Entropy **2015**, *17*, 5353-5381; doi:10.3390/e17085353

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Approximate Methods for Maximum Likelihood Estimation of Multivariate Nonlinear Mixed-Effects Models

Wan-Lun Wang

Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taichung 40724, Taiwan; E-Mail: wlunwang@fcu.edu.tw; Tel.: +886-4-24517250 (ext. 4407); Fax: +886-4-24517092

Academic Editors: Carlos Alberto De Bragança Pereira and Adriano Polpo

Received: 21 April 2015 / Accepted: 21 July 2015 / Published: 29 July 2015

Abstract: Multivariate nonlinear mixed-effects models (MNLMM) have received increasing use due to their flexibility for analyzing multi-outcome longitudinal data following possibly nonlinear profiles. This paper presents and compares five different iterative algorithms for maximum likelihood estimation of the MNLMM. These algorithmic schemes include the penalized nonlinear least squares coupled to the multivariate linear mixed-effects (PNLS-MLME) procedure, Laplacian approximation, the pseudo-data expectation conditional maximization (ECM) algorithm, the Monte Carlo EM algorithm and the importance sampling EM algorithm. When fitting the MNLMM, it is rather difficult to exactly evaluate the observed log-likelihood function in a closed-form expression, because it involves complicated multiple integrals. To address this issue, the corresponding approximations of the observed log-likelihood function under the five algorithms are presented. An expected information matrix of parameters is also provided to calculate the standard errors of model parameters. A comparison of computational performances is investigated through simulation and a real data example from an AIDS clinical study.

Keywords: importance sampling; Laplacian approximation; Monte Carlo EM; penalized nonlinear least squares; pseudo expectation conditional maximization

MSC Classifications: 62H12; 62J02

---

1. Introduction

Analysis of multi-outcome longitudinal data with various features has attracted considerable interest in clinical trials, biological psychology, environmental science and medical research, to name a few. The methodology of multivariate linear mixed-effects models (MLMM) [1] and multivariate nonlinear mixed-effects models (MNLMM) [2] has been developed for related work. A comprehensive study of the MLMM along with its applications can be found in [3–7], among others. Nonlinear models for repeated-measures data rest on more complicated mathematical derivations and heavier computational requirements than linear models, but they can offer flexibility in capturing a broader range of data patterns. Several approaches to carrying out maximum likelihood (ML) estimation of nonlinear mixed-effects models (NLMM) for single-outcome longitudinal data have been studied; see, for example, [8–12]. Bayesian inference in NLMM via Markov chain Monte Carlo (MCMC) procedures can be found, for instance, in [13–15]. Although the use of the NLMM, as well as its extensions in other families of distributions have been pretty well established in the literature, to the best of our knowledge, exploration of the inference on MNLMM is relatively rare so far. Analyzing each response variable of the data by fitting the NLMM separately might be inappropriate and fail to take account of the between-variable association, as well as its evolution.

For the general NLMM, the linearization method [8,16] that exploits a first-order Taylor expansion to approximate the nonlinear function in terms of a linear pseudo-data model is by far the most widely-used approach due to its numerical simplicity. Despite its popularity, [17] argued that the linearization method may produce substantial bias in parameter estimation, as the number of observations per subject is too small, and the variability of random effects tends to be large at the same time. Although computationally much simpler, the Laplace approximation method [10] can also lead to considerably-biased parameter estimates, depending on the quality of the mode. As an alternative to the pseudo-data and Laplace approximation approaches, the integral approximation methods that use Monte Carlo integration [18] or importance sampling [19] to approximate the observed likelihood may provide more accurate estimates than the linearization method. However, the numerical integration methods are generally inefficient to implement and become computationally prohibitive when the dimension of random effects increases [20]. Over the past few decades, several estimation algorithms for NLMM have been developed and implemented in different software. For example, the linearization methods using the first-order Taylor expansion [21] or the first-order conditional estimation (FOCE) [8,16] are embedded in R function `nlme`, while the Laplace approximation method is implemented in `NONMEM` [22] and the SAS macro `NLIMIX` [23]. A new SAS macro `NLMIXED` incorporating adaptive Gaussian quadrature has shown considerable improvement [24]. The other improved procedure based on the stochastic approximation expectation maximization [25] was implemented in `MONOLIX` [26], `NONMEM` [27] and R package `saemix` [28]. Multivariate nonlinear mixed-effects models can be fitted using *ad hoc* manipulation by expanding the design matrix with extra columns of dummy covariates flagging each element of the original multivariate responses.

Consider the multiple repeated measures $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i),\ i = 1, \ldots, N\}$, where $\boldsymbol{Y}_i$ is a $s_i \times r$ response matrix composed of r response vectors $\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,s_i})^{\mathrm{T}}$, $j = 1, \ldots, r$, and $\boldsymbol{X}_i$ is the covariate matrix for the i-th subject. Let $\boldsymbol{E}_i = [\boldsymbol{e}_{i1} : \boldsymbol{e}_{i2} : \cdots : \boldsymbol{e}_{ir}]$ be the $s_i \times r$ matrix of within-subject errors

associated with $\boldsymbol{Y}_i$, where $\boldsymbol{e}_{ij} = (e_{ij,1}, \ldots, e_{ij,s_i})$. Let $\boldsymbol{y}_i = \text{vec}(\boldsymbol{Y}_i)$ and $\boldsymbol{e}_i = \text{vec}(\boldsymbol{E}_i)$ denote the stacked $s_i r \times 1$ vectors of all responses and within-subject errors, respectively.

In general, the MNLMM takes the form of:

$$\boldsymbol{y}_i = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{X}_i) + \boldsymbol{e}_i, \quad i = 1, \ldots, N \tag{1}$$

where $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{X}_i)$ is a nonlinearly-differentiable function of a subject-specific parameter $\boldsymbol{\eta}_i$ governing the within-profile behaviors and $\boldsymbol{e}_i$ is a vector containing normally-distributed error components. Moreover, the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}_i$ can be incorporated into the model by letting:

$$\boldsymbol{\eta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \tag{2}$$

where $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are design matrices of size $s \times p$ and $s \times q$, respectively. We assume that $\boldsymbol{b}_i$ follows a multivariate normal distribution with mean vector $\boldsymbol{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{D}$, denoted by $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$, and independent of $\boldsymbol{e}_i \sim \mathcal{N}_{s_i r}(\boldsymbol{0}, \boldsymbol{R}_i)$. The joint distributions of $(\boldsymbol{b}_i^{\mathrm{T}}, \boldsymbol{e}_i^{\mathrm{T}})^{\mathrm{T}}$ for distinct subjects are independent. To reduce the number of parameters in $\boldsymbol{R}_i$, we assume that the k-th row of $\boldsymbol{E}_i$, say $\boldsymbol{e}_{i\cdot k}$, follows $\mathcal{N}_r(\boldsymbol{0}, \boldsymbol{\Sigma})$, and the j-th column of $\boldsymbol{E}_i$, say $\boldsymbol{e}_{ij\cdot}$, follows $\mathcal{N}_{s_i}(\boldsymbol{0}, \boldsymbol{C}_i)$, such that $\boldsymbol{R}_i = \boldsymbol{\Sigma} \otimes \boldsymbol{C}_i$. This specification implies that within-subject errors for all responses measured at the same occasion have variance-covariance $\boldsymbol{\Sigma}$. To capture the extra autocorrelation of a given response among irregularly-observed occasions, some parsimonious dependence structures can be made on $\boldsymbol{C}_i$, such as the compound symmetry, the p-order autoregressive model [29,30] and the damped exponential correlation [31]. For simplicity, we write $\boldsymbol{C}_i = \boldsymbol{C}_i(\boldsymbol{\phi})$, which depends on subject i according to its dimension s_i with each entry being a function of a small set of parameters $\boldsymbol{\phi}$ describing within-subject autocorrelation.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$ be the entire model parameters. According to Model Equation (1) with Assumption Equation (2), the marginal density of $\boldsymbol{y}_i$ is:

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}) = \int \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i, \boldsymbol{R}_i\big)\phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D})d\boldsymbol{b}_i, \tag{3}$$

where $\phi_d(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega})$ denotes the probability density function (pdf) of a d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Omega}$. Typically, this integral cannot yield a closed-form expression when the vector-valued function $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{X}_i)$ is nonlinear in random effects $\boldsymbol{b}_i$. Thus, the log-likelihood function of $\boldsymbol{\theta}$ for $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ is given by:

$$\begin{aligned}\ell(\boldsymbol{\theta}|\boldsymbol{y}) \;=\; & \sum_{i=1}^{N} \log\Big\{ \int (2\pi)^{-(s_i r+q)/2}|\boldsymbol{\Sigma}|^{-s_i/2}|\boldsymbol{C}_i|^{-r/2}|\boldsymbol{D}|^{-1/2} \\ & \times \exp\Big\{ -\frac{1}{2}\big[\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i\big)^{\mathrm{T}}\boldsymbol{R}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i\big) + \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i\big]\Big\}d\boldsymbol{b}_i\Big\}. \end{aligned} \tag{4}$$

The purpose of this article is to consider five different methods for carrying out ML estimation of the MNLMM described in Equation (1) along with Equation (2) and for approximating the observed log-likelihood Function Equation (4). The methods include the penalized nonlinear least squares coupled to multivariate linear mixed effects (PNLS-MLME) approximation [8], Laplacian approximation [32], a pseudo-data version of the expectation conditional maximization (ECM) algorithm [33], the Monte

Carlo EM (MCEM) algorithm [34] and the importance sampling EM (ISEM) algorithm [35]. The approximation to the observed log-likelihood is based on the standard Taylor expansion and is easy to calculate within the algorithms. A simple way of computing standard errors of parameters via the information-based method is provided.

The article is organized as follows. In Section 2, we describe the five computational procedures for ML estimation of the MNLMM together with the calculation of standard errors of parameters. In Section 3, the proposed methodology is illustrated with the analysis of HIV-AIDS data. Section 4 presents a comparison of the five approximation methods through simulation studies. We summarize and discuss implications in Section 5. The technical derivations are collected in the Appendix.

2. Five Approximate ML Procedures

From Model Equation (1), the j-th column (outcome) of $\boldsymbol{Y}_i$, say $\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,s_i})^{\mathrm{T}}$, can be formulated as:

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu}_{ij}(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}) + \boldsymbol{e}_{ij},$$

where $\boldsymbol{\mu}_{ij}(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}) = (\mu_j(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij,1}), \ldots, \mu_j(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij,s_i}))^{\mathrm{T}}$ and $\boldsymbol{e}_{ij} = (e_{ij,1}, \ldots, e_{ij,s_i})^{\mathrm{T}}$. Analogously, the model for the k-th row (occasion) can be expressed as:

$$\boldsymbol{y}_{i,k} = \boldsymbol{\mu}_i^k(\boldsymbol{\eta}_i, \boldsymbol{x}_{ik}) + \boldsymbol{e}_{i,k},$$

where $\boldsymbol{y}_{i,k} = (y_{i1,k}, \ldots, y_{ir,k})^{\mathrm{T}}$, $\boldsymbol{\mu}_i^k(\boldsymbol{\eta}_i, \boldsymbol{x}_{ik}) = (\mu_1(\boldsymbol{\eta}_i, \boldsymbol{x}_{i1,k}), \ldots, \mu_r(\boldsymbol{\eta}_i, \boldsymbol{x}_{ir,k}))^{\mathrm{T}}$ and $\boldsymbol{e}_{i,k} = (e_{i1,k}, \ldots, e_{ir,k})^{\mathrm{T}}$. We present five algorithms for employing ML estimation of Model Equation (1). The approximation to the observed log-likelihood Function Equation (4) and the calculation of standard errors of parameters are discussed, as well.

2.1. PNLS-MLME Procedure

Following the linear mixed-effects (LME) approximation method suggested by [8], the first procedure consists of two steps: a penalized nonlinear least squares (PNLS) step and a multivariate LME (MLME) step. The basic idea behind this procedure is that we estimate the unobservable random effects $\boldsymbol{b}_i$ via the PNLS step and then update the ML estimates of parameters $\boldsymbol{\theta}$ based on the formulation of MLMM for the pseudo-data. Specifically, the proposed PNLS-MLME procedure is sketched below.

In the PNLS step, first define:

$$g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}) = \big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\big)^{\mathrm{T}}(\boldsymbol{\Sigma} \otimes \boldsymbol{C}_i)^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\big) + \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i, \tag{5}$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{X}_i)$, for $i = 1, 2, \cdots, N$, is a function of fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{b}_i$. Fixing the current estimates of parameters $\hat{\boldsymbol{\theta}}^{(h)} = (\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{D}}^{(h)}, \hat{\boldsymbol{\Sigma}}^{(h)}, \hat{\boldsymbol{\phi}}^{(h)})$, the conditional modes of random effects $\boldsymbol{b}_i$ are obtained through minimizing a penalized nonlinear least-squares objective function:

$$\{\hat{\boldsymbol{b}}_i^{(h)}\}_{i=1}^N = \arg\min \sum_{i=1}^N g(\boldsymbol{y}_i, \boldsymbol{b}_i, \hat{\boldsymbol{\theta}}^{(h)}). \tag{6}$$

The joint distributions $(\boldsymbol{b}_i^{\mathrm{T}}, \boldsymbol{e}_i^{\mathrm{T}})^{\mathrm{T}}$ for distinct subjects are independent, and thus, all $\boldsymbol{y}_i$ are independent of each other. In practice, solving over $\hat{\boldsymbol{b}}_i^{(h)}$ for each subject can be implemented by minimizing $g(\boldsymbol{y}_i, \boldsymbol{b}_i, \hat{\boldsymbol{\theta}}^{(h)})$ with respect to q-dimensional random effects of one subject at a time, rather than finding the solutions with respect to those of all subjects simultaneously.

