mathematics

MDPI

Article

Stochastic EM Algorithm for Joint Model of Logistic Regression and Mechanistic Nonlinear Model in Longitudinal Studies

Hongbin Zhang

Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY 40536, USA; hongbin.zhang@uky.edu

Abstract: We study a joint model where logistic regression is applied to binary longitudinal data with a mismeasured time-varying covariate that is modeled using a mechanistic nonlinear model. Multiple random effects are necessary to characterize the trajectories of the covariate and the response variable, leading to a high dimensional integral in the likelihood. To account for the computational challenge, we propose a stochastic expectation-maximization (StEM) algorithm with a Gibbs sampler coupled with Metropolis–Hastings sampling for the inference. In contrast with previous developments, this algorithm uses single imputation of the missing data during the Monte Carlo procedure, substantially increasing the computing speed. Through simulation, we assess the algorithm's convergence and compare the algorithm with more classical approaches for handling measurement errors. We also conduct a real-world data analysis to gain insights into the association between CD4 count and viral load during HIV treatment.

Keywords: logistic regression; longitudinal binary data; measurement error; time-varying covariate; mechanistc nonelinear model; stochastic EM; random effects; Gibbs sampler; Metropolis–Hastings sampling; HIV treatment

MSC: 62P10

Citation: Zhang, H. Stochastic EM Algorithm for Joint Model of Logistic Regression and Mechanistic Nonlinear Model in Longitudinal Studies. *Mathematics* **2023**, *11*, 2317. https://doi.org/10.3390/math11102317

Academic Editor: Konstantin Kozlov

Received: 28 March 2023
Revised: 12 May 2023
Accepted: 15 May 2023
Published: 16 May 2023

1. Introduction

In medical and public health studies, binary data are frequently employed to capture disease status, treatment endpoints, and other dichotomous outcomes. When data are clustered, e.g., in longitudinal studies, the correlation among the repeatedly measured outcomes must be accounted for to avoid biased inference [1]. Logistic regression with random effects is a flexible analytic approach to analyze correlated binary data [2]. A potential difficulty in making inferences in such models is that the likelihood analysis is burdened by intractable integration. When the dimension of the random effects is high, numerical methods, e.g., Gauss–Hermite Quadrature [3], can be prohibitively intensive for the computation.

Another common problem in analyzing correlated data is the presence of covariates subject to measurement error [4]. Measurement error with independent observations has been extensively reviewed in the literature [5]. The random effects logistic regression model with time-constant covariates subject to measurement error has been also studied (see, e.g., [4]). Ref. [6] tackled the problem when one mismeasured covariate is time-varying. The authors proposed a maximum likelihood (ML) approach where a joint model framework was used to simultaneously infer the covariate and response processes. Motivated by the real-world example, a mechanistic nonlinear mixed effects model was used for the measurement error model. The Monte Carlo expectation and maximization (MCEM) implementation was based on samples from the conditional density, requiring Markov chain Monte Carlo (MCMC) procedures [7]. However, the choice of replicate size is the central issue in guaranteeing convergence, and this remains an open problem. In addition, the algorithm is very slow to converge. These issues can be even more severe

for the type of joint model we are considering as the structure of random effects can be complex, leading to a high dimensional integral.

As an alternative to address both the point-wise convergence and the computational problems related to MCEM, we consider the stochastic EM (StEM) algorithm [8]. The algorithm replaces the E-step by approximating the expectation with only one realization of the missing data at each iteration, thereby significantly reducing the computational burden. Recently, the algorithm has been expanded to the standard joint model focusing on time-to-event outcomes with mismeasured covariates [9]. The algorithm's feasibility and performance in our joint model have yet to be established.

The first objective of the present paper (Section 2) is, thus, to extend the StEM algorithm for the joint model of random effects logistic regression and mechanistic nonlinear random effects model. We use the Gibbs sampler to alternatively simulate the (lower dimension) full conditionals of the random effects structure. We approximate the full conditionals with instrumental distributions by carefully implementing the Metropolis–Hastings sampling. The second objective of this paper (Section 3) is to assess the convergence of the StEM algorithm and compare the algorithm with more classical approaches to handle measurement error, such as the two-step method. We also illustrate the algorithm with a real-world study in HIV dynamics where we explore, e.g., the relationship between CD4 count and viral load during antiretroviral treatment (Section 4).

2. Joint Model of Logistic Regression and Mechanistic Nonlinear Model with Random Effects

2.1. Model Specification

For individual i, $i = 1, \ldots, n$, consider two longitudinal processes $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$, measured at t_{ij} $j = 1, \ldots, n_i$, for the covariate and response, respectively. We assume that $\boldsymbol{x}_i$ are continuous and $\boldsymbol{y}_i$ are discrete (e.g., binary) measurements.

We consider a mechanistic nonlinear random effects model for the covariate process such that

$$x_{ij} = h(t_{ij}, \boldsymbol{a}_i) + e_{ij} \equiv x^*_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \quad \boldsymbol{a}_i \sim N(\boldsymbol{\alpha}, \mathbf{A}), \tag{1}$$

where $h(\cdot)$ is a known nonlinear function, $\boldsymbol{a}_i$ is a $p \times 1$ vector of random effects centered at $\boldsymbol{\alpha}$ (the fixed effects), $\mathbf{A}$ is the variance–covariance matrix, and $e_{ij}{}'s$ are measurement errors. Note that this specification may be viewed as a classical measurement error model where $\boldsymbol{x}^*_{\boldsymbol{i}} = (x^*_{i1}, \ldots, x^*_{in_i})^T$ is the unobserved true value of the error-prone covariate $\boldsymbol{x}_i$.

For the response process, we consider random effects logistic regression as

$$E(y_{ij}) = g(x^*_{ij}, t_{ij}, \boldsymbol{b}_i), \quad \boldsymbol{b}_i \sim N(\boldsymbol{\beta}, \mathbf{B}), \tag{2}$$

where $g(\cdot)$ is the inverse of the logit link function, $\boldsymbol{b}_i$ is a $q \times 1$ vector of random effects centered at $\boldsymbol{\beta}$, and $\mathbf{B}$ is a variance–covariance matrix. For simplicity, we suppress the error-free covariates in the model since no distributional assumptions are needed for these covariates. We assume that conditioning on the random effects $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, $y_{ij}{}'s$ are independent and follow a Bernoulli distribution. Model (2) extends the standard random effects logistic regression in the sense of including the unknown quantities x^*_{ij} in the regressors. We let x_{ij} be the current covariate value for individual i at measurement j, but the model can be easily extended to the case where x_{ij} is a summary of the covariate history values up to measurement j, such as a cumulative average.

