}_k(i), \hat{\boldsymbol{\eta}}_k(i)\right)$;

3. Calculate the ith row of I_o:

$$I_o(i,) = \frac{S_{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}_1(i)) - 8S_{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}_2(i)) + 8S_{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}_3(i)) - S_{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}}_4(i))}{12h}; \tag{3.3.2}$$

4. Symmetrize I_o as in the PREM algorithm.

3.4 *Comparison of the Methods: Implementation Considerations*

Each of the methods presented for obtaining SE estimates for joint modeling has its implementation trade-offs. In this section we compare implementation and computational difficulties as a precursor to the simulation study results of Section 4.1.

Out of all the methods to compute SEs, the bootstrap is the simplest to implement, requiring only trivial additional code beyond the code needed to fit the EM algorithm. Despite its simplicity, however, the bootstrap procedure requires running the full EM algorithm for each bootstrapped dataset. When the computation time to fit a single dataset is substantial, this can either severely limit the bootstrap sample size, or necessitate the use of parallel/batch computing for the bootstrap datasets. In addition, depending on the computational stability and implementation details of the EM algorithm used, convergence problems may arise for a subset of the bootstrap samples. It is also worth noting that multi-modality of the observed-data likelihood would be highly problematic for the bootstrap procedure, although in practice we have not found multi-modality to be a problem in the joint-modeling context presented in Section 2.

For the profile likelihood method, we need to compute the estimate of the nuisance parameter $\hat{\Lambda}(\boldsymbol{\theta})$ when fixing the parameter of interest $\boldsymbol{\theta}$, as defined by (3.2.3). Unfortunately, $\hat{\Lambda}(\boldsymbol{\theta})$ cannot be calculated via direct maximization. Therefore, in addition to the original EM algorithm for parameter estimation, another EM algorithm is required to obtain $\hat{\Lambda}(\boldsymbol{\theta})$. This algorithm, which is abbreviated PEME (Partial Expectation, Maximization and Evaluation) was presented in **?**. With the PEME algorithm, we can apply (3.1.2) to obtain the observed-data information matrix I_o. I_o is a $p \times p$ symmetric matrix, so the number of entries that need calculation is $\frac{1}{2}p(p+1)$ and the calculation of each entry requires evaluating the profile likelihood function at different $\boldsymbol{\theta}$'s.

For the PFDM and PREM methods, we also need the PEME algorithm since a profile technique is proposed in their application. Moreover, code for computing I_{oc} (as defined in Section 3.2) is required. Fortunately, I_{oc} is only calculated once and these methods actually save computation time compared to the profile likelihood method due to how the $DM_{\boldsymbol{\theta}^*}$ matrix is evaluated. Although $DM_{\boldsymbol{\theta}^*}$ is a $p \times p$ matrix as I_o, it can be evaluated row by row rather than entry by entry as shown by (3.2.1) and (3.2.2). Notice that, despite the saving in computation time, the $DM_{\boldsymbol{\theta}^*}$ matrix may not be symmetric which would lead to asymmetry in our target variance-covariance matrix as indicated by (3.2.4). This asymmetry is a result of the numerical approximation used to compute $DM_{\boldsymbol{\theta}^*}$, and is why a symmetrization step is added in step 4 of Section 3.2.

Finally, for the PFDS and PRES methods, again the PEME algorithm is required. Furthermore, these methods need the code for the Fisher score vector $S_{\boldsymbol{\theta}}(\boldsymbol{\eta})$. Then the observed-data information matrix I_o can be obtained row by row instead of entry by entry as shown by (3.3.2). This row-vectorization makes the PFDS and PRES methods much faster than the profile likelihood method.

Table 1 summarizes the above discussion. One thing that should be pointed out explicitly is that, unlike that bootstrap method, the profile likelihood, PFDM, PREM, PFDS and PRES methods all utilize numerical differentiation which introduces numerical error. As a result, these methods are not guaranteed to provide us with valid positive-definite variance-covariance matrix estimates. In particular, it is possible for the diagonal entries of the resulting variance-covariance matrix to be negative. In contrast, the bootstrap is subject to resampling error but has the advantage of ensuring both a non-negative-definite full covariance matrix as well as non-negative SE estimates for all parameters.