In the MLME step, which allows updating the parameter estimates, we utilize the first-order Taylor expansion of Model Equation (1) around the current estimates $\hat{\boldsymbol{\eta}}_i^{(h)} = \boldsymbol{A}_i\hat{\boldsymbol{\beta}}^{(h)} + \boldsymbol{B}_i\hat{\boldsymbol{b}}_i^{(h)}$, that is,

$$y_{ij,k} - \mu_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k}) + \dot{\mu}_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k})^{\mathrm{T}}\hat{\boldsymbol{\eta}}_i^{(h)} = \dot{\mu}_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k})^{\mathrm{T}}\boldsymbol{\eta}_i + e_{ij,k},$$

where $\dot{\mu}_j$, $j = 1, \ldots, r$, are the first partial derivatives of μ_j with respect to $\boldsymbol{\eta}_i$ and $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ are replaced by $\hat{\boldsymbol{\beta}}^{(h)}$ and $\{\hat{\boldsymbol{b}}_i^{(h)}\}_{i=1}^N$, respectively. Denote the pseudo-data by:

$$\tilde{y}_{ij,k} = y_{ij,k} - \mu_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k}) + \tilde{\boldsymbol{x}}_{ijk}\hat{\boldsymbol{\beta}}^{(h)} + \tilde{\boldsymbol{z}}_{ij,k}\hat{\boldsymbol{b}}_i^{(h)}, \tag{7}$$

where $\tilde{\boldsymbol{x}}_{ijk} = \dot{\mu}_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k})^{\mathrm{T}}\boldsymbol{A}_i$ and $\tilde{\boldsymbol{z}}_{ijk} = \dot{\mu}_j(\hat{\boldsymbol{\eta}}_i^{(h)}, \boldsymbol{x}_{ij,k})^{\mathrm{T}}\boldsymbol{B}_i$. Consequently, Model Equation (1) can be rewritten as:

$$\tilde{y}_{ij,k} = \tilde{\boldsymbol{x}}_{ijk}\boldsymbol{\beta} + \tilde{\boldsymbol{z}}_{ijk}\boldsymbol{b}_i + e_{ij,k}.$$

The model for the super vector of the pseudo-data for the i-th subject is:

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{8}$$

where $\tilde{\boldsymbol{y}}_i$ is a $s_i r \times 1$ vector composed of r pseudo-response vectors $\tilde{\boldsymbol{y}}_{ij} = (\tilde{y}_{ij,1}, \cdots, \tilde{y}_{ij,s_i})^{\mathrm{T}}$, $\tilde{\boldsymbol{X}}_i$ is a $s_i r \times p$ matrix with rows made up of $p \times 1$ vector $\tilde{\boldsymbol{x}}_{ijk}$ and $\tilde{\boldsymbol{Z}}_i$ is a $s_i r \times q$ matrix with rows made up of $q \times 1$ vector $\tilde{\boldsymbol{z}}_{ijk}$. Obviously, Model Equation (8) for the pseudo-data is shown in an LME representation, so the estimation procedure becomes much simpler.

Therefore, the log-likelihood function of $\boldsymbol{\theta}$ according to Model Equation (8) can be approximated by:

$$\begin{aligned}\ell_{\mathrm{PD}}(\boldsymbol{\theta}|\boldsymbol{y}) \;\cong\; & -\frac{1}{2}\sum_{i=1}^{N}\Big\{s_i r\log(2\pi) + \log\big|\tilde{\boldsymbol{Z}}_i\boldsymbol{D}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} + \boldsymbol{\Sigma}\otimes\boldsymbol{C}_i\big| \\ & + \big(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\big)^{\mathrm{T}}\big(\tilde{\boldsymbol{Z}}_i\boldsymbol{D}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} + \boldsymbol{\Sigma}\otimes\boldsymbol{C}_i\big)^{-1}\big(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\big)\Big\}.\end{aligned} \tag{9}$$

In the MLME step, we update $\hat{\boldsymbol{\beta}}^{(h)}$ by a generalized least-squares approach, which yields:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(h+1)} \;=\; & \left(\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i^{\mathrm{T}}\Big(\tilde{\boldsymbol{Z}}_i\hat{\boldsymbol{D}}^{(h)}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}^{(h)}\otimes\hat{\boldsymbol{C}}_i^{(h)}\Big)^{-1}\tilde{\boldsymbol{X}}_i\right)^{-1} \\ & \times\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i^{\mathrm{T}}\Big(\tilde{\boldsymbol{Z}}_i\hat{\boldsymbol{D}}^{(h)}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}^{(h)}\otimes\hat{\boldsymbol{C}}_i^{(h)}\Big)^{-1}\tilde{\boldsymbol{y}}_i.\end{aligned} \tag{10}$$

Denote the half-vectorization operator by $\mathrm{vech}(\cdot)$, which represents a column vector obtained by vectorizing only the lower triangular entries of a symmetric matrix. Given the current estimate $\hat{\boldsymbol{\beta}}^{(h+1)}$, we update $\hat{\boldsymbol{\alpha}}^{(h)} = (\mathrm{vech}(\hat{\boldsymbol{D}}^{(h)}), \mathrm{vech}(\hat{\boldsymbol{\Sigma}}^{(h)}), \hat{\boldsymbol{\phi}}^{(h)})$ by the Newton–Raphson method:

$$\hat{\boldsymbol{\alpha}}^{(h+1)} = \hat{\boldsymbol{\alpha}}^{(h)} - \hat{\mathbf{H}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{(h+1/2)^{-1}}\hat{\mathbf{s}}_{\boldsymbol{\alpha}}^{(h+1/2)}, \tag{11}$$

where $\hat{\mathbf{s}}_{\boldsymbol{\alpha}}^{(h+1/2)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\alpha\alpha}}^{(h+1/2)}$ are the score vector $\mathbf{s}_{\boldsymbol{\alpha}}$ and Hessian matrix $\mathbf{H}_{\boldsymbol{\alpha\alpha}}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(h+1)}$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(h)}$. Explicit expressions for elements in $\mathbf{s}_{\boldsymbol{\alpha}}$ and $\mathbf{H}_{\boldsymbol{\alpha\alpha}}$ are given in Appendix.

Iterations of Equations (6), (10) and (11) continue until either the maximum number of iterations or the user-specified convergence tolerance has been achieved.

2.2. Laplacian Procedure

From Function Equation (3) and Definition Equation (5), we have the joint density of $(\boldsymbol{y}_i, \boldsymbol{b}_i)$, denoted by $f(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{\theta}) = \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i, \boldsymbol{R}_i\big)\phi_q(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{D})$, and the marginal density of $\boldsymbol{y}_i$, given by:

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}) = \int (2\pi)^{-(s_i r+q)/2}|\boldsymbol{R}_i|^{-1/2}|\boldsymbol{D}|^{-1/2}\exp\Big\{-\frac{1}{2}g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta})\Big\}d\boldsymbol{b}_i. \tag{12}$$

Laplacian approximation [32,36] is an alternative technique to estimate the marginal densities or posterior predictive densities, which involve integrating out all non-target variables. We next discuss how to adopt the Laplacian approximation to evaluate Equation (12) and develop the corresponding estimation algorithm.

Set an initial guess of random effects $\boldsymbol{b}_i$ to be:

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i(\boldsymbol{y}_i, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{b}_i} f(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{\theta}) = \arg\min_{\boldsymbol{b}_i} g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}).$$

Consider the second-order Taylor expansion of $g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta})$ around $\hat{\boldsymbol{b}}_i$. It yields:

$$\begin{aligned} g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}) &\approx g(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta}) - \dot{g}(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta})(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i) + \frac{1}{2}(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i)^{\mathrm{T}}\ddot{g}(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta})(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i) \\ &\approx g(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i)^{\mathrm{T}}\ddot{g}(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta})(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i), \end{aligned}$$

because $\dot{g}(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta}) = 0$, where the first two partial derivatives of $g(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta})$ with respect to $\boldsymbol{b}_i$ are:

$$\dot{g}(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}) = -2\left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^{\mathrm{T}}}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i}\boldsymbol{R}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i)\big) - \boldsymbol{D}^{-1}\boldsymbol{b}_i\right),$$

and:

$$\ddot{g}(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}) = -2\left(\frac{\partial^2 \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^{\mathrm{T}}}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i}\boldsymbol{R}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i\big) - \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i^{\mathrm{T}}}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i}\boldsymbol{R}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i} - \boldsymbol{D}^{-1}\right),$$

respectively. Notice that the contribution of the term involving the second derivative of $\boldsymbol{\mu}_i$ in $\ddot{g}(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta})$ is usually negligible compared to that involving the product of the first derivative of $\boldsymbol{\mu}_i$ at $\boldsymbol{b}_i = \hat{\boldsymbol{b}}_i$ [37]. We hereby define:

$$\ddot{g}(\boldsymbol{y}_i, \hat{\boldsymbol{b}}_i, \boldsymbol{\theta}) \cong \boldsymbol{G}(\boldsymbol{y}_i, \boldsymbol{\theta}) = 2\left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i^{\mathrm{T}}}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i}^{\mathrm{T}}\boldsymbol{R}_i^{-1}\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i} + \boldsymbol{D}^{-1}\right). \tag{13}$$

Consequently, the Laplacian approximation to log-likelihood Equation (4) is:

$$\begin{aligned}\ell_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) &\cong \log\Big\{\prod_{i=1}^{N}(2\pi)^{-\frac{s_i r+q}{2}}|\boldsymbol{R}_i|^{-\frac{1}{2}}|\boldsymbol{D}|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}g(\boldsymbol{y}_i,\hat{\boldsymbol{b}}_i,\boldsymbol{\theta})\Big\}\\ &\quad\times\int\exp\Big\{-\frac{1}{4}(\boldsymbol{b}_i-\hat{\boldsymbol{b}}_i)^{\mathrm{T}}\ddot{g}(\boldsymbol{y}_i,\hat{\boldsymbol{b}}_i,\boldsymbol{\theta})(\boldsymbol{b}_i-\hat{\boldsymbol{b}}_i)\Big\}d\boldsymbol{b}_i\Big\} \qquad (14)\\ &= -\frac{1}{2}\sum_{i=1}^{N}\Big\{s_i r\log(2\pi)+\log|\boldsymbol{R}_i|+\log|\boldsymbol{D}|+\log\Big|\frac{1}{2}\boldsymbol{G}(\boldsymbol{y}_i,\boldsymbol{\theta})\Big|\\ &\quad+\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta},\hat{\boldsymbol{b}}_i)\big)^{\mathrm{T}}\boldsymbol{R}_i^{-1}\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta},\hat{\boldsymbol{b}}_i)\big)+\hat{\boldsymbol{b}}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\hat{\boldsymbol{b}}_i\Big\}. \qquad (15)\end{aligned}$$

with regard to ML estimation of $\boldsymbol{\theta}$, we can treat it as an optimization problem based on $\ell_{LA}(\boldsymbol{\theta}|\boldsymbol{y})$. Subsequently, we estimate $\boldsymbol{D}$ by taking the first partial derivative of Equation (15) with respect to $\boldsymbol{D}^{-1}$ and setting it to zero, yielding:

$$\hat{\boldsymbol{D}}=N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^{\mathrm{T}}.$$

By maximizing Equation (15), the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\phi}$ react with one another, and thus, we perform an iterative algorithm that proceeds as follows. Given $\hat{\boldsymbol{D}}$ and the current estimates $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\phi}}^{(h)}$, we update the diagonal elements in $\hat{\boldsymbol{\Sigma}}^{(h)}$ by:

$$\hat{\sigma}_{jj}^{(h+1)}=\Big(\sum_{i=1}^{N}s_i\Big)^{-1}\sum_{i=1}^{N}\mathrm{tr}\Big(\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}^{(h)})^{-1}\big(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}_{ij}^{(h+1)}\big)\big(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}_{ij}^{(h+1)}\big)^{\mathrm{T}}\Big),$$

and the off-diagonal elements by:

$$\begin{aligned}\hat{\sigma}_{jl}^{(h+1)} &= \Big(2\sum_{i=1}^{N}s_i\Big)^{-1}\sum_{i=1}^{N}\mathrm{tr}\Big(\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}^{(h)})^{-1}\Big[\big(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}_{ij}^{(h+1)}\big)\big(\boldsymbol{y}_{il}-\hat{\boldsymbol{\mu}}_{il}^{(h+1)}\big)^{\mathrm{T}}\\ &\quad+\big(\boldsymbol{y}_{il}-\hat{\boldsymbol{\mu}}_{il}^{(h+1)}\big)\big(\boldsymbol{y}_{ij}-\hat{\boldsymbol{\mu}}_{ij}^{(h+1)}\big)^{\mathrm{T}}\Big]\Big),\end{aligned}$$

for $j,l=1,\ldots,r$, where $\hat{\boldsymbol{\mu}}_{ij}^{(h+1)}$ is an $s_i\times 1$ subvector consisting of the $((j-1)s_i+1)$-th to the (js_i)-th entries of $\hat{\boldsymbol{\mu}}_i^{(h+1)}=\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}^{(h+1)},\hat{\boldsymbol{b}}_i)$. Unfortunately, equating the first partial derivatives of Equation (15) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, respectively, to zero cannot deduce the updated estimators in closed form. Therefore, we use the nlminb routine [38] to perform a numerical search for updating $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\phi}}^{(h)}$ sequentially. Specifically,

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \arg\min_{\boldsymbol{\beta}}\Big\{\sum_{i=1}^{N}\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta},\hat{\boldsymbol{b}}_i)\big)^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}^{(h+1)}\otimes\boldsymbol{C}_i(\hat{\boldsymbol{\phi}}^{(h)}))^{-1}\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta},\hat{\boldsymbol{b}}_i)\big)\Big\},$$

and:

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{(h+1)} &= \arg\min_{\boldsymbol{\phi}}\Big\{\sum_{i=1}^{N}\Big[\log\Big|\frac{1}{2}\boldsymbol{G}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{(-\boldsymbol{\phi})}^{(h+1)})\Big|+r\log\big|\boldsymbol{C}_i(\boldsymbol{\phi})\big|\\ &\quad+(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i^{(h+1)})^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}^{(h+1)}\otimes\boldsymbol{C}_i(\boldsymbol{\phi}))^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i^{(h+1)})\Big]\Big\}.\end{aligned}$$

2.3. Pseudo-ECM Algorithm

According to the pseudo-data model specified in Equation (8), treating the random effects $\{\boldsymbol{b}_i\}_{i=1}^N$ as latent data, we establish a complete-data framework of the model:

$$\tilde{\boldsymbol{y}}_i|\boldsymbol{b}_i \sim \mathcal{N}_{s_{ir}}(\tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i, \boldsymbol{R}_i), \quad \boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D}), \quad i = 1, \ldots, N.$$

Given the pseudo-complete data $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}_i\}_{i=1}^N$ and $\boldsymbol{b} = \{\boldsymbol{b}_i\}_{i=1}^N$, the log-likelihood function of $\boldsymbol{\theta}$ is:

$$\ell_c^P(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}, \boldsymbol{b}) = \sum_{i=1}^N \log\Big(\phi_{s_{ir}}(\tilde{\boldsymbol{y}}_i|\tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i, \boldsymbol{R}_i)\phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D})\Big). \tag{16}$$

To carry out ML estimation for the MNLMM, we develop an ECM algorithm [33], which is a variant of EM [39], replacing its M steps by several computationally-simpler conditional maximization (CM) steps. It has several appealing features, such as stability of monotone convergence and simplicity of implementation. Hereafter, the procedure is referred to as the pseudo-ECM algorithm, because it is developed under the pseudo-data defined in Equation (7). The proposed implementation approach is outlined below.