For the joint model of (1) and (2), letting $\boldsymbol{\theta}$ be the collection of all parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \mathbf{A}, \mathbf{B}\}$, we have the joint log-likelihood for the observed data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n\}$ as

$$l_o(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{ \int \int f(\boldsymbol{y}_i | \boldsymbol{x}^*_i, \boldsymbol{b}_i) f(\boldsymbol{x}_i | \boldsymbol{x}^*_i; \sigma^2) f(\boldsymbol{a}_i; \boldsymbol{\alpha}, \mathbf{A}) f(\boldsymbol{b}_i; \boldsymbol{\beta}, \mathbf{B}) d\boldsymbol{a}_i d\boldsymbol{b}_i \right\} \tag{3}$$

where

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i^*, \boldsymbol{b}_i) = \prod_{j=1}^{m_i} f(y_{ij}|x_{ij}^*, \boldsymbol{b}_i) = \prod_{j=1}^{m_i} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}},$$

(with $\mu_{ij} = E(y_{ij})$ as in (2)),

$$\begin{aligned} f(\boldsymbol{x}_i|\boldsymbol{x}_i^*; \sigma^2) &= \prod_{j=1}^{n_i} f(x_{ij}|x_{ij}^*; \sigma^2) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}[x_{ij} - x_{ij}^*]^2/\sigma^2\}, \\ f(\boldsymbol{a}_i; \boldsymbol{\alpha}, \mathbf{A}) &= (2\pi)^{-p/2} |\mathbf{A}|^{-1/2} \exp\{-(\boldsymbol{a}_i - \boldsymbol{\alpha})^T \mathbf{A}^{-1} (\boldsymbol{a}_i - \boldsymbol{\alpha})/2\}, \\ f(\boldsymbol{b}_i; \boldsymbol{\beta}, \mathbf{B}) &= (2\pi)^{-q/2} |\mathbf{B}|^{-1/2} \exp\{-(\boldsymbol{b}_i - \boldsymbol{\beta})^T \mathbf{B}^{-1} (\boldsymbol{b}_i - \boldsymbol{\beta})/2\}. \end{aligned}$$

This likelihood function does not have a closed-form expression since the functions in the integral can be nonlinear in both the random effects $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. Direct maximization of the likelihood is challenging as it will require numerical evaluation of an integral whose dimension is equal to the dimension of the random effects $\dim(\boldsymbol{a}_i) + \dim(\boldsymbol{b}_i) = p + q$. We resort to the EM algorithms by considering the unobserved random effects $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ as "missing data". For the "complete data" $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{a}_i, \boldsymbol{b}_i), i = 1, \ldots, n\}$, we have the complete data likelihood function for individual i as

$$L_c^{(i)}(\boldsymbol{\theta}|(\boldsymbol{a}_i, \boldsymbol{b}_i)) = f(\boldsymbol{y}_i|\boldsymbol{x}_i^*, \boldsymbol{b}_i; \boldsymbol{\beta}) \times f(\boldsymbol{x}_i|\boldsymbol{x}_i^*; \sigma^2) \times f(\boldsymbol{a}_i; \mathbf{A}) \times f(\boldsymbol{b}_i; \mathbf{B}). \tag{4}$$

2.2. Estimation Procedure

The EM algorithm introduced by [10] is a classical approach to estimate parameters of models with non-observed or incomplete data. Let us briefly cover the principle. Denote $\mathbf{z}$ as the vector of non-observed data, $(\mathbf{y}, \mathbf{z})$ as the complete data and $L_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$ as the log-likelihood of the complete data; the EM algorithm maximizes the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = E(L_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta})|\mathbf{y}; \boldsymbol{\theta}')$ function in two steps. At the k^{th} iteration, the E-step is the evaluation of $Q^{(k)}(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)})$, where the M-step updates $\boldsymbol{\theta}^{(k-1)}$ by maximizing $Q^{(k)}(\boldsymbol{\theta})$.

For cases where the E-step has no analytic form, Ref. [7] introduced the MCEM algorithm, which calculates the conditional expectations at the E-step via many simulations within each iteration and, hence, is quite computationally intensive. The choice of replicate size is the central issue in guaranteeing convergence. Ref. [8] introduced a stochastic version of the EM algorithm, namely, the StEM, which replaces the E-step with a single imputation of the complete data and then averages the last batch of M estimates in the Markov chain iterative sequence to obtain the point estimate of the parameters. The imputed data $\mathbf{z}^{(k)}$ at the k^{th} iteration are a random draw from the conditional distribution of the missing data given the observed data and the estimated parameter values at the $(k-1)^{th}$ iteration, $f(\mathbf{z}^{(k)}|\mathbf{y}, \boldsymbol{\theta}^{(k-1)})$. As $\mathbf{z}^{(k)}$ only depends on $\mathbf{z}^{(k-1)}$, $\{\mathbf{z}^{(k)}\}_{k\geq 1}$ is a Markov chain. Assuming that $\mathbf{z}^{(k)}$ takes values in a compact space and the kernel of the Markov chain is positive continuous for a Lebesgue measure, the Markov chain is ergodic, and that ensures the existence of a unique stationary distribution (see, e.g., [11,12]).

We will now describe the estimation procedure for our joint model. For the imputation step in StEM, we use the Gibbs sampler to break down our joint model's high dimension random effects structure into lower dimension full conditionals. Sampling from the full conditionals of the random effects can be accomplished using a multivariate rejection algorithm (see, e.g., [13,14]). The samples are independent and identically distributed; however, the computation can be intensive. For example, a large size of simulations within each iteration may be required to obtain an accepted sample, especially for individuals whose random effects distribution is skewed.

To avoid simulation from the full conditional distributions directly, we use a hybrid Gibbs sampler, also called "Metropolis-within-Gibbs", which consists of substituting simulations from the full conditional distributions of a Gibbs sampler with simulations from an instrumental distribution. Ref. [15] discussed such a method in the context of complete data nonlinear mixed models, which can be extended to our models as follows.

Regarding the proposal distribution for the full conditionals under the Metropolis–Hastings (M–H) sampling, we consider a marginal distribution. Specifically, at the k^{th}, $k = 1, \ldots,$ iteration, we update $\boldsymbol{a}_i^{(k)}$ as follows ($\boldsymbol{b}_i^{(k)}$ is updated similarly in the alternative step under the Gibbs sampler):

Step 1: simulate $\boldsymbol{a}_i^* \sim N(\boldsymbol{\alpha}^{(k-1)}, \mathbf{A}^{(k-1)})$, and, independently, sample η from the uniform (0, 1) distribution;

Step 2: calculate $\rho = \dfrac{L_c^{(i)}\left(\boldsymbol{\theta}^{(k-1)} \middle| \left(\boldsymbol{a}_i^*, \boldsymbol{b}_i^{(k-1)}\right)\right)}{L_c^{(i)}\left(\boldsymbol{\theta}^{(k-1)} \middle| \left(\boldsymbol{a}_i^{(k-1)}, \boldsymbol{b}_i^{(k-1)}\right)\right)}$;

Step 3: if $\eta \leq \rho$, we update $\boldsymbol{a}_i^{(k)}$ by $\boldsymbol{a}_i^*$; otherwise, $\boldsymbol{a}_i^{(k)} = \boldsymbol{a}_i^{(k-1)}$.

Note that part of the samples from the k^{th} iteration are re-used, avoiding the within-iteration simulation as in a multivariate rejection sampling. On the other hand, such property of the sampling requires close monitoring on the acceptance rate during Monte Carlo runs.