4. Simulation study and substantive data application

4.1 *Simulation study*

Consider two time-independent covariates X_{1i} and X_{2i} $(i = 1, 2, \cdots, n)$, where X_{1i} is a binary covariate from the Bernoulli distribution with success probability $\frac{1}{2}$ and X_{2i} is a continuous covariate from the uniform distribution $\mathcal{U}(0,1)$. The longitudinal model is

$$Y_{ij} = Y_i(t_{ij}) = m_i(t_{ij}) + e_{ij} = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 t_{ij} + \beta_4 X_{1i} t_{ij} + \beta_5 X_{2i} t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij},$$

with $\mathbf{b}_i = (b_{1i}, b_{2i})^\top \sim \mathcal{N}(0, \Sigma_b)$, $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ and $t_{ij} = 0.25(j-1)$ for $i = 1, 2, \ldots, n$. The hazard function for the survival time is

$$\lambda(t|\mathbf{b}_i, X_{1i}, X_{2i}) = \lambda(t) \exp\left\{\gamma_1 X_{1i} + \gamma_2 X_{2i} + \alpha m_i(t)\right\}.$$

The true values of the parameters are chosen as below. Case II is considered in addition to Case I to explore the effect of magnified measurement error to the SE estimation methods.

I: $\beta_1 = -1.0,\ \beta_2 = -1.5,\ \beta_3 = 1.0,\ \beta_4 = -0.5,\ \beta_5 = 0.5, \Sigma_b = \begin{pmatrix} 0.5 & -0.1 \\ -0.1 & 0.16 \end{pmatrix}$,
$\gamma_1 = -0.5,\ \gamma_2 = 1.5,\ \alpha = 0.5,\ \sigma_e^2 = 0.1.$

II: $\sigma_e^2 = 0.3$ and the other parameters remain the same as in case I.

The true baseline hazard function is given below which starts high and gradually decreases, then stays at certain level for some period of time and finally goes up:

$$\lambda(t) = \begin{cases} \exp\{-0.3t\}, & \text{if } 0 < t \leqslant 1 \\ \exp\{-0.3\}, & \text{if } 1 < t \leqslant 2.5 \\ \exp\{0.3(t-3.5)\}. & \text{if } t > 2.5 \end{cases}$$

The censoring time is simulated from an exponential distribution with mean 2.5 which yields a censoring proportion of approximately 30%. Moreover, the average number of longitudinal measurements is 3.5 for each subject. The simulation is repeated 500 times with sample size $n = 200$. Table 2 shows the results of our joint model with completely unspecified baseline hazard function along with those of the joint models with piecewise constant and spline-based baseline hazard functions fitted by the `JM` package, where the follow-up period is divided into 7 intervals with knots placed at the corresponding quantiles. The "MCSE" row in Table 2, which are the empirical standard errors of the parameter estimates from the 500 simulations, serves as the benchmark of comparison with the SE estimation methods. The corresponding SE estimates from all SE estimation methods are reported in Table 3 (for case I) which include results using different choices of h's. Due to limited space, the results for $h = 10^{-2}$ and $h = 10^{-5}$ are presented while additional results for $h = 10^{-3}$, $h = 10^{-4}$ and those for case II are available in the supplementary materials (web Tables 1-2). The purpose of choosing different h's is to illustrate the effect of the choice of h on the SE estimation procedures. Moreover, the bootstrap method is implemented with $B = 50$ and $B = 100$ bootstrap samples. As illustrated in Table 3 the computation time for the bootstrap greatly exceeds that of the other methods, and we restrict to $B \leqslant 100$ to ensure comparability and avoid prohibitive runtimes. For case I, we construct the 95% confidence intervals using the corresponding SE estimates and present the empirical coverage of each method in Figure 1. Another simulation setting where the longitudinal and survival processes only share the same random effects (as stated by (2.3)) is presented in Appendix A.3 of the Supplementary Materials.

4.2 *Discussion of the simulation results*

Comparing the results of case II with those of case I, we observe that increasing the measurement error increases the empirical standard error (Monte Carlo Standard Error, "MCSE") of all the parameters, as expected. Moreover, with greater measurement error, the EM algorithm takes longer to converge (on average, it takes 45.50 steps in case II while only 28.47 steps in case I).