E step: Evaluate the expected complete-data log-likelihood Function Equation (16) conditioning on the current estimates $\hat{\boldsymbol{\theta}}^{(h)}$ and the pseudo-responses $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{b}}_i^{(h-1)})$, which linearize the regression function around the previous estimates of mixed effects $(\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{b}}_i^{(h-1)})$ and should be updated at each iteration. This gives rise to the so-called Q-function:

$$\begin{aligned} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(h)}) &= -\frac{1}{2}\sum_{i=1}^N \Big\{ \log|\boldsymbol{\Sigma} \otimes \boldsymbol{C}_i| + \log|\boldsymbol{D}| + \mathrm{tr}\Big((\boldsymbol{\Sigma} \otimes \boldsymbol{C}_i)^{-1}\hat{\boldsymbol{\Omega}}_i^{(h)}\Big) \\ &\quad + \mathrm{tr}(\boldsymbol{D}^{-1}\hat{\boldsymbol{\Psi}}_i^{(h)})\Big\}, \end{aligned} \tag{17}$$

where:

$$\begin{aligned} \hat{\boldsymbol{\Psi}}_i^{(h)} &= E[\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}|\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{\theta}}^{(h)}] = \tilde{\boldsymbol{b}}_i^{(h)}\tilde{\boldsymbol{b}}_i^{(h)^{\mathrm{T}}} + (\hat{\boldsymbol{D}}^{(h)^{-1}} + \tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\hat{\boldsymbol{R}}_i^{(h)^{-1}}\tilde{\boldsymbol{Z}}_i)^{-1}, \\ \hat{\boldsymbol{\Omega}}_i^{(h)} &= E[\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i^{\mathrm{T}}|\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{\theta}}^{(h)}] = \tilde{\boldsymbol{e}}_i^{(h)}\tilde{\boldsymbol{e}}_i^{(h)^{\mathrm{T}}} + \tilde{\boldsymbol{Z}}_i(\hat{\boldsymbol{D}}^{(h)^{-1}} + \tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\hat{\boldsymbol{R}}_i^{(h)^{-1}}\tilde{\boldsymbol{Z}}_i)^{-1}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} \end{aligned}$$

with $\hat{\boldsymbol{R}}_i^{(h)} = \hat{\boldsymbol{\Sigma}}^{(h)} \otimes \boldsymbol{C}_i(\hat{\boldsymbol{\phi}}^{(h)})$, $\tilde{\boldsymbol{b}}_i^{(h)} = E[\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{\theta}}^{(h)}] = \hat{\boldsymbol{D}}^{(h)}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}}(\tilde{\boldsymbol{Z}}_i\hat{\boldsymbol{D}}^{(h)}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} + \hat{\boldsymbol{R}}_i^{(h)})^{-1}(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}^{(h)})$, and $\tilde{\boldsymbol{e}}_i^{(h)} = E[\tilde{\boldsymbol{e}}_i|\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{\theta}}^{(h)}] = \tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\tilde{\boldsymbol{b}}_i^{(h)}$, where $\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_i(\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{b}}_i^{(h-1)})$ represents the updated pseudo-responses.

CM step: Update the current estimates $\hat{\boldsymbol{\beta}}^{(h)}$, $\hat{\boldsymbol{D}}^{(h)}$, $\hat{\boldsymbol{\Sigma}}^{(h)}$ and $\hat{\boldsymbol{\phi}}^{(h)}$ by maximizing the Q-function Equation (17). We obtain:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(h+1)} &= \Big(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\mathrm{T}} \hat{\boldsymbol{R}}_i^{(h)^{-1}} \tilde{\boldsymbol{X}}_i\Big)^{-1}\Big(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\mathrm{T}} \hat{\boldsymbol{R}}_i^{(h)^{-1}} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{Z}}_i \tilde{\boldsymbol{b}}_i^{(h)})\Big),\\
\hat{\boldsymbol{D}}^{(h+1)} &= N^{-1}\sum_{i=1}^{N} \hat{\boldsymbol{\Psi}}_i^{(h)},\\
\hat{\sigma}_{jl}^{(h+1)} &= \begin{cases} \Big(\sum_{i=1}^{N} s_i\Big)^{-1} \sum_{i=1}^{N} \mathrm{tr}\Big(\hat{\boldsymbol{C}}_i(\hat{\boldsymbol{\phi}}^{(h)})^{-1} \hat{\boldsymbol{\omega}}_{ijl}^{(h+1/2)}\Big), & \text{for } j = l,\\ \Big(2\sum_{j=1}^{N} s_i\Big)^{-1} \sum_{i=1}^{N} \mathrm{tr}\Big(\hat{\boldsymbol{C}}_i(\hat{\boldsymbol{\phi}}^{(h)})^{-1}\Big[\hat{\boldsymbol{\omega}}_{ijl}^{(h+1/2)} + \hat{\boldsymbol{\omega}}_{ilj}^{(h+1/2)}\Big]\Big), & \text{for } j \neq l, \end{cases}\\
\hat{\boldsymbol{\phi}}^{(h+1)} &= \arg\min_{\boldsymbol{\phi}} \Big\{ r\sum_{i=1}^{N} \log|\boldsymbol{C}_i| + \sum_{i=1}^{N} \mathrm{tr}\Big((\hat{\boldsymbol{\Sigma}}^{(h+1)} \otimes \boldsymbol{C}_i)^{-1} \hat{\boldsymbol{\Omega}}_i^{(h+1/2)}\Big)\Big\},
\end{aligned}$$

where $\hat{\boldsymbol{\omega}}_{ijl}^{(h+1/2)}$ is an $s_i \times s_i$ matrix consisting of the $((j-1)s_i+1)$-th to the (js_i)-th columns and rows of $\hat{\boldsymbol{\Omega}}_i^{(h)}$ in which $\boldsymbol{\beta}$ and $\boldsymbol{D}$ have been replaced by their updated estimates at the $h+1$ iteration. Besides, $\hat{\boldsymbol{\Omega}}_i^{(h+1/2)}$ in the above optimization function for $\hat{\boldsymbol{\phi}}^{(h+1)}$ is $\hat{\boldsymbol{\Omega}}_i^{(h)}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(h+1)}$, except for $\boldsymbol{\phi}$.

Given $\{\hat{\boldsymbol{b}}_i^{(0)}\}_{i=1}^{N}$ and $\hat{\boldsymbol{\theta}}^{(0)}$, we implement the pseudo-ECM algorithm until the user's specified convergence criterion satisfies. Analogous to the PNLS-MLME method, this algorithm is established under the pseudo-data scenario. Hence, the resulting approximate log-likelihood value can be obtained by using Equation (9).

2.4. Monte Carlo EM Algorithm

We offer a Monte Carlo (MC) version of the EM algorithm [40] for ML estimation of Model Equation (1) and evaluate the observed log-likelihood Equation (4) via the MC integration. The MCEM is a modification of the EM algorithm in which the E step is computed numerically through a large number of simulated samples.

Given the complete data $(\boldsymbol{y}, \boldsymbol{b})$, the log-likelihood function of $\boldsymbol{\theta}$ for the MNLMM can be expressed as:

$$\ell_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{b}) = \sum_{i=1}^{N} \log\Big(\phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i), \boldsymbol{R}_i\big)\phi_q\big(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D}\big)\Big). \tag{18}$$

In the E step, we compute the expectation of complete data log-likelihood Function Equation (18) to yield the Q-function:

$$\begin{aligned}
Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(h)}) &= \sum_{i=1}^{N} \int \log \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i), \boldsymbol{R}_i\big) P(\boldsymbol{b}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}) d\boldsymbol{b}_i\\
&\quad + \sum_{i=1}^{N} \int \log \phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D}) P(\boldsymbol{b}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}) d\boldsymbol{b}_i.
\end{aligned} \tag{19}$$

Obviously, Equation (19) cannot be written in closed form, since the conditional distribution of $\boldsymbol{b}_i$ given $\boldsymbol{y}_i$:

$$P(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) \propto \exp\Big\{-\frac{1}{2}\Big[\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\big)^{\mathrm{T}} \boldsymbol{R}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\big) + \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{D}^{-1} \boldsymbol{b}_i\Big]\Big\} \tag{20}$$

has no standard form. To simulate random samples from Equation (20), we perform the Metropolis–Hastings (M-H) algorithm [41] with the proposal distribution:

$$\boldsymbol{b}_i^{(m+1)} \sim \mathcal{N}_q\big(\boldsymbol{b}_i^{(m)}, \boldsymbol{G}^{-1}\big(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big)\big), \tag{21}$$

where $\boldsymbol{G}^{-1}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)})$ is the inverse matrix of $\boldsymbol{G}(\boldsymbol{y}_i, \boldsymbol{\theta})$ given in Equation (13) and evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(h)}$. Note that the idea of considering such a proposal distribution comes from the integration of Equation (14) over $\boldsymbol{b}_i$, which is, up to a multiplicative constant, approximately equal to a $\mathcal{N}(\hat{\boldsymbol{b}}_i, \boldsymbol{G}^{-1}\big(\boldsymbol{y}_i, \boldsymbol{\theta}\big))$. We have the probability $\min\big\{1,\ P(\boldsymbol{b}_i^{(m+1)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)})/P(\boldsymbol{b}_i^{(m)}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)})\big\}$ to accept the new generation $\boldsymbol{b}_i^{(m+1)}$, but otherwise to set $\boldsymbol{b}_i^{(m+1)} = \boldsymbol{b}_i^{(m)}$. After having a set of converged MC samples $\{\boldsymbol{b}_i^{(m)}\}_{m=1}^M$, the random effects $\boldsymbol{b}_i$, as well as their function $f(\boldsymbol{b}_i)$ in Equation (19) can be estimated by $\hat{\boldsymbol{b}}_i^{(h)} = \sum_{m=1}^M \boldsymbol{b}_i^{(m)}/M$ and $E[f(\boldsymbol{b}_i)|\boldsymbol{y}_i, \boldsymbol{\theta}] = \sum_{m=1}^M f(\boldsymbol{b}_i^{(m)})/M$, respectively, at each iteration.

In the M step, we find the limited value of the obtained Q-function Equation (19) by equating the following functions:

$$\frac{\partial Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(h)})}{\partial \boldsymbol{D}} = \sum_{i=1}^N \frac{\partial}{\partial \boldsymbol{D}} E\big[\log \phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D})|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big] \tag{22}$$

and:

$$\frac{\partial Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(h)})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^N \frac{\partial}{\partial \boldsymbol{\alpha}} E\big[\log \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i), \boldsymbol{R}_i\big)|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big] \tag{23}$$

to zeros, where $\boldsymbol{\alpha} = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\phi}\}$. By allowing differentiation under the integral sign for Equation (22), we update the estimate of $\boldsymbol{D}$ by:

$$\hat{\boldsymbol{D}}^{(h+1)} = \frac{1}{N}\sum_{i=1}^N E\big[\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big] \cong \frac{1}{N}\sum_{i=1}^N \Big\{\frac{1}{M}\sum_{i=1}^M \boldsymbol{b}_i^{(m)}\boldsymbol{b}_i^{(m)^{\mathrm{T}}}\Big\}.$$

Since solving Equation (23) is analytically intractable, we perform a profile approximate Q-function approach, which updates $\hat{\boldsymbol{\beta}}^h$, $\hat{\boldsymbol{\Sigma}}^{(h)}$ and $\hat{\boldsymbol{\phi}}^{(h)}$ by a sequential optimization procedure as the Laplacian method described in Section 2.2. It gives:

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^N E\big[\log \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i), \hat{\boldsymbol{R}}_i^{(h)}\big)|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big], \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}^{(h+1)} = \arg\max_{\boldsymbol{\Sigma}} \sum_{i=1}^N E\big[\log \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}^{(h+1)}, \boldsymbol{b}_i), \boldsymbol{\Sigma} \otimes \hat{\boldsymbol{C}}_i^{(h)}\big)|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big], \tag{25}$$

and:

$$\hat{\boldsymbol{\phi}}^{(h+1)} = \arg\max_{\boldsymbol{\phi}} \sum_{i=1}^N E\big[\log \phi_{s_i r}\big(\boldsymbol{y}_i|\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}^{(h+1)}, \boldsymbol{b}_i), \hat{\boldsymbol{\Sigma}}^{(h+1)} \otimes \boldsymbol{C}_i(\boldsymbol{\phi})\big)|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)}\big]. \tag{26}$$

Consequently, the marginal log-likelihood can be approximated as:

$$\begin{aligned}\ell_{\mathrm{MC}}(\boldsymbol{\theta}|\boldsymbol{y}) &= -\frac{1}{2}\sum_{i=1}^N \Big\{(s_i r + q)\log(2\pi) + s_i \log|\boldsymbol{\Sigma}| + r\log|\boldsymbol{C}_i| + \log|\boldsymbol{D}|\Big\} \\ &\quad -\frac{1}{2M}\sum_{i=1}^N\sum_{m=1}^M g(\boldsymbol{y}_i, \boldsymbol{b}_i^{(m)}, \boldsymbol{\theta}).\end{aligned}$$

According to an alternative hierarchy of the MNLMM,

$$\boldsymbol{y}_i|\boldsymbol{\eta}_i \sim \mathcal{N}_{s_i r}(\boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{x}_i), \boldsymbol{R}_i), \quad \boldsymbol{\eta}_i \sim \mathcal{N}_s(\boldsymbol{A}_i\boldsymbol{\beta}, \boldsymbol{B}_i\boldsymbol{D}\boldsymbol{B}_i^{\mathrm{T}}), \quad \text{for } i = 1, \ldots, N,$$

the MCEM algorithm that deals with Monte Carlo integration directly on the individual parameters $\boldsymbol{\eta}_i$ rather than subject-specific random effects $\boldsymbol{b}_i$ can yield an explicit estimator for the fixed effects $\boldsymbol{\beta}$. However, such an implementation may not be feasible in the framework of MNLMMs due to the possible singularity of $\boldsymbol{B}_i\boldsymbol{D}\boldsymbol{B}_i^{\mathrm{T}}$.