The M-step maximizes $\prod_{i=1}^n L_c^{(i)}(\boldsymbol{\theta}|(\boldsymbol{a}_i, \boldsymbol{b}_i))$ for parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \mathbf{A}, \mathbf{B}\}$. As $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ are considered data, the complete log-likelihood no longer involves integrals, substantially simplifying the maximization. Solving the score function for the variance components yields the following equations:

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\{x_{ij} - x_{ij}^*\}^2, \\ \mathbf{A} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_i - \boldsymbol{\alpha})^T(\boldsymbol{a}_i - \boldsymbol{\alpha}), \\ \mathbf{B} &= \frac{1}{n}(\boldsymbol{b}_i - \boldsymbol{\beta})^T(\boldsymbol{b}_i - \boldsymbol{\beta}). \end{aligned} \tag{5}$$

A Nelder–Mead method can be used to search for the solution numerically to obtain the MLEs for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ that appear in $\prod_{i=1}^n L_c^{(i)}(\boldsymbol{\theta}|(\boldsymbol{a}_i, \boldsymbol{b}_i))$ (for example, R function "optim"). Estimators for (σ^2, A, B) and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be solved iteratively between (5) and the optimization. In general, increasing the number of random effects does not substantially increase the complexity of the maximization step. However, the imputation step will be more complicated as this increases the dimension of missing data that need to be imputed.

Determining convergence appears to be an open question for StEM. In the literature, the commonly used approach for convergence diagnostics is through visual examination of the trace plots (see, e.g., [16–18]). Recently, Ref. [19] proposed a Geweke-statistics-based method. We adopt this approach in our implementation and claim that convergence is achieved when the Geweke statistics are smaller than a designated threshold.

Specifically, once starting the Markov chain, as described in the section above with the initial values, the stationarity is determined by using a batch procedure based on the Geweke statistics [20]. Let batch size be M, which can be tuned for specific data. We use a moving window for the Markov chain and compute the Geweke statistics at each increment of w iterations, another tuning parameter. More precisely, the batch procedure goes in the following steps:

1. **Initialization**. Set $\xi = 0$. Run the StEM algorithm to obtain the initial series of the estimates $\{\boldsymbol{\theta}^{(w*\xi+1)}, \ldots, \boldsymbol{\theta}^{(w*\xi+M)}\}$.

2. **Check stationarity**. For each entry p in $\boldsymbol{\theta}$, we compute the Geweke statistic z_p from the Markov chain $\{\theta_p^{(w*\xi+1)}, \ldots, \theta_p^{(w*\xi+M)}\}$ based on the standardized mean difference between the first φ portion and last ψ portion of the chain. We regard stationarity as being reached when all $|z_p|s$ are sufficiently small, i.e.,

$$\sum_{p=1}^{P} z_p^2 < \delta P,$$

where P is the total number of parameters and δ is a tuning parameter.

3. **Update.** If stationarity is not reached, execute w additional runs of the chain, increase the number ξ by 1, and then repeat Step 2.

The Fisher information matrix of our joint model has no closed-form solution due to the nonlinearity. Approximations to the Fisher information matrix by linearization have been proposed in the optimal design context by Mentre and others [21,22] for the nonlinear mixed effects model. In this paper, we compute the Fisher information matrix by linearizing the function $h(\cdot)$ and $g(\cdot)$ around the fixed effects and the conditional expectation of the individual random effects $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{a}_i, \boldsymbol{b}_i\}$. The models are approximated using first order Taylor expansions to arrive at a bivariate linear mixed effects model. The Fisher information matrix is proven to be asymptotically consistent [23]. Alternatively, the Louis principle could be used to compute the Fisher information matrix based on the complete data likelihood where the second derivatives of the likelihood are involved [24].

3. A Simulation Study

We conducted a simulation study to evaluate the StEM algorithm for a finite sample. We chose the same covariate model as in the data analysis section,

$$\begin{aligned} x_{ij} &= \log_{10}\{e^{a_{1i}} + e^{a_{2i} - a_{3i} t_{ij}}\} + e_{ij} \equiv x_{ij}^* + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \\ \boldsymbol{a}_i &= (a_{1i}, a_{2i}, a_{3i})^T \sim N((\alpha_1, \alpha_2, \alpha_3)^T, \mathbf{A}). \end{aligned}$$

The response model was also chosen to be the same as in the real-data analysis,

$$\begin{aligned} \text{logit}(E(y_{ij})) &= b_{1i} + b_{2i} * x_{ij}^* + b_{3i} * t_{ij}, \\ \boldsymbol{b}_i &= (b_{1i}, b_{2i}, b_{3i})^T \sim N((\beta_1, \beta_2, \beta_3)^T, \mathbf{B}). \end{aligned}$$

For comparison, we fit the response process data with the observed covariate, assuming no measurement error, also known as the naive approach. We also applied the commonly used two-step method that fits the covariate model first and then plugs the predicted covariate into a standard random effects logistic regression. The advantage of the two-step method is that standard software can be used, but the drawback is that the estimates can be biased [25]. We used R (4.1.0) for the entire implementation where the saemix (3.1) package and the lme4 (1.1-15) packages are used to fit the covariate model and response model, respectively.

The convergence of the StEM algorithm was ensured under the ergodic assumption of the Markov chain, which requires careful choices of the instrumental distributions to approximate the full conditionals in the M–H sampling and multi-stage sampling. Under the context of the nonlinear mixed effects model, Ref. [26] discussed a three-stage sampling strategy which can be extended to our joint model. Specifically, besides the Stage 1 sampling described in Section 2, the following two additional stages are also implemented to obtain $\boldsymbol{a}_i^{(k)}$.

Stage 2: simulate from $N(\boldsymbol{a}_i^{(k-1)}, \gamma_1 A^{(k-1)})$, i.e., a multivariate random walk as the candidate proposal for the M–H sampling.

Stage 3: simulate a succession of $\dim(\boldsymbol{a}_i)$ uni-dimensional Gaussian random walk $N(a_p^{(k-1)}, \gamma_2 \text{diag}(A_p^{(k-1)}))$ for each component of $\boldsymbol{a}_i^{(k-1)}$.

In all three stages of the M–H sampling, the target distribution is the joint distribution of $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ evaluated at the appropriate random effects and parameters. γ_1 and γ_2 are scaling parameters chosen to ensure a sufficient acceptance rate. In implementing the StEM algorithm, we tuned γ_1 and γ_2 as 0.4, leading to an acceptance rate between 20% and 40%. Following [19], we tuned the Geweke-related parameters with M = 300, w = 10, $\varphi = 0.1$, $\psi = 0.5$, and $\delta = 1.5$. Such settings led to iterations before 2000 convergence on average.

We set the true parameter values from our data analysis of the motivating dataset. Besides setting n = 388, the same as the real-data, we also simulated the scenario when n = 100 and n = 200. For simplicity, we set n_i to be 8, simulating a scenario representing an HIV trial in which the repeated measures are scheduled at weeks 0, 4, 8, 16, and every eight weeks until week 72. For each sample size, we assessed the performance when the measurement errors were at different levels $\sigma^2 = 0.20$ and $\sigma^2 = 0.40$ where the former is the estimated magnitude from the data.

We conducted 100 simulation runs for each scenario. Tables 1–3 show the simulation results for the fixed effects where we compared the estimate (Est) (defined as $\bar{\hat{\theta}} = \sum \hat{\theta}_s / S$, where $\hat{\theta}_s$ is the estimate in the s^{th} replicate, S = 100), standard deviation (SD), standard error (SE), and the coverage rate (CR). Across all three scenarios for n = 388, 100, and 200, we see that for the naive (NAI) and two-step (TS) methods, the SEs are generally smaller than the SDs, leading to poor coverage rates, especially for the parameters in the response model. The CR can be as low as 72% (n = 100, $\sigma^2 = 0.20$) for TS and 12% for NAI (n = 388, $\sigma^2 = 0.40$). On the other hand, the SEs and SDs are more agreeable under the StEM method. As a result, the coverages on the true parameters are all around the nominal level.