As illustrated by Table 2, comparing with the semiparametric joint model, the joint models with piecewise constant and spline-based baseline hazard yield satisfactory parameter estimates but the biases for α are 0.02306 and 0.02803, respectively, which are 188 % and 228 % of 0.01229, the bias under our semiparametric approach. For the standard error estimates, they are not quite comparable since different models are applied. This simulation demonstrates that parametric approaches could provide good approximations but the biases in the estimates could be improved by a nonparametric approach.

From Table 3 and web Tables 1-2, we observe that, by and large, the new approaches

(PFDM/PREM, PFDS/PRES) yield comparable SE estimates with the profile likelihood method. In the following, we discuss the performance of the methods in detail. First, for the choice of numerical differentiation approach, although the Richardson extrapolation is about four times slower than the forward difference, it is more stable in two aspects: 1. methods using the Richardson extrapolation (PREM and PRES) are more likely to produce positive variance estimates compared to those using the forward difference (PFDM and PFDS) as the "N. Posit." (number of positive variance estimates out of the 500 simulations) show, especially when the measurement error is large and h is small; 2. methods using the Richardson extrapolation are less sensitive to the choice of h (comparing the results from "$h = 10^{-2}$" with those from "$h = 10^{-5}$"). Therefore, our profile method using Richardson extrapolation is typically preferred for the stability of the SE estimates, despite a slight sacrifice in computation time.

Second, for the choice between PREM and PRES, although they yield almost identical results, PRES is in general preferred since numerically differentiating the Fisher score vector is more straightforward than numerically differentiating the EM operator and it avoids the problem of PREM that the error would be magnified when the EM algorithm is slow. From the perspective of writing computer code, PRES is also preferred. PRES only requires code for the Fisher score vector which relates to the first derivative of the complete-data log- likelihood while PREM requires code for I_{oc} which relates to the second derivative of the complete-data log-likelihood.

Third, the choice between PRES and PL is considered. The profile likelihood method appears to be more sensitive to the choice of h and may lead to negative or underestimation of the SE when h is small (by comparing the "PL($h = 10^{-2}$)" and "PL($h = 10^{-5}$)" results), particularly for the survival regression parameters (γ_1, γ_2 and α). Moreover, PRES is computationally faster than PL, particularly when the sample size of the data or the number of parameters grows larger. Hence, the PRES is also considered to outperform the profile likelihood method.

To be complete, results from the bootstrap method are also provided. The computation time of the bootstrap method for case II is much greater than that for case I, e.g. for $B = 50$, the average computation time for case I is 4702 seconds while that for case II is 8456 seconds. The reason is that the EM algorithm takes more iterations to converge under greater measurement error. Therefore, the bootstrap procedure suffers from an unappealing property that the computation time depends on the convergence speed of the EM algorithm. Moreover, the results of the additional simulation study show (in web Table 3) that the Bootstrap method seems to overestimate the standard errors of the survival regression parameters ($\gamma_1, \cdots, \gamma_4$, and α). This is possibly due to the bootstrap resampling (with replacement) scheme which leads to resampled covariates that have slightly smaller variations than the original covariates and hence slightly larger SE estimates. A similar phenomenon was observed in **?** and explained on page 1041. The key explanation, like in standard experimental designs, is that a design with larger variations leads to better precision in the estimation of regression coefficients. We would like to make an additional remark which is not displayed explicitly in the tables that a subset of the bootstrap samples may be subject to convergence problems. For case I, the convergence problem is negligible while for case II, 1.6% (averaging over the 500 simulations) of the bootstrap samples fail to converge.

In terms of coverage properties, the performance of all the SE estimation methods are illustrated in (a) and (b) of Figure 1. PREM and PRES are again shown to provide more accurate SE estimates and are more stable under different choices of h since they display smaller differences in coverage ratios to the theoretical true value than the other methods. As a result, we would like to recommend the PRES method as the best choice for the purpose of SE estimation.