2.5. Importance Sampling EM Algorithm

Importance sampling (IS) is an alternative way of performing MC integration. We provide an ISEM algorithm, which modifies MC approximation of Equation (19) in the E step of the MCEM algorithm by using the IS method. To implement the ISEM algorithm, we first choose an appropriate envelope distribution from which the samples are simulated and the importance weights calculated. Like that used in the M-H algorithm, Equation (21) is a natural consideration for the envelop distribution. As suggested by [35], an envelop distribution could be a mixture of two multivariate normal distributions with pdf:

$$\lambda(\boldsymbol{b}_i) = P_0\phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \hat{\boldsymbol{D}}^{(h)}) + (1 - P_0)\phi_q\big(\boldsymbol{b}_i|\hat{\boldsymbol{b}}_i^{(h)}, \boldsymbol{G}^{-1}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(h)})\big), \tag{27}$$

where the mixing proportion $0 \leq P_0 \leq 1$ is a pre-specified value.

Notably, ISEM can be performed to evaluate the expected values of any functions of unobservable $\{\boldsymbol{b}_i\}_{i=1}^N$, e.g., $f(\boldsymbol{b}_i) = \boldsymbol{b}_i$ and $f(\boldsymbol{b}_i) = \boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}$. It follows that:

$$E\big[f(\boldsymbol{b}_i)|\boldsymbol{y}_i, \boldsymbol{\theta}\big] = \int f(\boldsymbol{b}_i)f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\theta})d\boldsymbol{b}_i = \frac{\int f(\boldsymbol{b}_i)f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})f(\boldsymbol{b}_i|\boldsymbol{D})d\boldsymbol{b}_i}{\int f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})f(\boldsymbol{b}_i|\boldsymbol{D})d\boldsymbol{b}_i}. \tag{28}$$

Having obtained a sufficient number of random effects, denoted by $\{\boldsymbol{b}_i^{(m)}\}_{i=1}^N$, $m = 1, \ldots, M$, we adopt the ratio of two MC approximations using IS from Equation (27) to estimate Equation (28), given by:

$$E\big[f(\boldsymbol{b}_i)|\boldsymbol{y}_i, \boldsymbol{\theta}\big] \cong \frac{\sum_{m=1}^M f(\boldsymbol{b}_i^{(m)})f(\boldsymbol{y}_i|\boldsymbol{b}_i^{(m)}, \boldsymbol{\theta})f(\boldsymbol{b}_i^{(m)}|\boldsymbol{D})/\lambda(\boldsymbol{b}_i^{(m)})}{\sum_{m=1}^M f(\boldsymbol{y}_i|\boldsymbol{b}_i^{(m)}, \boldsymbol{\theta})f(\boldsymbol{b}_i^{(m)}|\boldsymbol{D})/\lambda(\boldsymbol{b}_i^{(m)})}. \tag{29}$$

In the E step, given the current estimates of parameters $\hat{\boldsymbol{\theta}}^{(h)}$, we compute Equation (19) in which the required conditional moments of latent data $\boldsymbol{b}$ can be approximated based on Equation (29). In the M step, we update each entry of $\hat{\boldsymbol{\theta}}^{(h)}$ by maximizing the Q-function. Indeed, the ISEM procedure works conceptually similarly to that of MCEM: only $\hat{\boldsymbol{D}}^{(h+1)}$ shows an explicit solution, while $\hat{\boldsymbol{\beta}}^{(h+1)}$, $\hat{\boldsymbol{\Sigma}}^{(h+1)}$ and $\hat{\boldsymbol{\phi}}^{(h+1)}$ are obtained through sequential optimization solutions via Equations (24)–(26). The IS approximation to the marginal log-likelihood is:

$$\begin{aligned}\ell_{IS}(\boldsymbol{\theta}|\boldsymbol{y}) \;\cong\; & -\frac{1}{2}\sum_{i=1}^N \Big\{ s_i \log|\boldsymbol{\Sigma}| + r\log|\boldsymbol{C}_i| + \log|\boldsymbol{D}| \Big\} \\ & + \sum_{i=1}^N \log\Big\{ \frac{1}{M}\sum_{m=1}^M \big[\exp\big\{ -\frac{1}{2} g(\boldsymbol{y}_i, \boldsymbol{b}_i^{(m)}, \boldsymbol{\theta}) \big\} f(\boldsymbol{b}_i^{(m)}|\boldsymbol{D})/\lambda(\boldsymbol{b}_i^{(m)}) \big] \Big\}.\end{aligned}$$

2.6. Expected Information Matrix

For Model Equation (8), denoting by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\text{vech}(\boldsymbol{D}), \text{vech}(\boldsymbol{\Sigma}), \boldsymbol{\phi})$, the expected information matrix of $\boldsymbol{\theta}$ obtained by taking the expectation of the negative Hessian matrix can be expressed as:

$$\mathbf{J}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\alpha}} \\ \mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\alpha}}^{\mathrm{T}} & \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \end{bmatrix}, \tag{30}$$

where $\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\mathrm{T}} \tilde{\boldsymbol{\Lambda}}_i^{-1} \tilde{\boldsymbol{X}}_i$, $\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\alpha}} = \mathbf{0}$, and $\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ is a $g \times g$ information matrix whose (l, s)-th entry is $[\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]_{ls} = 2^{-1} \sum_{i=1}^{N} \text{tr}(\tilde{\boldsymbol{\Lambda}}_i^{-1} \dot{\tilde{\boldsymbol{\Lambda}}}_{il} \tilde{\boldsymbol{\Lambda}}_i^{-1} \dot{\tilde{\boldsymbol{\Lambda}}}_{is})$, for $l, s = 1, \ldots, g$, $g = q(q+1)/2 + r(r+1)/2 + \dim(\boldsymbol{\phi})$, with $\dot{\tilde{\boldsymbol{\Lambda}}}_{il}$ being $\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)}$ given in (A.1) with $\hat{\boldsymbol{\theta}}^{(h)}$ replaced by $\boldsymbol{\theta}$. Consequently, the asymptotic variance-covariance matrix of $\boldsymbol{\theta}$ can be approximated by the inverse of information Matrix Equation (30), denoted by $\mathbf{J}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$. The resulting standard errors of parameters are the square roots of diagonal entries of $\mathbf{J}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

2.7. Initialization

When implementing iterative procedures, a common difficulty encountered in practice is that the algorithm is painfully slow or even non-convergent. Such a computational problem may occur in handling ML estimation of the MNLMM, especially when the data are too sparse or the dimension of random effects is over-specified. To overcome this potential problem, a default procedure of automatically creating a set of good initial values is summarized below.

(i) A direct way of obtaining the initial value for $\boldsymbol{\beta}$ is to fit the NLMMs to each outcome variable separately by using the nlme R package [12].

(ii) Using the fitting results of NLMMs for each outcome, we take the initial value $\hat{\boldsymbol{D}}^{(0)}$ as a (block) diagonal form with the diagonal entry being the variances (covariances) of random effects under the fitted NLMMs.

(iii) For the initial value for $\boldsymbol{\Sigma}$, we use the sample variance-covariance matrix of the data. That is, take $\hat{\boldsymbol{\Sigma}}^{(0)} = \sum_{i=1}^{N} \sum_{t=1}^{s_i} (\boldsymbol{y}_{i\cdot t} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{i\cdot t} - \bar{\boldsymbol{y}})^{\mathrm{T}} / (\sum_{i=1}^{N} s_i - 1)$, where $\boldsymbol{y}_{i\cdot t} = (y_{i1t}, \cdots, y_{irt})^{\mathrm{T}}$ and $\bar{\boldsymbol{y}} = (\sum_{i=1}^{N} s_i)^{-1} (\sum_{i=1}^{N} \sum_{t=1}^{s_i} y_{i1t}, \cdots, \sum_{i=1}^{N} \sum_{t=1}^{s_i} y_{irt})^{\mathrm{T}}$.

(iv) The initial values for $\boldsymbol{\phi}$, depending on the structure, are simply chosen to give a condition of nearly uncorrelated errors.

3. Application: ACTG 315 Data

We present a comparison of the five algorithms via a real data example from the AIDS Clinical Trial Group protocol 315 (ACTG 315) study developed by the Immunology Research Agenda Committee of the U.S. National Institute of Allergy and Infectious Disease, the ACTG sponsor. The study design and recruitment of participants (patients) were conducted by University Hospitals of Cleveland, Rush-Presbyterian-St. Luke's Medical Center and University of Colorado Health Science Center. In the study, 53 human immunodeficiency virus type 1 (HIV-1)-infected patients were recruited, and their plasma HIV-1 RNA (viral load) copies and $CD4^+$ T cell counts were repeatedly measured at Days

0, 2, 7, 10, 14, 28, 56, 84, 168 and 196 after the start of treatment. A more detailed description of the study can be found in [42,43].

HIV-1 infection is associated with progressive and profound loss of immune function that places infected persons at enhanced risk for opportunistic infections, and even death. A reaction in HIV-1-related immune deficiency can be characterized by decreases in the numbers of circulating $CD4^+$ T helper lymphocytes. $CD4^+$ T cells in blood decline to a lower level after HIV-1 infection and may recover to a high level after antiviral therapies suppress viral load. Generally, there is a negative correlation between the virologic marker (measured by HIV-1 RNA) and the immunologic marker (measured by $CD4^+$ T cells) during antiviral treatments. As a consequence, a joint analysis of HIV-1 RNA and $CD4^+$ counts is helpful to take the evolution of the correlation among responses over time into account. The data have been analyzed by [44–47] using different modeling approaches.

As a part of the clinical trial on 53 patients, a total of 48 patients were recruited in our analysis after excluding four early drop-out patients and one due to a plasma HIV-1 RNA pattern that suggested intermittent adherence to study therapy. To stabilize the variances and to reduce the strong skewness among the two makers, a base-10 logarithmic transformation is made for HIV-1 RNA and a square-root transformation for $CD4^+$ T cells. Both transformations are widely used in HIV-AIDS clinical trials. Let $y_{i1,k}$ and $y_{i2,k}$ be $\log_{10}$ RNA and $\text{CD4}^{0.5}$ markers, respectively, at the k-th time point for patient i. We consider the following bivariate nonlinear mixed-effects model for $y_{i1,k}$ and $y_{i2,k}$:

$$\begin{aligned} y_{i1,k} &= \log_{10}\left(\exp\{(\beta_1 + b_{i1}) + \beta_2 t_{ik}\} + \exp\{\beta_3 \text{rna}_i\}\right) + e_{i1,k}, \\ y_{i2,k} &= (\beta_4 + b_{i2})/(1 + \exp\{(\beta_5 - t_{ik})/\beta_6\}) + e_{i2,k}, \end{aligned} \tag{31}$$

where $t_{ik} = \text{day}_{ik}/7$ is the k-th visited time point (week) for patient i; rna_i is the $\log_{10}$ RNA levels for patient i at the start of the study; (b_{i1}, b_{i2}) are the bivariate normally-distributed random effects; and $(\boldsymbol{e}_{i1}^{\mathrm{T}}, \boldsymbol{e}_{i2}^{\mathrm{T}}) = (e_{i1,1}, \ldots, e_{i1,s_i}, e_{i2,1}, \ldots, e_{i2,s_i})$ are the within-subject errors following a multivariate normal distribution with zero mean and variance-covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{C}_i$. Because the baseline RNA is a significant covariate in the ACTG 315 study [47], it should be incorporated into the analysis. To account for the extra autocorrelation caused by within-patient dependence among unequally-spaced occasions, we employ a continuous order-one autoregressive structure, *i.e.*, $\boldsymbol{C}_i = [\phi^{|t_{ik} - t_{ik'}|}]$, for the across-occasion covariance matrix of within-subject errors.