Table 1. Simulation results on the performance of the naive (NAI), two-step (TS), and StEM methods under different magnitudes of measurement error (n = 388).

	$\boldsymbol{\theta}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
Method	True	3.32	12.04	3.91	−0.48	−0.62	6.73
				$\sigma^2 = 0.20$			
NAI	Est				−0.55	−0.60	6.12
	SD				0.62	0.26	0.82
	SE				0.51	0.09	0.76
	CR				0.79	0.47	0.94
TS	Est	3.32	12.04	3.91	−0.54	−0.61	6.11
	SD	0.04	0.07	0.03	0.62	0.26	0.81
	SE	0.04	0.07	0.03	0.54	0.13	0.75
	CR	0.93	0.92	0.97	0.90	0.85	0.82
StEM	Est	3.32	12.04	3.91	−0.45	−0.63	6.80
	SD	0.04	0.08	0.04	0.57	0.14	0.94
	SE	0.04	0.08	0.04	0.70	0.14	0.95
	CR	0.99	0.94	0.96	1.00	0.98	0.96
				$\sigma^2 = 0.40$			
NAI	Est				−0.58	−0.59	6.14
	SD				0.62	0.25	0.81
	SE				0.49	0.08	0.76
	CR				0.60	0.12	0.95

Table 1. *Cont.*

	θ	α_1	α_2	α_3	β_1	β_2	β_3
TS	Est	3.32	12.04	3.91	−0.58	−0.59	6.14
	SD	0.05	0.09	0.03	0.62	0.25	0.81
	SE	0.04	0.08	0.04	0.54	0.13	0.75
	CR	0.90	0.91	0.97	0.90	0.81	0.80
StEM	Est	3.32	12.03	3.91	−0.48	−0.62	6.71
	SD	0.05	0.09	0.03	0.63	0.12	1.00
	SE	0.05	0.10	0.05	0.70	0.14	0.94
	CR	0.96	0.96	0.99	0.99	0.96	0.95

Table 2. Simulation results on the performance of the naive (NAI), two-step (TS), and StEM methods under different magnitudes of measurement error (n = 100).

	θ	α_1	α_2	α_3	β_1	β_2	β_3
Method	True	3.32	12.04	3.91	−0.48	−0.62	6.73
				$\sigma^2 = 0.20$			
NAI	Est				−0.17	−0.67	6.07
	SD				0.82	0.34	1.34
	SE				0.85	0.17	1.20
	CR				0.96	0.81	0.87
TS	Est	3.33	12.02	3.90	−0.17	−0.67	6.07
	SD	0.07	0.11	0.07	0.82	0.34	1.34
	SE	0.07	0.13	0.07	0.80	0.33	1.12
	CR	0.97	0.93	0.97	0.90	0.97	0.72
StEM	Est	3.35	12.02	3.87	−0.36	−0.70	6.63
	SD	0.07	0.10	0.08	0.94	0.45	1.40
	SE	0.08	0.16	0.08	0.88	0.47	1.12
	CR	0.99	0.83	0.99	1.00	0.83	0.99
				$\sigma^2 = 0.40$			
NAI	Est				−0.16	−0.72	6.12
	SD				0.87	0.36	1.23
	SE				0.91	0.16	1.32
	CR				0.97	0.78	0.94
TS	Est	3.33	12.01	3.91	−0.15	−0.72	6.12
	SD	0.11	0.17	0.07	0.87	0.36	1.23
	SE	0.08	0.14	0.07	0.80	0.36	1.14
	CR	0.81	0.86	0.99	0.94	0.94	0.75
StEM	Est	3.35	12.06	3.91	−0.24	−0.69	6.69
	SD	0.10	0.13	0.07	0.95	0.37	1.38
	SE	0.09	0.17	0.09	0.92	0.54	1.20
	CR	0.99	0.98	0.99	0.99	0.99	0.80

Table 3. Simulation results on the performance of the naive (NAI), two-step (TS)s and StEM methods under different magnitudes of measurement error (n = 200).

	θ	α_1	α_2	α_3	β_1	β_2	β_3
Method	True	3.32	12.04	3.91	−0.48	−0.62	6.73
				$\sigma^2 = 0.20$			
NAI	Est				−0.39	−0.59	5.98
	SD				0.73	0.21	0.99
	SE				0.68	0.13	0.92
	CR				0.92	0.77	0.92
TS	Est	3.32	12.04	3.91	−0.30	−0.59	5.98
	SD	0.06	0.10	0.05	0.72	0.21	1.00
	SE	0.05	0.09	0.05	0.70	0.21	0.90
	CR	0.90	0.93	0.93	0.94	0.89	0.79
StEM	Est	3.33	12.04	3.91	−0.48	−0.64	6.74
	SD	0.05	0.09	0.06	0.81	0.18	1.38
	SE	0.06	0.11	0.06	0.85	0.21	1.02
	CR	0.97	0.98	0.97	0.98	0.98	0.86
				$\sigma^2 = 0.40$			
NAI	Est				−0.37	−0.59	6.03
	SD				0.74	0.21	0.98
	SE				0.66	0.12	0.93
	CR				0.89	0.51	0.93
TS	Est	3.33	12.03	3.91	−0.37	−0.59	6.03
	SD	0.08	0.13	0.06	0.74	0.21	0.98
	SE	0.06	0.11	0.05	0.71	0.17	0.91
	CR	0.83	0.87	0.91	0.96	0.89	0.81
StEM	Est	3.33	12.04	3.91	−0.37	−0.63	6.39
	SD	0.07	0.11	0.06	0.73	0.20	1.19
	SE	0.07	0.13	0.06	0.85	0.21	1.02
	CR	0.92	0.96	0.95	0.99	0.94	0.87

To assess the accuracy of the methods, following [27], we also calculated the standardized bias and root mean square error (RMSE) over the 100 replications for both the fixed effects parameters and the dispersion parameters in the response model.

$$\text{Standardized bias} = \left(\bar{\hat{\theta}} - \theta\right)\left\{\frac{\sum\left(\hat{\theta}_s - \bar{\hat{\theta}}\right)^2}{S-1}\right\}^{-1/2}$$

$$\text{RMSE} = \left\{\left(\bar{\hat{\theta}} - \theta\right)^2 + \frac{\sum\left(\hat{\theta}_s - \bar{\hat{\theta}}\right)^2}{S-1}\right\}^{1/2}$$

where θ is the true value for the parameter of interest. The standardized bias ranges from -1 to 1. Figures 1–3 show the results. There, we see that there are no distinguishable differences among the three methods in terms of the standardized bias and RMSE for β_1 and β_2 across the three settings n = 388, 100, and 200. Significant biases and RMSEs can be seen for β_3 under the NAI and TS methods. For those two methods, estimates become worse for the dispersion parameters. The severest case is at $\sigma^2_{B_{11}}$ for all three scenarios. Furthermore, there are noticeable deviations from zero for $\sigma^2_{B_{22}}$ and $\sigma^2_{B_{33}}$ where the StEM outperforms the classical methods, especially for n = 388 and n = 200 cases.