4.3 *HIV clinical trial data analysis*

To verify that the recommended PRES method can be applied practically, we now present a substantive data example. This set of HIV clinical trial data is originated from **?** and has been analyzed by **?** as an illustrative example. The clinical trial was called a ddI/ddC study which is aimed to compare the clinical efficacy and safety of two drugs, namely ddI (didanosine) and ddC (zalcitabine), in HIV infected patients intolerant or failing ZDV (zidovudine) therapy. There were 467 patients enrolled in the study with 230 of them randomized to receive ddI and 237 to ddC. The average follow-up time was 15.6 months and CD4 lymphocyte counts were recorded at study entry and at the 2-, 6-, 12- and 18-month visits (measurements may be missing due to patients' condition). 279 patients had not experienced death by the end of the study, resulting in approximately 60% right-censoring. The data is open to access in `JM`.

Figure 2 displays the cross-sectional mean curves of $Y_{ij} = \sqrt{CD4}$ (a square root transformation is put on the CD4 counts due to its right skewness) for the ddI and ddC treatment groups. Based on the patterns shown in the figure, a quadratic trend over time is included in our model for Y_{ij} while **?** assumed a linear trend over time. We also fitted the same model for Y_{ij} as **?** with the model and results provided in Appendix A.4 and Table 4 of the Supplementary Materials, respectively. Moreover, **?** opted for approximating $\lambda(t)$ with a piecewise constant function due to the underestimation problem of the SEs if $\lambda(t)$ is left completely unspecified. Since our methodology can handle this case, we opt for the semi-parametric model. Now, after proposing the new SE estimation approaches, we apply the most recommended PRES method to obtain the SE estimates while leaving $\lambda(t)$ completely unspecified. The model fitted is presented below:

$$Y_{ij} = m_i(t_{ij}) + e_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \mathrm{drug}_i t_{ij} + \beta_3 t_{ij}^2 + \beta_4 \mathrm{drug}_i t_{ij}^2 + b_{1i} + b_{2i} t_{ij} + e_{ij},$$

$$\lambda(t|\mathbf{b}_i, \mathrm{drug}_i) = \lambda(t) \exp\{\gamma \mathrm{drug}_i + \alpha m_i(t)\}.$$

The results are provided in Table 4. The p-value for α is very small (< 0.0001) which suggests that the CD4 lymphocyte count is an important covariate in the survival model. Moreover, the treatment effect on the CD4 counts seems to be moderately significant since the p-value for β_2 and β_4 are 0.0640 and 0.0862, respectively. This point is different with that from **?** where the treatment had little effect on the CD4 counts. Therefore, according to our results, the CD4 counts satisfy the first two criteria to be an adequate surrogate marker. However, with the CD4 counts in the survival model, the treatment effect is still statistically significant (since the p-value of γ is 0.0044). Hence, we reach the conclusion that CD4 count is not a useful surrogate marker for these patients.

5. Discussion and Conclusion

In this paper, we we have proposed two new SE estimation methods when using the EM algorithm in a semiparametric joint modeling setting by applying a profile technique to overcome the challenges of high-dimensional parameters brought upon by the nonparametric component. The performance of these methods are examined systematically through simulation studies. Simulation results verify that these methods produce accurate SE estimates and the PRES method is recommended as the best choice. We hope that the ability to rapidly obtain reliable SE estimates with high- or infinite-dimensional hazard functions can expand the types of models applied in practice. The HIV clinical trial data analysis shows that the PRES method also performs well when analyzing a realistically sized substantive dataset.

Finally, we would like to make a concluding remark that the efficient procedures to estimate the SEs of the parameters of interest introduced here are applicable whenever the EM algorithm is used and there exists a high or infinite dimensional nuisance parameter. Although the proposed methods are illustrated through the semiparametric joint modeling setting in this paper, their applications are potentially quite extensive. For instance, they can be readily implemented in settings with more complicated censoring scheme or in models other than the Cox model for survival time or the linear mixed effects model for longitudinal measurements. Even more generally, the same ideas can be extended beyond the joint survival-longitudinal modeling context. While we have focused on the SE estimation of the finite dimension parameter, our approach could be employed to estimate the SE of the cumulative baseline hard function at a computational cost.

Software

The simulation studies and substantive data analysis are implemented in R. Software in the form of R code is available on request from Cong Xu (cgxu@ucdavis.edu).

Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The research of Jane-Ling Wang was supported by the National Science Foundation (DMS-09006813) and the National Institutes of Health (1R01 AG025218-01A2) . The research of Cong Xu is supported by the same NIH grant. *Conflict of Interest*: None declared.