According to the standard formulation in Equation (2), we specify:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \text{rna}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_3 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^{\mathrm{T}}, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^{\mathrm{T}},$$

and $\boldsymbol{b}_i = (b_{i1}, b_{i2})^{\mathrm{T}}$, where $\boldsymbol{I}_d$ is a diagonal matrix of order d. Define:

$$\xi_1 = (\exp\{\eta_1 + \eta_2 t\} + \exp\{\eta_3\})^{-1}/\log(10), \text{ and } \xi_2 = (1 + \exp\{(\eta_5 - t)/\eta_6\})^{-1}, \tag{32}$$

where $\eta_1 = \beta_1 + b_{i1}$, $\eta_2 = \beta_2$, $\eta_3 = \beta_3 \text{rna}_i$, $\eta_4 = \beta_4 + b_{i2}$, $\eta_5 = \beta_5$ and $\eta_6 = \beta_6$. The first derivatives of μ_1 and μ_2 specified in Equation (31) with respect to $\boldsymbol{\eta}$ are:

$$\dot{\mu}_1 = \frac{\partial \mu_1}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \xi_1 \exp\{\eta_1 + \eta_2 t\} \\ \xi_1 t \exp\{\eta_1 + \eta_2 t\} \\ \xi_1 \exp\{\eta_3\} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \text{ and } \dot{\mu}_2 = \frac{\partial \mu_2}{\partial \boldsymbol{\eta}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \xi_2 \\ -\mu_2 \exp\{(\eta_5 - t)/\eta_6\}\xi_2/\eta_6 \\ \mu_2 \exp\{(\eta_5 - t)/\eta_6\}(\eta_5 - t)\xi_2/\eta_6^2 \end{bmatrix}.$$

The first derivative of mean function $\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) = (\mu_1, \mu_2)$ with respect to $\boldsymbol{b}_i$ is:

$$\frac{\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)}{\partial \boldsymbol{b}_i} = \begin{bmatrix} \boldsymbol{\xi}_1 \exp\{(\beta_1 + b_{i1}) + (\beta_2 + b_{i2})\boldsymbol{t}_i\} & \mathbf{0}_{s_i} \\ \mathbf{0}_{s_i} & \boldsymbol{\xi}_2 \end{bmatrix},$$

where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are $s_i \times 1$ vectors composed of ξ_1 and ξ_2 given by Equation (32) with t replaced by a $s_i \times 1$ occasion vector $\boldsymbol{t}_i$ of the i-th patient.

Table 1 presents the parameter estimates and their standard deviations (in parentheses) from the five computational methods, namely PNLS-MLMM, Laplacian, pseudo-ECM, MCEM with 500 Monte Carlo samples and ISEM with mixing proportion $P_0 = 0.5$. When employing the ISEM algorithm, several choices of the mixing proportion P_0, ranging from 0–1 with and increment of 0.1, are considered. To save space, we reported only the result for $P_0 = 0.5$, as it yields the maximized log-likelihood value. The results indicate that the five methods can give very similar estimates and the significance of model parameters. According to the estimates of $\boldsymbol{\Sigma} = [\sigma_{jl}]$, the estimated correlation of $\log_{10}$RNA and CD4$^{0.5}$ ranges from -0.13–-0.18 (around), confirming a negative relationship between the virologic and immunologic markers. The between-patient correlations of the two responses have no statistical significance based on the estimates of $\boldsymbol{D}$. The estimate of autoregressive parameter ϕ is significantly different from zero, revealing an existence of autocorrelation among the within-patient variability. Figure 1 displays the observations and estimated mean curves in which the covariate is set to be the average of baseline RNA values of all patients for the five computational methods. Judging from the figure, the considered logarithmic and logistic curves in Equation (31) are reasonable functions to describe the evolutions of RNA in the $\log_{10}$ scale and CD4 in the square-root scale over time. The trend of $\log_{10}$RNA decreases at the beginning due to the rapid growth of CD4$^{0.5}$ cells in the early days of antiviral therapies. After nearly four weeks, the decline pattern on $\log_{10}$RNA and the growth pattern on CD4$^{0.5}$ become slow and smooth. As an illustration, the fitted values obtained by the five methods together with the observations for seven randomly-selected patients are displayed in Figure 2. As anticipated, the fitted trajectories for each patient show the slight difference among the five estimating procedures. Generally, they adapt the trend along observed repeated measures, but some of configurations are not ideally captured. It is known that the viral load (RNA copies) and CD4 counts are highly variable immune system markers, making them difficult to fit.

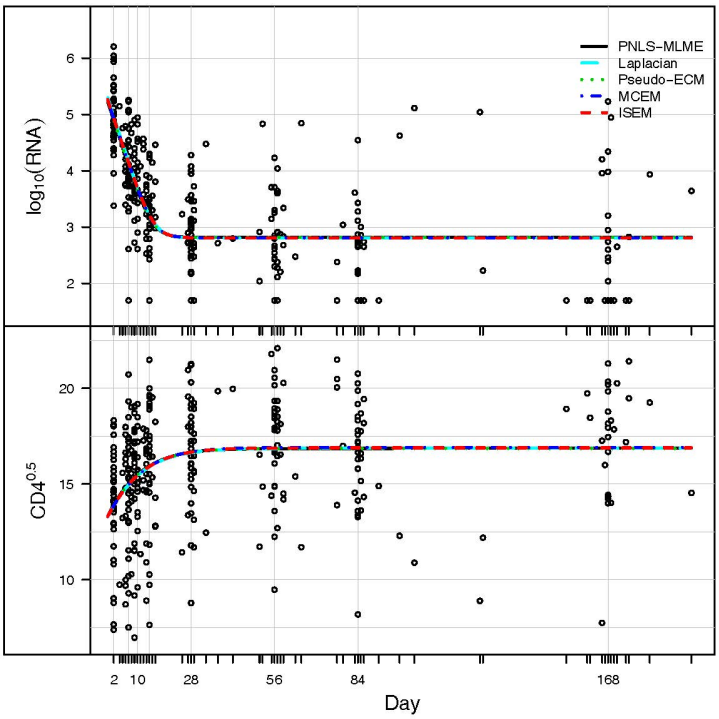

Figure 1. The $\log_{10}(\text{RNA})$ and $\text{CD4}^{0.5}$ observations ($\circ$) with the estimated mean curves against time (in days) from ML estimation using the five proposed procedures.

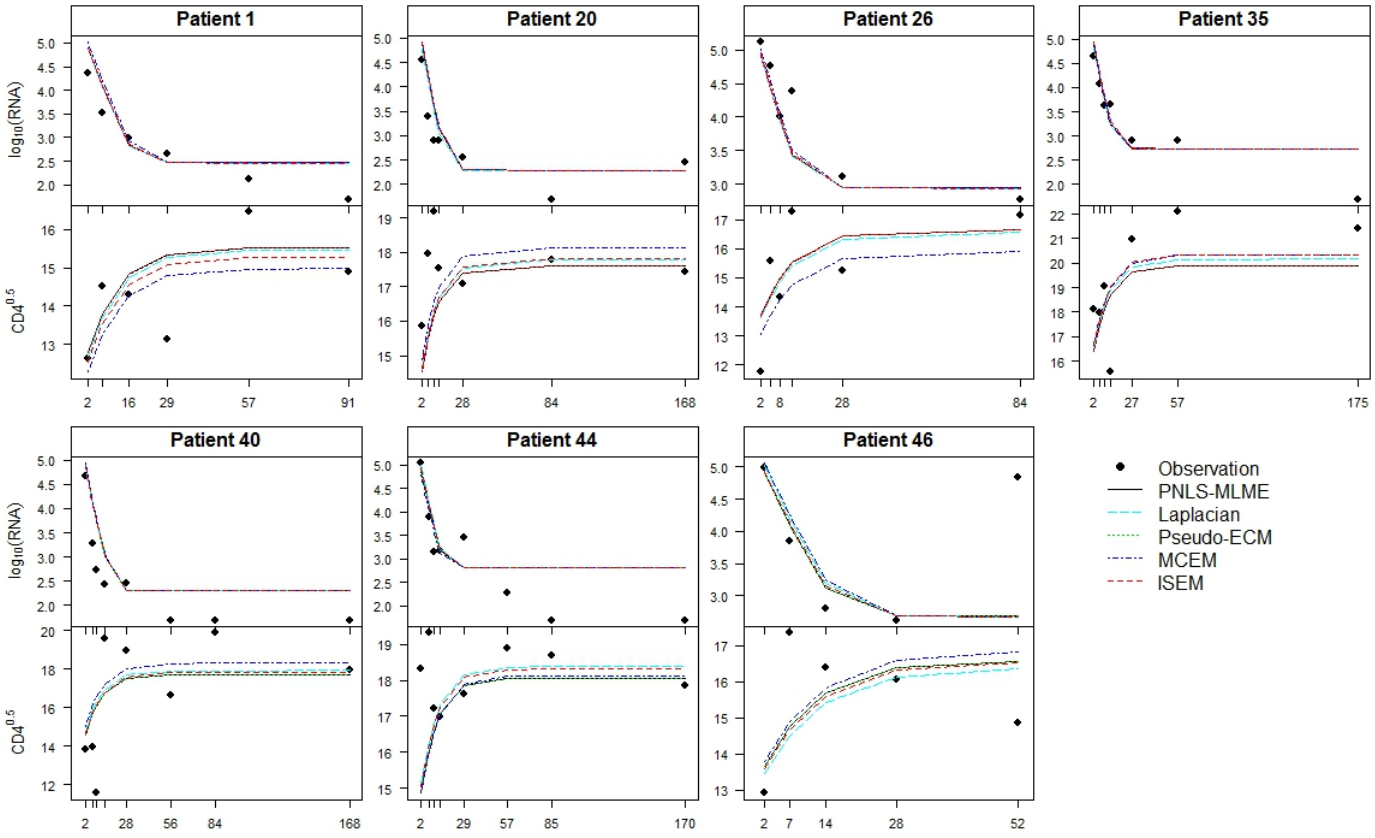

Figure 2. The fitted values obtained by the five proposed procedures together with the observations ($\bullet$) of $\log_{10}(\text{RNA})$ and $\text{CD4}^{0.5}$ for seven randomly-selected patients.

Table 1. Estimation results for AIDS Clinical Trial Group protocol 315 (ACTG 315) data. PNLS, penalized nonlinear least squares; MLME, multivariate linear mixed-effects; ECM, expectation conditional maximization; MCEM, Monte Carlo EM; ISEM, importance sampling EM.

Parameter	PNLS-MLME	Laplacian	Pseudo-ECM	MCEM	ISEM
β_1	12.0477	12.9800	12.0485	12.0784	12.114
	(0.2513)	(0.2858)	(0.2530)	(0.2626)	(0.2652)
β_2	−2.6558	−2.6476	−2.6543	−2.6198	−2.6069
	(0.1781)	(0.1970)	(0.1777)	(0.1950)	(0.1992)
β_3	1.3039	1.3001	1.3039	1.3012	1.3000
	(0.0274)	(0.0248)	(0.0273)	(0.0253)	(0.0249)
β_4	16.8604	16.8577	16.8605	16.8875	16.9058
	(0.3911)	(0.3340)	(0.3914)	(0.3863)	(0.3829)
β_5	−1.7324	−1.7791	−1.7312	−1.7721	−1.7643
	(0.4936)	(0.4590)	(0.4930)	(0.4632)	(0.4585)
β_6	1.3081	1.3514	1.3078	1.3604	1.3463
	(0.3262)	(0.2899)	(0.3259)	(0.2972)	(0.2896)
d_{11}	0.0000	0.7457	0.0583	0.1183	0.1398
	(0.4665)	(0.5763)	(0.4753)	(0.4673)	(0.4612)
d_{21}	−0.0020	−0.1400	0.0144	−0.2386	0.0838
	(0.5414)	(0.5203)	(0.5479)	(0.5401)	(0.5295)
d_{22}	4.7425	3.8251	4.7585	5.4602	5.4894
	(1.3803)	(0.9953)	(1.3826)	(1.3561)	(1.3361)
σ_{11}	0.4655	0.4267	0.4622	0.4379	0.4329
	(0.0458)	(0.0411)	(0.0455)	(0.0420)	(0.0414)
σ_{21}	−0.2232	−0.1738	−0.2164	−0.2185	−0.2225
	(0.0965)	(0.0747)	(0.0962)	(0.0786)	(0.0754)
σ_{22}	5.7063	3.5558	5.6929	3.8956	3.6033
	(0.5991)	(0.3520)	(0.5980)	(0.3874)	(0.3541)
ϕ	0.6824	0.5447	0.6818	0.5674	0.5343
	(0.0311)	(0.0422)	(0.0312)	(0.0400)	(0.0425)

Furthermore, the approximate values of log-likelihood function for Model Equation (31) evaluated at the ML estimates $\hat{\boldsymbol{\theta}}$ obtained respectively by the five estimation procedures are reported in Table 2. To assess the accuracy of the approximations of the log-likelihood function, we also perform the double integral in log-likelihood Function Equation (4) by plugging the corresponding $\hat{\boldsymbol{\theta}}$ into Equation (4) and using the integrate routine in the R package to get the exact log-likelihoods. The exact log-likelihood values together with the absolute differences (AD) between the approximate and exact values are also listed in Table 2. Roughly, the log-likelihood values under the five approximation methods are similar

and close to their corresponding exact values. In this example, the pseudo-ECM yields the most precise evaluation, followed by Laplacian, MCEM, ISEM and PNLS-MLME.

Table 2. Approximate and exact log-likelihood functions for the fitted Model Equation (31) under the five estimation methods. AD, absolute difference.