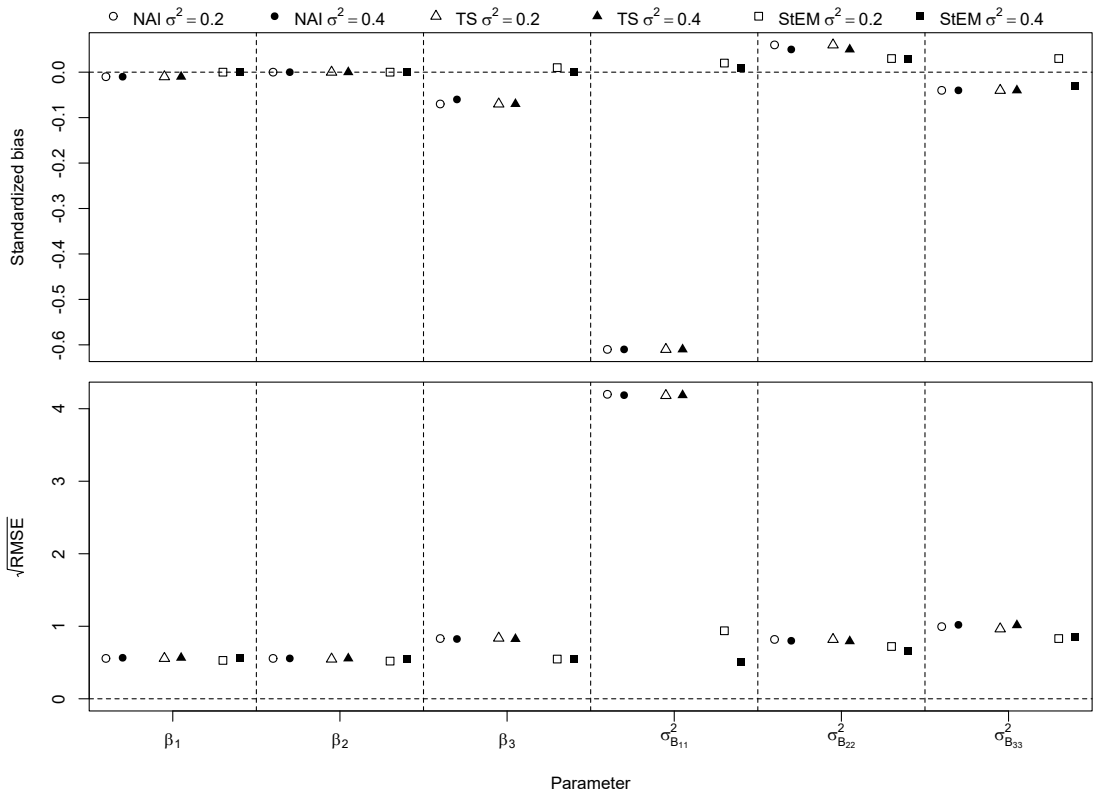

Figure 1. Simulation results, standard biases and RMSEs, for the parameters in the response model by the three methods: naive (NAI), two-step (TS) and StEM for n = 388.

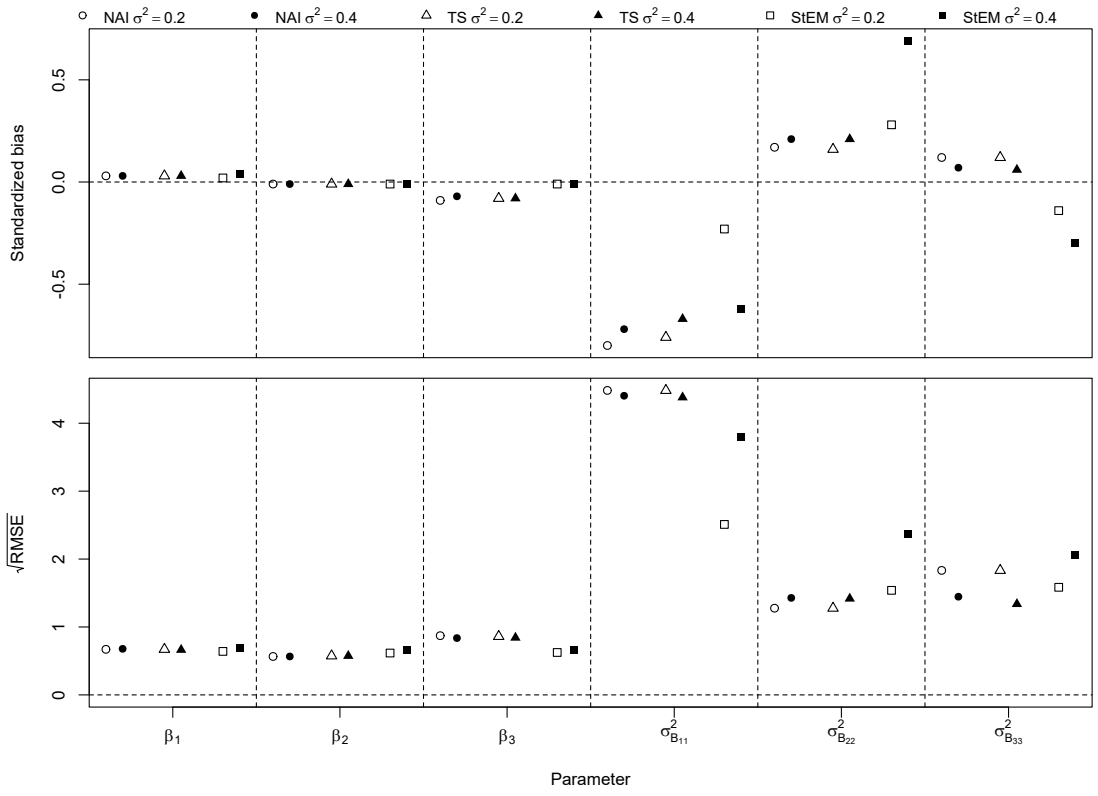

Figure 2. Simulation results of the standard biases and RMSEs for the parameters in the response model of the three methods: naive (NAI), two-step (TS), and StEM for n = 100.