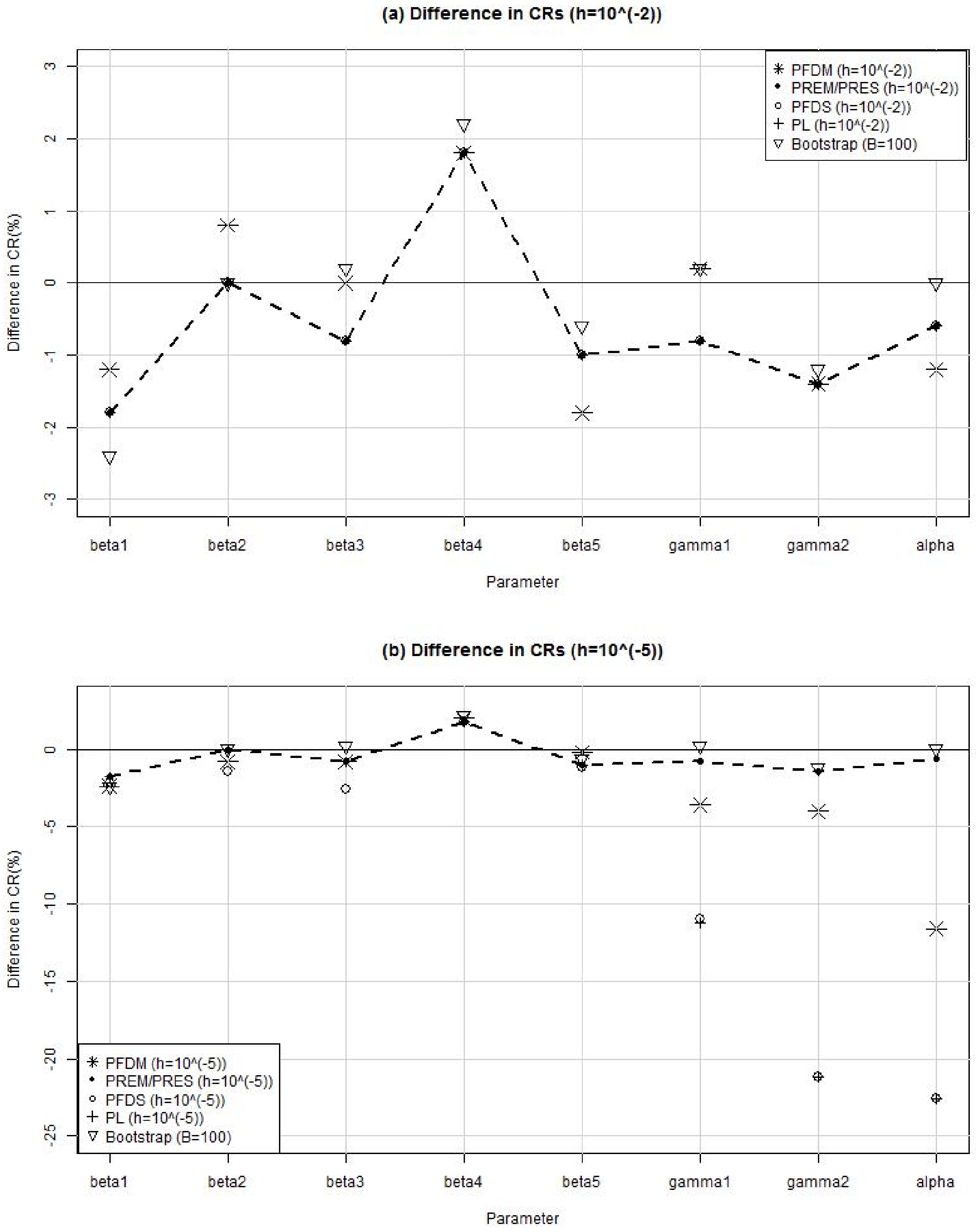

Fig. 1. The CRs (coverage ratios) of the 95% confidence interval (obtained using the SE estimates from each method) minus 95% for case I with (a) $h = 10^{-2}$; (b) $h = 10^{-5}$. The dashed lines are lines which connects the points for PREM/PRES methods. Some of the points are overlapped.