	PNLS-MLME	**Laplacian**	**Pseudo-ECM**	**MCEM**	**ISEM**
Approximate	−974.360	−986.794	−974.592	−966.763	−1010.370
Exact	−1063.338	−991.754	−978.269	−981.384	−978.758
AD	88.978	4.96	3.677	14.621	31.612

Although the proposed five algorithms can provide quite similar estimates of model parameters, as well as the fitted mean profiles shown in Figures 1 and 2, we should give the following remarks. The PNLS-MLMM and Laplacian methods involve solving the fixed effects $\boldsymbol{\beta}$ and the modes of random effects $\{\boldsymbol{b}_i\}_{i=1}^{N}$ by implementing optimal iterative procedures. Thus, the two methods are very sensitive to initial values and may suffer from slow or even non-convergence due to singularity of variance-covariance matrices, especially when unnecessary random effects are included in the model. The MCEM and ISEM methods spend more time in generating an adequate number of samples of random effects to evaluate the required conditional expectations. Overall, the pseudo-ECM algorithm is the best method in terms of computational efficiency in this study. However, all of the proposed methods may get trapped in one of many local maxima of the log-likelihood function. To assess the stability of the resulting estimates, a variety of initial values should be employed when implementing the algorithms. The global optimal solution is obtained by choosing the one with the largest log-likelihood value.

4. Simulation Study

In this section, two simulation studies with data generated from two models with linear and nonlinear profiles, respectively, are undertaken to compare the performance of the five algorithmic procedures for fitting the MNLMM. The performance comparison includes the convergence efficiency in terms of the number of iterations and consumed CPU time, the accuracy of parameter estimates and the precision of log-likelihood approximation. All computations were carried out by R package 2.13.1 in a Win32 environment of a desktop PC machine with a 3.40-GHz/Intel Core(TM) i7-2600 CPU Processor and 4.0 GB RAM.

4.1. Bivariate Linear Case

To perform an evaluation of the exact log-likelihood values that is tractable, in this simulation, we restrict ourselves to generating datasets from the following bivariate LMM:

$$\begin{aligned} y_{i1k} &= \beta_1 + b_{i1} + \beta_2 t_k + e_{i1k}, \\ y_{i2k} &= \beta_3 + (\beta_4 + b_{i2}) t_k + e_{i2k}, \end{aligned} \tag{33}$$

for $i = 1, \ldots, N$ and $k,\ t_k = 1, \ldots, 7$. Following the standard notation for Model Equation (1) along with Assumption Equation (2), we set $\boldsymbol{A}_i = \boldsymbol{I}_4$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$, $\boldsymbol{B}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$ and $\boldsymbol{b}_i = (b_{i1}, b_{i2})^{\mathrm{T}} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{D})$. The specific model parameters are:

$$\boldsymbol{\beta} = (1, 2, -2, 4)^{\mathrm{T}}, \quad \boldsymbol{D} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{and } \boldsymbol{C}_i = \boldsymbol{I}_7,$$

where the values of ρ are chosen as 0, 0.5 and 0.9 to reflect zero, middle and high correlations between outcome variables, respectively. The sample sizes N are set to 25 and 50, and a total of 100 replications are run for each combination of between-outcome correlation ρ and sample size N. Each simulated dataset is fitted by the MNLMM using the five computational procedures, say the PNLS-MLME, Laplacian, pseudo-ECM, MCEM and ISEM algorithms, described in Section 2. Initial values for the parameters are chosen to be the true values of parameters plus a random draw from the standard normal distribution. Note that the E step of the MCEM algorithm is undertaken with generating $M = 1000$ MC samples. When implementing the ISEM algorithm, the envelop distribution was multivariate normal mixtures with three different mixing proportions $P_0 = 0.1,\ 0.5$ and 0.9. Because all converged estimates are almost the same, we report only the result under $P_0 = 0.5$ for the sake of conciseness. The computational procedures achieve convergence when:

$$\max_{l=1,\ldots,m} \left(\left| (\hat{\theta}_l^{(h+1)} - \hat{\theta}_l^{(h)}) / \hat{\theta}_l^{(h)} \right| \right) < 0.01,$$

where m is the number of unknown parameters.

Table 3 summarizes the averages of CPU time (Time), numbers of iterations (Iter), converged log-likelihood values ($\ell_{\max}$), relative bias (RB) of log-likelihood functions and empirical sums of relative mean squared errors (RMSE) of parameter estimates obtained by five approximation methods over 100 replicates under all considered scenarios. The relative bias of log-likelihood values calculated as $(\ell_{\max} - \ell_{\text{true}})/|\ell_{\text{true}}|$ is used to evaluate the accuracy of the estimation of the log-likelihood function, where ℓ_{true} is the true value of the log-likelihood function and $\ell_{\max}$ is the converged maximized log-likelihood value. The empirical sums of RMSE for each case are calculated as $\sum_{l=1}^{m} (\hat{\theta}_l - \theta_l)^2 / \theta_l^2$, where θ_l and $\hat{\theta}_l$ are each the entry of the true value of the parameter and its estimate, respectively.

Based on the results shown in Table 3, we first compare the convergence speed of the five estimation procedures. Apparently, the pseudo-ECM method takes the least consumed CPU time, and it is followed by the PNLS-MLME, Laplacian, ISEM and then the MCEM methods. The fewest number of iterations is required by running the PNLS-MLME method followed by the pseudo-ECM, Laplacian, ISEM and MCEM methods, while the last four methods show negligible differences, especially for a large sample size and a high between-outcome correlation. Not surprisingly, the MCEM and ISEM methods require heavier computational cost, because they need to generate a great number of random samples of random effects to perform the MC integration in each iteration. We also find that the consumed CPU time and the required number of iterations decrease when the between-outcome correlation ρ increases. We remark that the PNLS-MLME method converges quickly, but it fails to converge unless the initial values are good enough. When the chosen starting point is far from optimum, it may cause divergence of the procedure, and thereby, another set of initial values should be reset.

Table 3. Simulation results for the computational performance of five approximation methods under each combination of correlations ρ and sample sizes N. Iter, iteration; RB, relative bias.

N	ρ		PNLS-MLME	Laplacian	Pseudo-ECM	MCEM	ISEM
25	0	Time	4.077	25.954	1.970	8789.093	5862.499
		Iter	2.150	12.140	9.800	138.440	58.390
		$\ell_{\max}$	−576.769	−610.274	−577.121	−556.914	−642.139
		RB	0.008	−0.033	0.008	0.045	−0.100
		RMSE	2.229	2.441	2.169	2.176	2.177
	0.5	Time	4.370	30.803	2.045	2403.145	1680.319
		Iter	2.120	11.430	9.930	35.650	15.750
		$\ell_{\max}$	−559.366	−582.622	−559.907	−536.608	−625.736
		RB	0.009	−0.022	0.008	0.052	−0.103
		RMSE	0.580	0.672	0.561	0.601	0.602
	0.9	Time	3.646	25.006	1.749	1252.625	1158.028
		Iter	2.000	8.940	8.570	18.330	10.760
		$\ell_{\max}$	−468.270	−474.786	−468.909	−423.555	−535.591
		RB	0.011	−0.003	0.009	0.118	−0.120
		RMSE	0.470	0.484	0.450	0.486	0.477
50	0	Time	8.365	41.545	8.927	6825.341	3967.824
		Iter	2.240	10.050	9.260	56.240	20.170
		$\ell_{\max}$	−1159.337	−1177.863	−1159.675	−1120.721	−1292.848
		RB	0.004	−0.010	0.004	0.039	−0.094
		RMSE	1.688	1.747	1.685	1.692	1.689
	0.5	Time	9.776	56.560	10.210	2112.857	1706.392
		Iter	2.140	9.760	9.530	11.800	9.690
		$\ell_{\max}$	−1124.354	−1140.195	−1124.911	−1079.401	−1258.644
		RB	0.004	−0.009	0.004	0.046	−0.098
		RMSE	0.277	0.324	0.270	0.313	0.315
	0.9	Time	8.185	34.382	6.666	1512.85	1091.661
		Iter	2.000	6.070	6.210	7.320	6.850
		$\ell_{\max}$	−933.662	−943.973	−934.566	−843.025	−1069.55
		RB	0.005	−0.006	0.004	0.113	−0.116
		RMSE	0.226	0.229	0.226	0.237	0.234

When assessing the approximated log-likelihood functions, we find that all approximation methods produce relative biases in log-likelihoods within ± 0.12 (the range is not quite large). Because the simulated datasets are generated from a linear scenario, *i.e.*, bivariate LMM specified in Equation (33), the pseudo-data model given in Equation (8) certainly satisfies the MLMM [1] framework. Therefore, the ML estimates of model parameters, as well as the maximized log-likelihood value obtained by the pseudo-ECM algorithm are exactly the same as those obtained by fitting the MLMM using the EM-based

algorithm. Besides, the PNLS-MLME method uses the same approximation of the log-likelihood function, say $\ell_{\mathrm{PD}}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, with that of pseudo-ECM. Thus, the values of relative biases in log-likelihoods obtained by the PNLS-MLME and pseudo-ECM algorithms are quite similar, and they are very close to zero. Additionally, the Laplacian approximation gives near-zero, but slightly under-estimated relative biases in log-likelihoods, and the relative biases are negligible when the sample size and between-outcome correlation are large. The log-likelihood values could be slightly over-estimated by using the MCEM method and slightly under-estimated by using the ISEM method. As anticipated, the approximations of log-likelihood functions will get close to the exact log-likelihood value when the sample size increases.

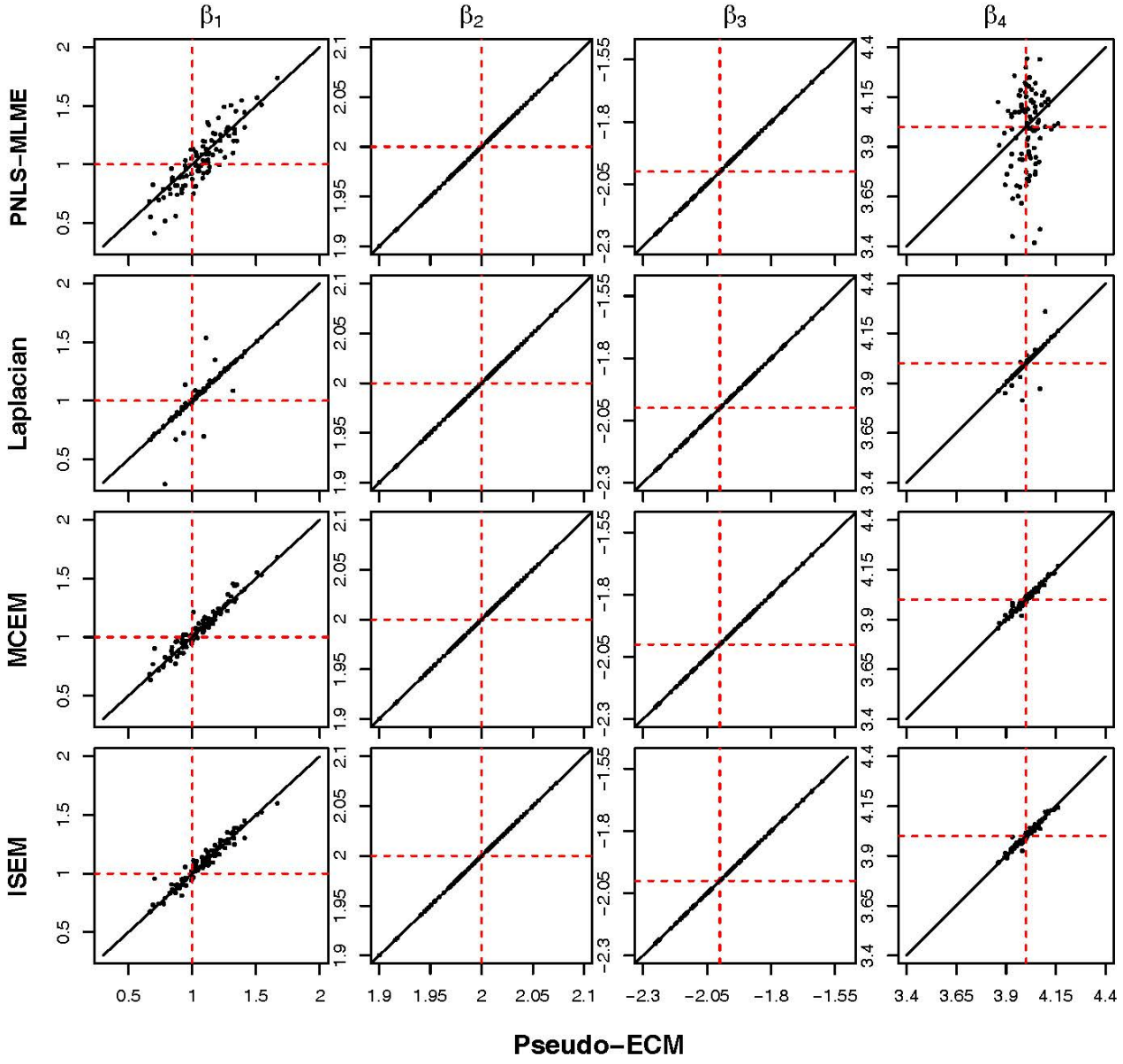

Figure 3. Scatter plots of fixed-effects estimates for PNLS-MLME, Laplacian, MCEM and ISEM against pseudo-ECM methods for the multivariate nonlinear mixed-effects model (MNLMM) under the case of $N = 25$, $\rho = 0.9$.