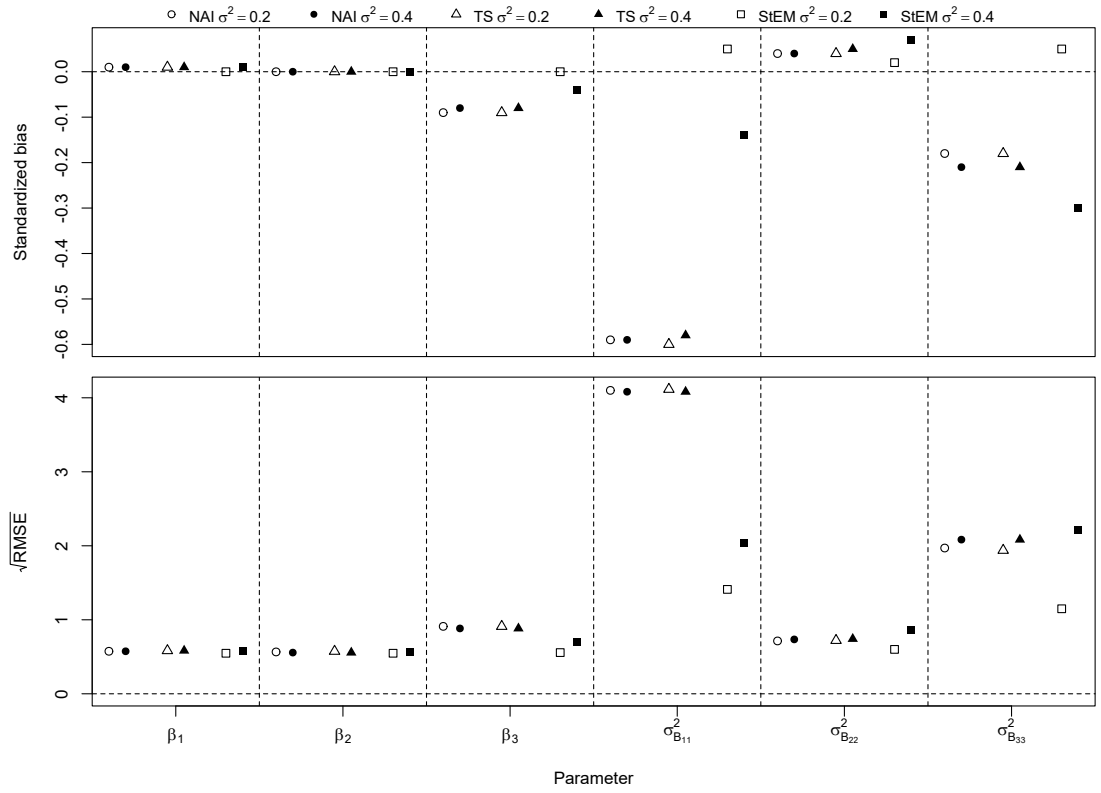

Figure 3. Simulation results of the standard biases and RMSEs for the parameters in the response model of the three methods: naive (NAI), two-step (TS), and StEM for n = 200.

4. Application to ACTG Trial

This section analyzes the AIDS clinical trials group (ACTG) 388 dataset [28]. ACTG 388 was a randomized, open-label study comparing two different four-drug regimens with a standard three-drug regimen for 512 subjects with no or limited previous experience with antiretroviral therapy. The plasma HIV-1 RNA (viral load) was repeatedly quantified at weeks 0, 4, 8, 16, and every eight weeks until week 72. The CD4 cell counts were also measured throughout this study on a similar schedule. We removed subjects with fewer than three viral load measurements and those who encountered viral rebound, leading to analytic data with 388 individuals. There are 3019 total observations with a median of 7.0 visits, ranging from 3 to 13. Figure 4 shows the trajectories of viral load and CD4 count in the analytic data.

We used regression models to investigate the association between viral load and CD4 restoration during the HIV treatment, accounting for the measurement errors in the time-dependent viral load. Due to the lack of mechanistic models for CD4 data, we considered empirical models for CD4. Since CD4 data are counts, it might be interesting to consider Poisson models since they are typical choices for count data. In AIDS studies, it is also interesting to see if CD4 values are above or below 200 cells/mm^3 and how the viral load impacts such transition; thus, it is useful to model CD4 as binary data. Specifically, we modeled CD4 with

$$E[y_{ij}|x^*_{ij}] = g(b_{1i} + b_{2i}x^*_{ij} + b_{3i}t_{ij})$$

where $g(\cdot)$ is the inverse of the logit link function, y_{ij} is the binary CD4 value of individual i at time t_{ij}, $\boldsymbol{b}_i = (b_{1i}, b_{2i}, b_{3i})^T$ is a vector of individual random effects, and x^*_{ij} is the unobserved true value of the observed viral load x_{ij} which is measured with errors. We assumed $\boldsymbol{b}_i \sim N((\beta_1, \beta_2, \beta_3)^T, \mathbf{B})$. We specified random intercept, random slope for viral load, random slope for time, and defined a diagonal covariance matrix $\mathbf{B}$ as such a model produced the smallest AIC in our preliminary data analysis.

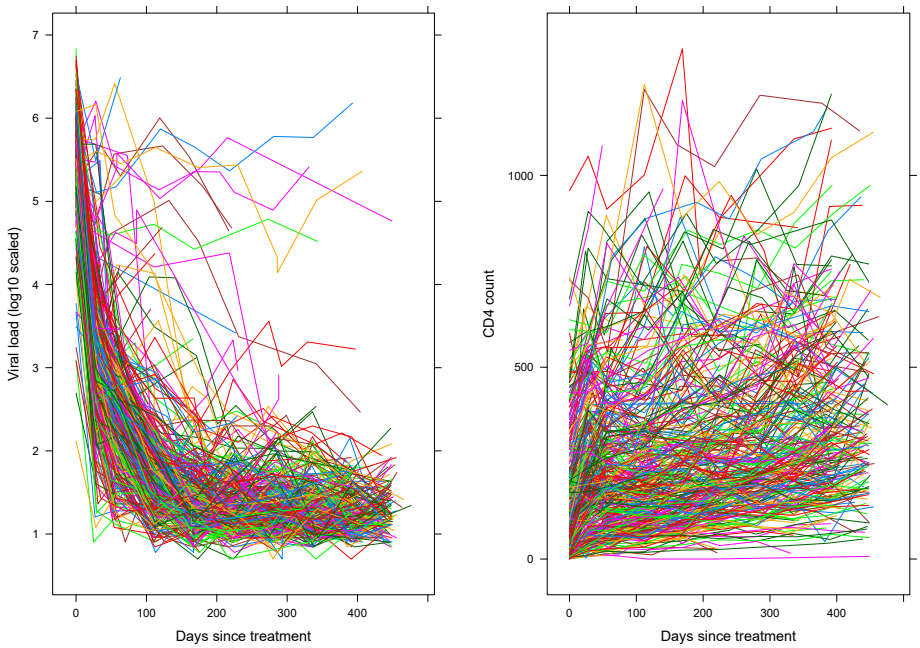

Figure 4. Trajectories of viral load and CD4 count after the initiation of HIV treatment in ACTG 388 data.

Since mechanistic nonlinear models are available and biologically justified for viral load data, we chose a one-compartment mechanistic nonlinear model proposed by [29] based on our preliminary analysis. Specifically, for x_{ij}, the observed viral load ($\log_{10}$ transformed) for individual i at time j, we have

$$x_{ij} = \log_{10}\{P_{1i} + P_{2i}e^{-\lambda_i t_{ij}}\} + e_{ij} \equiv x^*_{ij} + e_{ij},$$

where $\log(P_{1i}) = \alpha_1 + a_{1i}$, $\log(P_{2i}) = \alpha_2 + a_{2i}$, $\lambda_i = \alpha_3 + a_{3i}$. The individual-specific random effects characterize viral trajectory where P_{1i} is the baseline value, P_{2i} is the decay slope, and λ_{2i} is the decay rate. As usual, e_{ij} is the measurement error with $e_{ij} \sim N(0, \sigma^2)$. We assumed $\boldsymbol{a}_i = (a_{1i}, a_{2i}, a_{3i})^T \sim N(0, \mathbf{A})$ where $\mathbf{A}$ is unstructured to allow all possible correlations among the random effects. To avoid minimal (large) estimates, which may be unstable, we re-scaled the original time (range 0 to 504 days) to a range between 0 and 1.