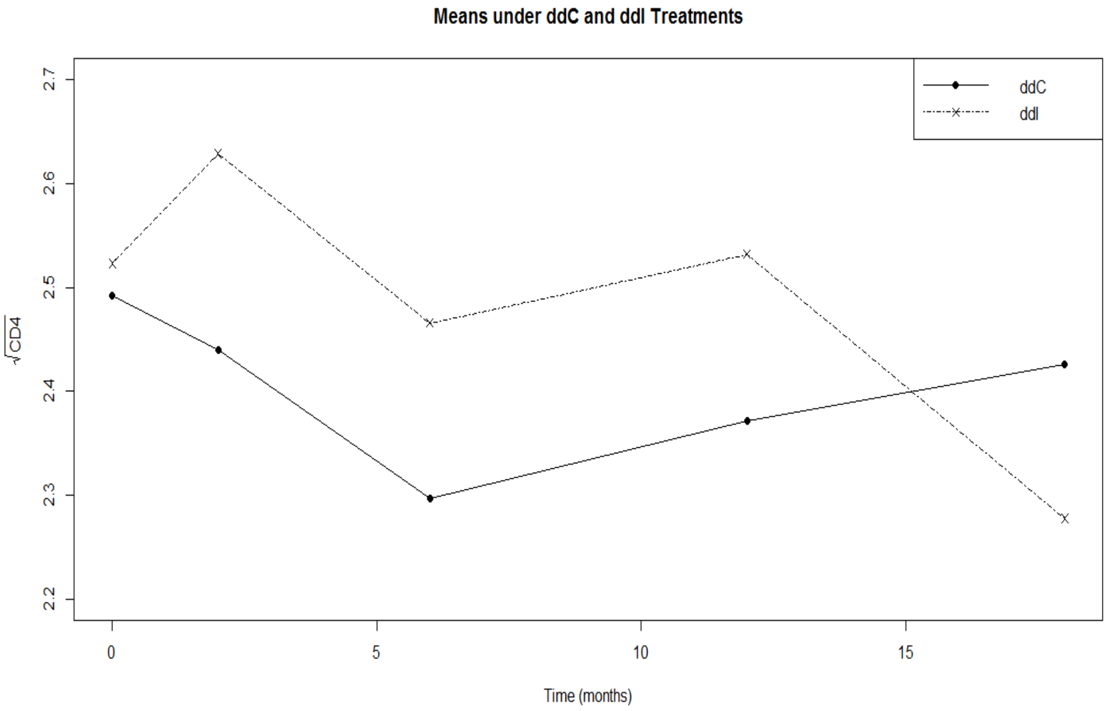

Fig. 2. Time plot that displays the cross-sectional mean curves of sqrt(CD4) for the ddC and ddI treatment groups.

Table 1. *Summary of the SE estimation methods. p is the length of the finite dimensional parameter $\boldsymbol{\theta}$; B is the number of Bootstrap samples; m is the average number of iterations for the EM algorithm to converge.*

Method	Calculate	Calculate	Evaluate	Computation
	I_{oc}	$S_{\boldsymbol{\theta}}(\boldsymbol{\eta})$	$pl_n(\boldsymbol{\theta})$	Demand
PFDM/PREM	$\surd$	×	×	$O(p)$
PFDS/PRES	×	$\surd$	×	$O(p)$
Profile Lik.	×	×	$\surd$	$O(p^2)$
Bootstrap	×	×	×	$O(Bm)$

Table 2. *Estimates from the EM algorithm for case I in Section 4.1. θ_0: true value of the parameters. "Mean" and "MCSE": empirical means and standard errors of the parameter estimates from the 500 simulations. "CR": coverage ratios of the 95% confidence interval constructed using the "MCSE" column.*