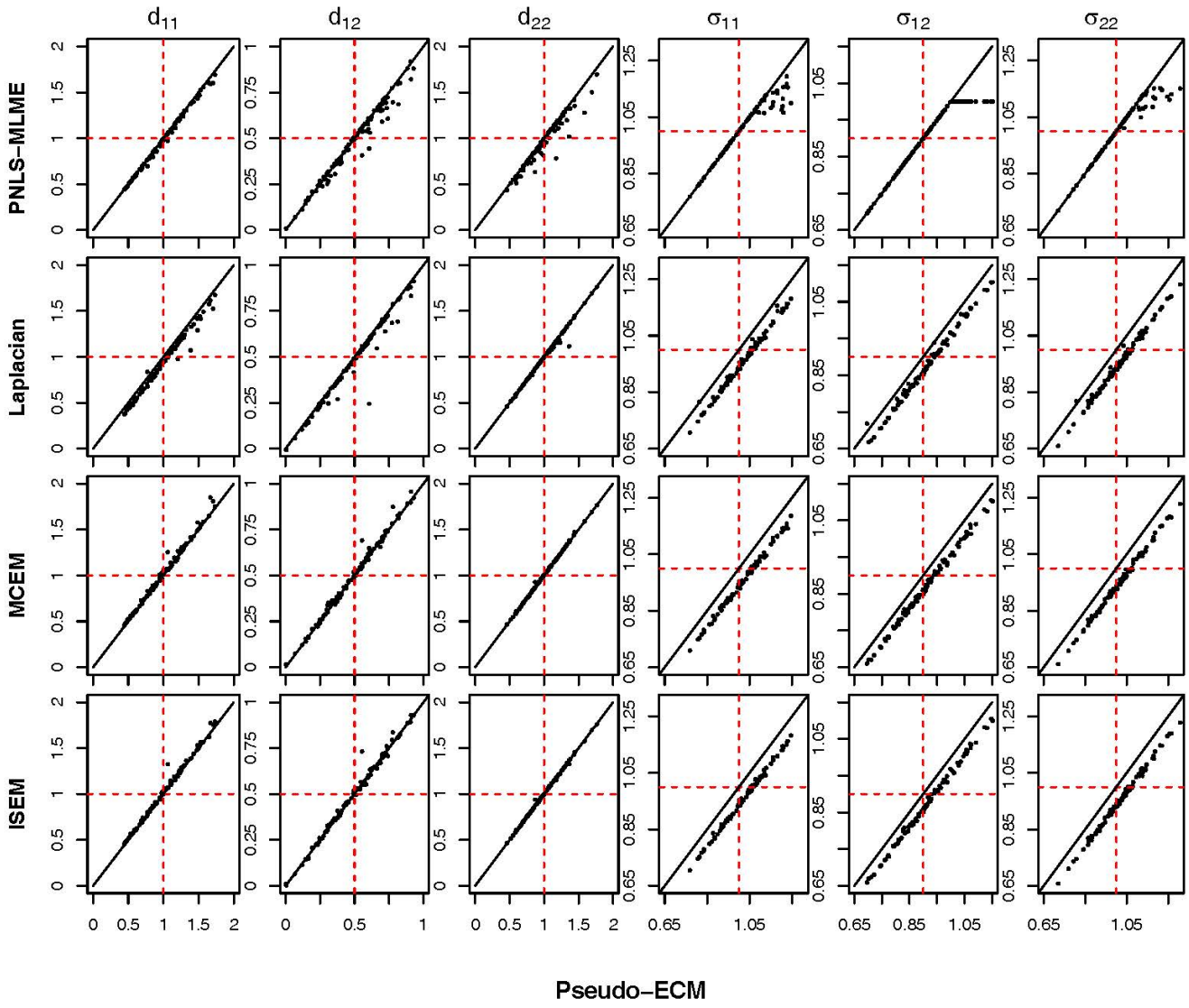

Figure 4. Scatter plots of variance-covariance components estimates for PNLS-MLME, Laplacian, MCEM and ISEM against pseudo-ECM methods for the MNLMM under the case of $N = 25$, $\rho = 0.9$.

We now turn our attention to observing the estimation performance for model parameters under the five computational methods. From the RMSE rows of Table 3, typically, the five methods give comparable results for estimation accuracy due to negligible differences in RMSE scores. The RMSE decreases as the sample size increases, confirming the good asymptotic properties of ML estimators, at least for the setting of parameters used in this simulation. As mentioned above, the pseudo-ECM method implemented for linear models produces the same results as the EM-type algorithm for MLMM. Judging from Table 3, the pseudo-ECM method has the smallest RMSE among the five computational methods. Furthermore, we compare the estimates of each parameter obtained by PNLS-MLME, Laplacian, MCEM and ISEM against those obtained by pseudo-ECM one-by-one in detail. Figures 3 and 4 display the scatter plots of the estimates of fixed effects ($\boldsymbol{\beta}$) and variance-covariance components ($\boldsymbol{D}$ and $\boldsymbol{\Sigma}$) separately for the pseudo-ECM method (in the X-axes) *versus* the other four procedures (in the Y-axes). The dashed lines indicate the true values of parameters. To save space, we present only the case of

$N = 25$ and $\rho = 0.9$, because the other five cases exhibit almost a similar pattern. It can be seen from the two figures that the estimates are all located in the neighborhood of the true values, indicating that all five computational procedures yield very precise estimates of model parameters. In general, there is a strong agreement in the estimates obtained through the five methods, because the point estimates fall close to the 45-degree line. However, for the estimate of β_4, PNLS-MLME appears to have a slightly large variability. For the estimates of σ_{11}, σ_{12} and σ_{22}, the other four methods tend to give estimates smaller than does the pseudo-ECM algorithm.

4.2. Bivariate Nonlinear Case

In the simulation, the data were generated from the MNLMM with nonlinear mean curves Equation (31). The presumed model parameters are:

$$\boldsymbol{\beta} = (12, -2.7, 1.3, 16.9, -1.7, 1.3)^{\mathrm{T}}, \quad \boldsymbol{D} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 4 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & -0.2 \\ -0.2 & 5 \end{bmatrix}, \quad \boldsymbol{C}_i = \boldsymbol{I}_{10}.$$

Each simulated dataset is fitted by the MNLMM using the five approximation methods described in Section 2. To investigate the effect of the size of MC samples for MCEM and mixing proportions of the envelope distribution for ISEM, we consider MC sample sizes $M = 500,\ 1000,\ 2000$ and the mixing proportions $P_0 = 0.1,\ 0.5,\ 0.9$. A total of 100 replications are run for each of sample sizes $N = 25$ and 50 across nine computational procedures. The convergence rule is the same as the previous simulation. Note that numerical double-integration is performed to calculate the exact log-likelihood, such that the evaluation of the accuracy of the approximate log-likelihood is tractable.

In this simulation study, there are 18 (10) and 12 (7) non-convergence cases out of 100 trials for the PNLS-MLME and Laplacian methods, respectively, under sample size $N = 25\ (50)$. To ensure that we are comparing estimates of different methods based on the same simulated data and initial values, an additional dataset will be regenerated in the procedure if one of the methods did not converge for a particular dataset. This can be done by using the R `try()` function to handle the error-recovery. Table 4 reports the computing results, including the averages of CPU time (Time), numbers of iterations (Iter), converged log-likelihood values ($\ell_{\max}$), RB of log-likelihood functions and empirical sums of the RMSE of parameter estimates for each sample size and each algorithm. The results indicate that the pseudo-ECM spent the least CPU time, followed by the PNLS-MLME, Laplacian, ISEM with $P_0 = 0.1,\ 0.5$, MCEM with $M = 500,\ 1000,\ 2000$ and, then, ISEM with $P_0 = 0.9$. The PNLS-MLME demands the fewest numbers of iterations, followed by Laplacian, pseudo-ECM, ISEM with $P_0 = 0.1,\ 0.5$, MCEM with $M = 2000, 1000, 500$ and, then, ISEM with $P_0 = 0.9$. The performance of the five methods under the bivariate nonlinear model is conceptually similar to that under the bivariate linear model shown in Section 4.1. It makes sense that the consumed CPU time increases with the size of MC samples M for MCEM, but the required iteration number decreases with MC sample size M. Besides, for the ISEM method, when the proportion of importance samples of random effects drawing from the posterior of $\boldsymbol{b}_i$ increases (say P_0 decreases), both the CPU time and iteration number decrease.

It can be seen from the RB column of Table 4 that all methods except for the three ISEM procedures provide comparable accuracy for approximate observed log-likelihood values, while the ISEM method tends to get a relatively large bias. Observing the empirical sums of RMSE, the PNLS-MLME and

pseudo-ECM methods can yield more accurate estimates of model parameters as $N = 25$ and $N = 50$, respectively, while the others show minor difference in RMSE scores. The MCEM method generally offers better precision of the parameter estimates when the size of generated MC samples increases. Although the MCEM spent much CPU time and had larger iteration numbers to achieve convergence, it can produce relatively small bias for the approximation of observed log-likelihood and smaller RMSE for estimates of model parameters, especially for large sizes of sample $N = 50$ and MC samples $M = 2000$. Additionally, among the three settings of P_0 for ISEM, the case of equal weights (say $P_0 = 0.5$) gives smaller RB and RMSE scores. If we want to obtain more accurate results of approximate log-likelihood using the ISEM algorithm, probably a larger number of samples of random effects might be necessary, but it seems inefficient. As expected, when the sample size N increases, the required CPU time and iteration number increase, and the RB and RMSE decrease, confirming the large sample properties of ML estimation. In addition, the RMSE ($\times 10^2$) for the estimates of each parameter under the nine considered estimating procedures are listed in Table 5. It seems that the estimators for β_5, β_6, d_{11}, d_{21}, d_{22} and σ_{21} show somewhat less precise point estimates as opposed to the other parameters in the setting of this simulation. Observing the table, there are remarkable differences in the magnitude of RMSE values as the precision of parameter estimates depends heavily on the specification of nonlinear mean functions. Moreover, there are no consistent rankings of precision among the nine considered procedures for each parameter. Although this is a limited study, it demonstrates that all five approximation methods can give reasonable results for parameter estimation.

Table 4. Simulation results for nine estimating procedures under the bivariate nonlinear case.

Sample Size N	Methods	Comparison Criteria				
		Time	Iter	$\ell_{\max}$	RB	RMSE
25	PNLS-MLME	5.071	3.533	−847.968	0.009	1.671
	Laplacian	21.199	7.133	−860.383	−0.012	2.000
	Pseudo-ECM	2.709	12.000	−847.994	0.009	1.967
	MCEM ($M = 500$)	9062.743	380.000	−847.217	0.010	2.099
	MCEM ($M = 1000$)	9569.619	213.733	−847.346	0.010	2.072
	MCEM ($M = 2000$)	11,375.297	131.400	−847.896	0.009	2.029
	ISEM ($P_0 = 0.9$)	17,008.449	333.733	−887.996	−0.028	1.999
	ISEM ($P_0 = 0.5$)	4635.601	93.400	−881.169	−0.018	1.882
	ISEM ($P_0 = 0.1$)	1086.651	22.200	−862.842	−0.020	2.077
50	PNLS-MLME	14.149	3.940	−1710.123	0.007	1.119
	Laplacian	53.066	7.690	−1763.046	−0.010	1.134
	Pseudo-ECM	11.331	13.070	−1710.216	0.007	1.110
	MCEM ($M = 500$)	15,860.866	392.595	−1713.939	0.005	1.184
	MCEM ($M = 1000$)	24,077.335	238.470	−1714.151	0.005	1.157
	MCEM ($M = 2000$)	26,328.930	134.750	−1714.447	0.004	1.151
	ISEM ($P_0 = 0.9$)	31,224.663	386.120	−1789.168	−0.021	1.255
	ISEM ($P_0 = 0.5$)	7065.363	106.350	−1780.396	−0.015	1.138
	ISEM ($P_0 = 0.1$)	2805.677	26.870	−1779.298	−0.018	1.153

Table 5. Relative mean squared errors ($\times 10^2$) for the estimates of model parameters under nine iterative procedures.

Sample Size N	Methods	Parameter											
		β_1	β_2	β_3	β_4	β_5	β_6	d_{11}	d_{21}	d_{22}	σ_{11}	σ_{21}	σ_{22}
25	PNLS-MLME	0.046	0.960	0.046	0.041	18.505	11.559	21.698	87.774	4.565	0.998	20.003	0.909
	Laplacian	0.045	0.965	0.046	0.033	18.600	11.766	20.875	120.340	4.609	1.412	20.013	1.325
	Pseudo-ECM	0.043	0.964	0.046	0.026	18.501	11.570	20.066	118.786	4.759	0.988	20.010	0.909
	MCEM ($M = 500$)	0.046	0.956	0.045	0.039	18.736	11.781	20.926	130.477	4.549	1.389	19.682	1.299
	MCEM ($M = 1000$)	0.047	0.969	0.045	0.038	18.589	11.668	21.160	127.668	4.797	1.383	19.559	1.314
	MCEM ($M = 2000$)	0.047	0.970	0.046	0.036	18.602	11.669	20.402	123.817	4.606	1.404	19.935	1.315
	ISEM ($P_0 = 0.9$)	0.046	0.960	0.046	0.028	18.590	11.666	20.510	120.740	4.609	1.400	20.013	1.315
	ISEM ($P_0 = 0.5$)	0.047	0.969	0.046	0.040	18.420	11.476	19.930	110.919	4.000	1.463	19.619	1.271
	ISEM ($P_0 = 0.1$)	0.043	0.993	0.045	0.021	18.785	11.899	26.451	122.587	4.407	1.631	19.474	1.377
50	PNLS-MLME	0.053	0.433	0.019	0.083	8.038	6.437	9.609	62.998	3.126	0.355	20.445	0.290
	Laplacian	0.054	0.433	0.019	0.040	8.056	6.501	10.172	63.165	3.121	0.787	20.025	1.051
	Pseudo-ECM	0.053	0.432	0.019	0.043	8.055	6.452	8.858	62.921	3.013	0.355	20.493	0.289
	MCEM ($M = 500$)	0.052	0.420	0.019	0.087	8.117	6.505	10.334	67.836	3.055	0.875	20.054	1.033
	MCEM ($M = 1000$)	0.054	0.420	0.019	0.085	8.149	6.508	10.099	65.263	3.113	0.881	20.003	1.063
	MCEM ($M = 2000$)	0.054	0.418	0.019	0.075	8.120	6.494	10.185	64.350	3.120	0.892	20.274	1.070
	ISEM ($P_0 = 0.9$)	0.059	0.415	0.019	0.080	8.034	6.429	18.313	67.054	3.142	0.924	20.045	1.011
	ISEM ($P_0 = 0.5$)	0.055	0.429	0.019	0.053	8.131	6.508	9.040	64.614	3.143	0.861	19.819	1.080
	ISEM ($P_0 = 0.1$)	0.052	0.431	0.019	0.035	8.194	6.554	10.182	64.165	3.011	0.987	20.542	1.147

5. Discussion and Conclusions

In this article, we describe and compare five approximation methods to carry out ML estimation of the MNLMM, as well as the evaluation of the observed log-likelihood function. The methods, namely PNLS-MLME, Laplacian approximation, pseudo-ECM, MCEM and ISEM algorithms, depend on the result of the first two order Taylor expansions. The PNLS-MLME and pseudo-ECM methods use a linearization of nonlinear mean functions, while the other three methods rely on an approximation of the observed likelihood. Numerical results indicate that the five methods can give comparable accuracy of the estimation of model parameters, as well as approximation of the observed log-likelihood function of the MNLMM.