Figure 5 displays the trace plot for all the parameters in the joint model when fitting the StEM algorithm to the ACTG 388 data. Table 4 shows the estimates and standard errors from the three approaches. Note that a value of -0.48, -0.51, and -0.60 is estimated for β_2 from the naive and two-step methods and the StEM algorithm, respectively, reflecting the negative association between viral load and CD4. In addition, the standard errors are estimated larger in StEM method compared to the other methods. Furthermore, higher variability is seen for the variance–covariance matrix of the response process by the StEM algorithm, likely attributed to the fact that the joint model incorporates more uncertainties due to the mismeasured covariate. We also calculated the deviance of the response model using the conditional mode of the random effects upon convergence. There is no improvement over the classical models, but this is understandable as a true version of the covariate might not explain more variation of the response. Albeit, based on the findings from simulation, we believe the results from the StEM model are more accurate regarding the magnitude of the relationship between viral load and CD4 as well as the associated uncertainty.

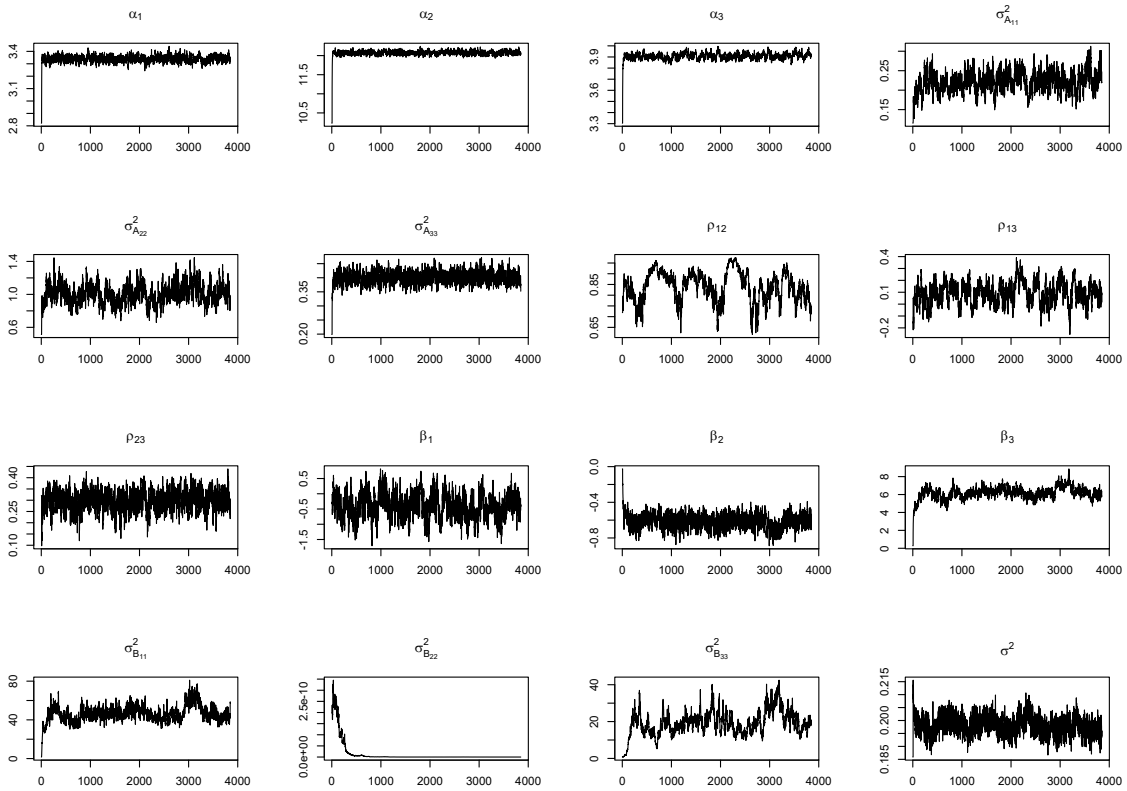

Figure 5. Trace plots of all parameters in the joint model of random effects logistic regression and mechanistic nonlinear random effects model for ACTG 388 data using StEM algorithm which stops at iteration 3790.

Table 4. Estimates in fitting the naive (NAI), two-step (TS), and the StEM methods to the ACTG 388 data.

Method	Parameter	α_1	α_2	α_3	β_1	β_2	β_3
NAI	est				−0.39	−0.48	5.50
	se				0.49	0.09	0.59
TS	est	3.32	12.04	3.89	−0.54	−0.51	5.66
	se	0.03	0.07	0.03	0.49	0.10	0.59
StEM	est	3.34	12.07	3.91	−0.48	−0.62	6.73
	se	0.04	0.08	0.04	0.56	0.11	0.61

Estimates of the variance components: NAI: $\mathbf{B}: \begin{pmatrix} 32.31 & & \\ & 0.01 & \\ & & 16.24 \end{pmatrix}$; TS: $\sigma^2 = 0.20$, $\mathbf{A}: \begin{pmatrix} 0.23 & 0.35 & 0.01 \\ & 1.02 & 0.20 \\ & & 0.40 \end{pmatrix}$ $\mathbf{B}: \begin{pmatrix} 31.97 & & \\ & 0.01 & \\ & & 16.39 \end{pmatrix}$; StEM: $\sigma^2 = 0.20$, $\mathbf{A}: \begin{pmatrix} 0.23 & 0.37 & 0.02 \\ & 1.07 & 0.10 \\ & & 0.40 \end{pmatrix}$ $\mathbf{B}: \begin{pmatrix} 49.38 & & \\ & 0.01 & \\ & & 26.33 \end{pmatrix}$.

5. Discussion

To analyze longitudinal binary data with mismeasured time-varying covariate, we proposed an ML estimation method that may be preferred over methods classically used. We extended the StEM algorithm proposed by [8] by including the Gibbs sampling step of the StEM algorithm coupled with a three-stage M–H sampling strategy [26]. We adopted the Geweke-statistics-based stopping criterion [19] and applied the linearization approach [23] at the convergence to obtain the Fisher information matrix. We applied the StEM algorithm to model the HIV viral load and CD4 count processes to assess their relationship. The simulation study illustrates the precision and accuracy of our approach. We showed that the coverage

rate, bias, and RMSE obtained by the StEM algorithm are highly satisfactory for all parameters, outperforming the other two classical approaches in handling measurement error.

The one-compartment model we used is deduced from a differential equation model proposed by [30] describing the HIV dynamics with both the viral load decrease and the CD4+ increase under treatment. Ref. [29] showed that this differential system has an analytic solution under the assumption that the non-infected CD4+ cell concentration is constant. Because this assumption is valid before any viral rebound due to multiple virus mutations, we only considered the subset of the ACTG 388 data without viral rebound. The StEM algorithm could also be extended to cases with viral rebound (see, e.g., [17]), but that is out of the scope of this paper.

To take into account the measurement error problem with longitudinal binary data, Ref. [6] proposed an exact MCEM algorithm and emphasized computational problems, such as slow or even no convergence, especially when the dimension of the random effects is not small. The main problem of the MCEM is the simulation of large independent samples of the random effects at each iteration. Our implementation of the StEM algorithm is a more adapted tool to estimate joint models as it avoids the computational problem in the E-step. Indeed, only one realization of the random effects simulated is needed at each iteration with an incremental updating of the samples rather than generating an entire new sample through rejection sampling, thereby side-stepping the computational problem of the E-step of the MCEM. As an example, it takes 48 h for the exact MCRM algorithm to converge when using the same viral load dynamic model and response model (i.e., a joint model with a random effect vector of size $p + q = 3 + 3$), whereas the StEM algorithm takes about 30 min to converge. The StEM, which is a true ML estimation method, is about 96 times faster than the MCEM algorithm.