θ	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	α
θ_0	-1.0	-1.5	1.0	-0.5	0.5	-0.5	1.5	0.5
With Unspecified Baseline Hazard Function								
Mean	-0.99922	-1.50637	1.00145	-0.50416	0.49678	-0.49800	1.54721	0.51229
Bias	0.00078	0.00637	0.00145	0.00416	0.00322	0.00200	0.04721	0.01229
MCSE	0.09939	0.11760	0.12354	0.10917	0.18441	0.24130	0.37139	0.13989
CR	0.942	0.950	0.952	0.958	0.940	0.948	0.940	0.950
With Piecewise Constant Baseline Hazard Function								
Mean	-0.99920	-1.50711	1.00316	-0.50514	0.49655	-0.49118	1.56189	0.52306
Bias	0.00080	0.00711	0.00316	0.00514	0.00345	0.00882	0.06189	0.02306
MCSE	0.09942	0.11758	0.12367	0.10918	0.18429	0.23876	0.36891	0.13948
CR	0.942	0.950	0.952	0.958	0.940	0.950	0.944	0.948
With Spline-based Baseline Hazard Function								
Mean	-0.99841	-1.50683	1.00594	-0.50420	0.49653	-0.49117	1.58066	0.52803
Bias	0.00159	0.00683	0.00594	0.00420	0.00347	0.00883	0.08066	0.02803
MCSE	0.09940	0.11730	0.12325	0.10906	0.18395	0.24321	0.37447	0.14406
CR	0.942	0.950	0.952	0.960	0.942	0.956	0.950	0.960

Table 3. *SE estimates for case I in Section 4.1. "PFDM/PREM": numerically differentiate the EM update operator with forward difference/Richardson extrapolation. "PFDS/PRES": numerically differentiate the Fisher score vector with forward difference/Richardson extrapolation. "PL": profile likelihood method. "MCSE": empirical standard deviations of the parameter estimates from the 500 simulations; "N. Posit." row: number of positive variance estimates out of the 500 simulations for each method; "C. Time" row: average computation time (in seconds) for each method.*

θ	MCSE	PFDM	PREM	PFDS	PRES	PL	Bootstrap
		$(h = 10^{-2})$	$(h = 10^{-2})$	$(h = 10^{-2})$	$(h = 10^{-2})$	$(h = 10^{-2})$	$(B = 50)$
β_1	0.09939	0.09675	0.09529	0.09527	0.09527	0.09528	0.09532
β_2	0.11760	0.12278	0.11908	0.11826	0.11821	0.11821	0.11947
β_3	0.12354	0.12251	0.12007	0.11979	0.11968	0.11990	0.12266
β_4	0.10917	0.11591	0.11498	0.11487	0.11482	0.11490	0.11753
β_5	0.18441	0.17368	0.18310	0.18312	0.18311	0.18318	0.18727
γ_1	0.24130	0.24637	0.24144	0.23871	0.23849	0.23784	0.24470
γ_2	0.37139	0.36045	0.35430	0.35327	0.35264	0.35208	0.36223
α	0.13989	0.13208	0.13595	0.13561	0.13585	0.13508	0.14017
N.	Posit.	243	500	500	500	500	500
C.	Time	56	213	35	131	178	4702
θ	MCSE	PFDM	PREM	PFDS	PRES	PL	Bootstrap
		$(h = 10^{-5})$	$(h = 10^{-5})$	$(h = 10^{-5})$	$(h = 10^{-5})$	$(h = 10^{-5})$	$(B = 100)$
β_1	0.09939	0.09426	0.09529	0.09507	0.09527	0.09524	0.09471
β_2	0.11760	0.11236	0.11908	0.10917	0.11821	0.11796	0.11881
β_3	0.12354	0.11999	0.12008	0.11065	0.11968	0.11963	0.12294
β_4	0.10917	0.11640	0.11498	0.11360	0.11482	0.11476	0.11820
β_5	0.18441	0.19045	0.18310	0.17709	0.18311	0.18073	0.18778
γ_1	0.24130	0.20927	0.24144	0.16850	0.23849	0.16789	0.24645
γ_2	0.37139	0.31366	0.35430	0.20435	0.35264	0.20523	0.36383
α	0.13989	0.10381	0.13594	0.08203	0.13585	0.08213	0.14048
N.	Posit.	500	500	490	500	499	500
C.	Time	56	213	35	131	178	9175

Table 4. *Results for the HIV clinical trial data analysis. "Est. SE" are the estimated SEs obtained using the PRES method with $h = 10^{-5}$.*

θ	β_0	β_1	β_2	β_3	β_4	γ	α
Est. Value	2.5210	-0.0582	0.0013	0.0251	-0.0016	0.5137	-1.0631
Est. SE	0.04315	0.00992	0.00074	0.01323	0.00096	0.18059	0.11531
p-value	< 0.0001	< 0.0001	0.0640	0.0580	0.0862	0.0044	< 0.0001