We only focused on the measurement error problem in the context of viral load observations as the single covariate for the dichotomous CD4 count which is concurrently measured, i.e., matched, with viral load. Still, StEM can be extended to multiple time-varying covariates, other distributions of the response, e.g., Poisson, and other data complexes such as mismatched data or left censoring which [6] included in the MCEM. For example, when there is more than one error-prone covariate, the methods are similar: we can assume a multivariate LME model [31] for the covariate processes and then proceed in a similar way. When a covariate is missing at a certain time point, we could replace it with a predicted value during implementation. A missing covariate with a response variable can be more involved where we could expand the joint model by adding a missing data model; a model, when missing, is non-ignorable [32]. Note that when the covariates are time-independent and are measured with errors, replications or validation data are needed to address measurement errors [4]. When the error-prone covariates are time-dependent, however, statistical analysis can proceed without replications or validation data if we assume that the true covariate value changes smoothly over time and model the covariate process empirically as shown above since, in this case, we can view the repeated measurements as replicates.

One additional note is on the choice of tuning parameters in our implementation of the StEM algorithm. Following the guideline from [19], those values lead to satisfying results although empirical in nature. Further research in this area is warranted.

In conclusion, our StEM algorithm combines the statistical properties of several methods together with computational efficiency, estimation accuracy, and precision. We thus recommend using this method to model longitudinal binary data with mismeasured time-varying covariates.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bolder, B.M.; Brooks, M.; Clark, C.; Geange, S.; Pouls, J.; Stevens, M.; White, J. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.* **2009**, *24*, 127–141.
2. McCulloch, C.; Searle, S.; Neuhaus, J. *Generalized, Linear, and Mixed Models*; John Wiley and Sons: New York, NY, USA, 2008.
3. Olver, F.; Lozier, D.; Boisver, R.; Clark, C. *Quadrature: Gauss-Hermit Formula: NIST Handbook of Mathematical Functions*; Cambridge University Press: London, UK, 2010.
4. Carroll, R.; Ruppert, D.; Stefanski, L. *Measurement Error in Nonlinear Models: A Modern Perspective*; Chapman and Hall: Boca Raton, FL, USA, 2006.
5. Fuller, W. *Measurement Error Models*; John Wiley and Sons: New York, NY, USA, 1987.
6. Zhang, H.; Wong, H.; Wu, L. A mechanistic nonlinear model for censored and mis-measured covariates in longitudinal models, with application in AIDS studies. *Stat. Med.* **2018**, *37*, 167–178. [CrossRef] [PubMed]
7. Wei, G.; Tanner, M. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Am. Stat. Assoc.* **1990**, *85*, 699–704. [CrossRef]
8. Diebolt, J.; Celeus, G. Asymptotic properties of a stochastic EM algorithm for estimating mixture proportions. *Stoch. Model.* **1996**, *9*, 599–613.
9. Mbognig, C.; Bleakley, K.; Lavielle, M. Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation-maximization algorithm. *J. Stat. Comput. Simul.* **2015**, *85*, 1512–1528. [CrossRef]
10. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **1977**, *39*, 1–38.
11. Ip, E.H. A Stochastic EM Estimator in the Presence of Missing Data: Theory and Application. Ph.D. Dissertation, Standford University, Standford, CA, USA, November 1994.
12. Nielsen, S. The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **2000**, *6*, 457–489. [CrossRef]
13. Geweke, J. *Handbook of Computational Economics*; North-Holland: Amsterdam, The Netherlands, 1996.
14. Zhang, H.; Wu, L. A nonlinear model for truncated and mis-measured time-varying covariates in survival models, with applications in HIV/AIDS studies. *J. R. Stat. Soc.—Appl. Stat. Ser.* **2018**, *67*, 1437–1450. [CrossRef]
15. Baey, C.; Trevezas, S.; Cournede, P. A non linear mixed effects model of plant growth and estimation via stochastic variants of the EM algorithm. *Commun. Stat.—Theory Methods* **2016**, *45*, 1643–1669. [CrossRef]
16. Huang, R.; Xu, R.; Dulai, P.S. Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risk outcomes. *Stat. Med.* **2020**, *39*, 3397–3411. [CrossRef]
17. Wang, R.; Bing, A.; Wang, C.; Hu, Y.; Bosch, R.; DeGruttola, V. A flexible nonlinear mixed effects model for HIV viral load rebound after treatment interruption. *Stat. Med.* **2020**, *39*, 2051–2066. [CrossRef] [PubMed]
18. Yang, F. A stochastic EM algorithm for quantile and censored quantile regression models. *Compt. Econ.* **2018**, *52*, 555–582. [CrossRef]
19. Zhang, S.; Chen, Y.; Liu, Y. An improved stochastic EM algorithm for large-scale full-information item factor analysis. *Br. J. Math. Stat. Psychol.* **2020**, *73*, 44–71. [CrossRef] [PubMed]
20. Gelman, A.; Rubin, D. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457–472. [CrossRef]
21. Mentre, F.; Mallet, A.; Baccar, D. Optimal design in random-effects regression models. *Biometrika* **1997**, *84*, 429–442. [CrossRef]
22. Retout, S.; Comets, E.; Samson, A.; Mentre, F. Design in nonlinear mixed effects models: Optimization using the Fedorov-Wynn algorithtm and power of the Wald test for binary covariates. *Stat. Med.* **2007**, *26*, 5162–5179. [CrossRef]
23. Zhang, H.; Wu, L. An approximate method for generalized linear and nonlinear mixed effects models with mechanistic nonlinear covariate measurement error model. *Metrika* **2019**, *82*, 471–499. [CrossRef]
24. Louis, T. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc.* **1982**, *44*, 226–233.
25. Wu, L. *Mixed Effects Models for Complex Data*; Chapman and Hall: London, UK, 2010.
26. Samson, A.; Lavielle, M.; Mentre, F. Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Comput. Stat. Data Anal.* **2006**, *51*, 1562–1574. [CrossRef]
27. Burton, A.; Altman, D.; Royston, P.; Holder, R. The design of simulation studies in medical statistics. *Stat. Med.* **2006**, *25*, 4279–4292. [CrossRef]
28. Fischl, M.; Ribaudo, H.; Collier, A. A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. *J. Infect. Dis.* **2003**, *188*, 625–634. [CrossRef] [PubMed]
29. Wu, H.; Ding, A. Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **1999**, *55*, 410–418. [CrossRef] [PubMed]
30. Perelson, A.; Essunger, P.; Cao, Y.; Vesanen, M.; Hurley, A.; Saksela, K.; Markowitz, M.; Ho, D. Decay characteristics of HIV-1 infected compartments during combination therapy. *Nature* **1997**, *387*, 188–191. [CrossRef] [PubMed]
31. Shah, A.; Laird, N.; Schoenfeld, D. A random-effects model for multiple characteristics with possibly missing data. *Am. Stat. Assoc.* **1997**, *92*, 775–779. [CrossRef]
32. Little, R.; Rubin, D. *Statistical Analysis with Missing Data*; Wiley: New York, NY, USA, 2002.