In summary, the five algorithmic schemes preserve flexibility and simplicity in carrying out ML estimation of the MNLMM. The pseudo-ECM method can offer relatively better efficiency compared to the other four methods. For the PNLS-MLME and Laplacian methods, a poor initial guess of $\boldsymbol{\theta}$ can result in poor estimates of $\{\boldsymbol{b}_i\}_{i=1}^N$, and thereby, the accuracy of parameter estimates and the performance of convergence become worse. To overcome this weakness, the consideration of different starting values for $\hat{\boldsymbol{D}}^{(0)}$ is recommended by specifying $c\hat{\boldsymbol{D}}^{(0)}$, where c is a random draw from the standard normal distribution and the original $\hat{\boldsymbol{D}}^{(0)}$ is given in Section 2.7. The MCEM and ISEM methods appear to be less efficient, because both of them spend much time to generate MC samples for evaluating the required conditional expectations in each iteration. For the implementation of the ISEM algorithm, the specification of the mixing proportion P_0 depends on the data at hand. We suggest trying a variety of settings and choose the optimal P_0 corresponding to the maximized approximate observed log-likelihood. An R package for fitting MNLMM based on the proposed techniques will be released in the near future.

However, the multivariate normality assumption in the MNLMM might not provide robust inference if the data, even after being transformed, and exhibit fat tails and/or skewness [48–50]. To alleviate such limitations, it is natural to replace the multivariate normally-distributed random effects and within-subject errors of the MNLMM by a broader family, such as the multivariate skew-normal distribution [51], the multivariate skew-t distribution [52], the multivariate skew-elliptical distribution [53], or the multivariate skew-normal independent distribution [54,55]. The proposed methods are readily extendable to carry out ML estimation of the multivariate version of skew-family nonlinear mixed models. This leads to valuable further research on the issue of developing multivariate skew-family nonlinear mixed models together with their ML inference.

Acknowledgments

The author would like to express her deepest gratitude to the Chief Editor, the Associate Editor and two anonymous reviewers for their insightful comments and suggestions that greatly improved this article. This work was partially supported by the Ministry of Science and Technology under Grant No. MOST 103-2118-M-035-001-MY2 of Taiwan.

Conflicts of Interest

The author declares no conflict of interest.

Appendix

A. Score Vector and Hessian Matrix

The score vector $\mathbf{s}_{\boldsymbol{\alpha}}$ calculated as the first derivatives of $\ell_{\mathrm{PD}}(\boldsymbol{\theta}|\boldsymbol{y})$ in Equation (9) with respect to each entry of $\boldsymbol{\alpha}$ can be expressed by:

$$[\mathbf{s}_{\boldsymbol{\alpha}}]_l = \frac{1}{2}\sum_{i=1}^{N}\Big\{\big(\tilde{\boldsymbol{y}}_i^{(h)} - \tilde{\boldsymbol{X}}_i^{(h)}\boldsymbol{\beta}\big)^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}_i^{(h)^{-1}}\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)}\tilde{\boldsymbol{\Lambda}}_i^{(h)^{-1}}\big(\tilde{\boldsymbol{y}}_i^{(h)} - \tilde{\boldsymbol{X}}_i^{(h)}\boldsymbol{\beta}\big) - \mathrm{tr}\big(\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)}\big)\Big\},$$

for $l = 1, \ldots, g$, $g = q(q+1)/2 + r(r+1)/2 + \dim(\boldsymbol{\phi})$, where $\tilde{\boldsymbol{\Lambda}}_i^{(h)} = \tilde{\boldsymbol{Z}}_i^{(h)}\boldsymbol{D}\tilde{\boldsymbol{Z}}_i^{(h)^{\mathrm{T}}} + \boldsymbol{\Sigma}\otimes\boldsymbol{C}_i(\boldsymbol{\phi})$,

$$\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)} = \frac{\partial\tilde{\boldsymbol{\Lambda}}_i^{(h)}}{\partial w_l} = \begin{cases} \tilde{\boldsymbol{Z}}_i^{(h)}\frac{\partial\boldsymbol{D}}{\partial w_l}\tilde{\boldsymbol{Z}}_i^{(h)^{\mathrm{T}}} & \text{if } w_l = \mathrm{vech}(\boldsymbol{D}), \\ \frac{\partial\boldsymbol{\Sigma}}{\partial w_l}\otimes\boldsymbol{C}_i(\boldsymbol{\phi}) & \text{if } w_l = \mathrm{vech}(\boldsymbol{\Sigma}), \\ \boldsymbol{\Sigma}\otimes\frac{\partial\boldsymbol{C}_i(\boldsymbol{\phi})}{\partial w_l} & \text{if } w_l = \boldsymbol{\phi}. \end{cases} \tag{A.1}$$

Here, $\partial\boldsymbol{D}/\partial\omega_l$ is one in the (j,l)-th and the (l,j)-th elements of $\boldsymbol{D}$ as $\omega_l = d_{jl}$, say the distinct element of $\boldsymbol{D}$, and zero otherwise; similarly for $\partial\boldsymbol{\Sigma}/\partial\omega_l$ when $\omega_l = \sigma_{jl}$. Besides, the Hessian matrix calculated as the second derivatives of $\ell_{\mathrm{PD}}(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to each entry of $\boldsymbol{\alpha}$ is:

$$\begin{aligned}\Big[\mathbf{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\Big]_{lu} = \frac{1}{2}\sum_{i=1}^{N}\Big\{&\mathrm{tr}\Big[\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}\Big(\dot{\tilde{\boldsymbol{\Lambda}}}_{iu}^{(h)}\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)} - \ddot{\tilde{\boldsymbol{\Lambda}}}_{ilu}^{(h)}\Big)\Big] + \mathrm{tr}\Big[\Big(\tilde{\boldsymbol{y}}_i^{(h)} - \tilde{\boldsymbol{X}}_i^{(h)}\boldsymbol{\beta}\Big) \\ &\times\Big(\tilde{\boldsymbol{y}}_i^{(h)} - \tilde{\boldsymbol{X}}_i^{(h)}\boldsymbol{\beta}\Big)^{\mathrm{T}}\Big(\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}(\ddot{\tilde{\boldsymbol{\Lambda}}}_{ilu}^{(h)} - 2\dot{\tilde{\boldsymbol{\Lambda}}}_{iu}^{(h)}\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}\dot{\tilde{\boldsymbol{\Lambda}}}_{il}^{(h)})\tilde{\boldsymbol{\Lambda}}_i^{-1^{(h)}}\Big)\Big]\Big\},\end{aligned}$$

where:

$$\ddot{\tilde{\boldsymbol{\Lambda}}}_{ilu}^{(h)} = \frac{\partial \dot{\tilde{\boldsymbol{\Lambda}}}_i^{(h)}}{\partial w_l} = \begin{cases} \frac{\partial \boldsymbol{\Sigma}}{\partial w_l} \otimes \frac{\partial \boldsymbol{C}_i(\boldsymbol{\phi})}{\partial w_u} & \text{if } w_l = \text{vech}(\boldsymbol{\Sigma}), w_u = \boldsymbol{\phi}, \\ 0 & \text{otherwise.} \end{cases}$$

References

1. Shah, A.; Laird, N.; Schoenfeld, D. A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. *J. Am. Stat. Assoc.* **1997**, *92*, 775–779.
2. Marshall, G.; de la Cruz-Mesía, R.; Barón, A.E.; Rutledge, J.H.; Zerbe, G.O. Non-linear Random Effects Model for Multivariate Responses with Missing Data. *Statist. Med.* **2006**, *25*, 2817–2830.
3. Sammel, M.; Lin, X.; Ryan, L. Multivariate Linear Mixed Models for Multiple Outcomes. *Statist. Med.* **1999**, *18*, 2479–2492.
4. Song, X.; Davidian, M.; Tsiatis, A.A. An Estimator for the Proportional Hazards Model with Multiple Longitudinal Covariates Measured with Error. *Biostatistics* **2002**, *3*, 511–528.
5. Roy, J.; Lin, X. Analysis of Multivariate Longitudinal Outcomes with Nonignorable Dropouts and Missing Covariates: Changes in Methadone Treatment Practices. *J. Am. Stat. Assoc.* **2002**, *97*, 40–52.
6. Roy, A. Estimating Correlation Coefficient between Two Variables with Repeated Observations Using Mixed Effects Model. *Biom. J.* **2006**, *48*, 286–301.
7. Wang, W.L.; Fan, T.H. ECM-Based Maximum Likelihood Inference for Multivariate Linear Mixed Models with Autoregressive Errors. *Comput. Stat. Data Anal.* **2010**, *54*, 1328–1341.
8. Lindstrom, M.J.; Bates, D.M. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* **1990**, *46*, 673–687.
9. Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurements Data*; Chapman & Hall: London, UK, 1995.
10. Pinheiro, J.C.; Bates, D.M. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *J. Comput. Graph. Stat.* **1995**, *4*, 12–35.
11. Pinheiro, J.C.; Bates, D.M. *Mixed-Effects Models in S and S-PLUS*; Springer: Berlin, Germany 2000.
12. Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D.; R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*; R package version 3.1-104; Available online: http://CRAN.R-project.org/package=nlme (accessed on 24 July 2015).
13. Dey, D.K.; Chen, M.H.; Chang, H. Bayesian Approach for Nonlinear Random Effects Models. *Biometrics* **1997**, *53*, 1239–1252.
14. Huang, Y.; Liu, D.; Wu, H. Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System. *Biometrics* **2006**, *62*, 413–423.
15. Lachosa, V.H.; Castrob, L.M.; Dey, D.K. Bayesian Inference in Nonlinear Mixed-Effects Models Using Normal Independent Distributions. *Comput. Stat. Data Anal.* **2013**, *64*, 237–252.
16. Wolfinger, R.D.; Lin, X. Two Taylor-Series Approximation Methods for Nonlinear Mixed Models. *Comput. Stat. Data Anal.* **1997**, *25*, 465–490.

17. Ge, Z.; Bickel, J.P.; Rice A.J. An Approximate Likelihood Approach to Nonlinear Mixed Effects Models via Spline Approximation. *Comput. Stat. Data Anal.* **2004**, *46*, 747–776.
18. Walker, S.G. An EM Algorithm for Nonlinear Random Effects Models. *Biometrics* **1996**, *52*, 934–944.
19. Wang, J. EM Algorithms for Nonlinear Mixed Effects Models. *Comput. Stat. Data Anal.* **2007**, *51*, 3244–3256.
20. Vonesh, E.F.; Wang, H.; Nie, L.; Majumdar, D. Conditional Second-order Generalized Estimating Equations for Generalized Linear and Nonlinear Mixed-Effects Models. *J. Am. Stat. Assoc.* **2002**, *97*, 271–283.
21. Vonesh, E.F. Non-linear Models for the Analysis of Longitudinal Data. *Stat. Med.* **1992**, *11*, 1929–1954.
22. Beal, S.; Sheiner, L. The NONMEM System. *Am. Stat.* **1980**, *34*, 118–199.
23. Wolfinger, R.D. Comment: Experiences with the SAS Macro NLINMIX. *Stat. Med.* **1997**, *16*, 1258–1259.
24. Wolfinger, R.D. Fitting Nonlinear Mixed Models with the New NLMIXED Procedure. In Proceedings of the 99 Joint Statistical Meetings, Miami Beach, FL, USA, 11–14 April 1999.
25. Kuhn, E.; Lavielle, M. Maximum Likelihood Estimation in Nonlinear Mixed Effects Models. *Comput. Stat. Data Anal.* **2005**, *49*, 1020–1038.
26. Lavielle, M. *MONOLIX (MOdelès NOn LInéaires à effets miXtes)*; MONOLIX Group: Orsay, France, 2008.
27. Beal, S.; Sheiner, L.; Boeckmann, A.; Bauer, R. *NONMEM User's Guides (1989–2009)*; Icon Development Solutions: Ellicott City, MD, USA, 2009.
28. Comets, E.; Lavenu, A.; Lavielle, M. Saemix: Stochastic Approximation Expectation Maximization (SAEM) Algorithm. R package version 1, 2011.
29. Wang, W.L.; Fan, T.H. Estimation in Multivariate t Linear Mixed Models for Multiple Longitudinal Data. *Statist. Sinica* **2011**, *21*, 1857–1880.
30. Wang, W.L.; Fan, T.H. Bayesian Analysis of Multivariate t Linear Mixed Models Using a Combination of IBF and Gibbs Samplers. *J. Multivar. Anal.* **2012**, *105*, 300–310.
31. Wang, W.L. Multivariate t Linear Mixed Models for Irregularly Observed Multiple Repeated Measures with Missing Outcomes. *Biom. J.* **2013**, *55*, 554–571.
32. Tierney, L.; Kadane, J.B. Accurate Approximations for Posterior Moments and Densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86.
33. Meng, X.L.; Rubin, D.B. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* **1993**, *80*, 267–278.
34. Booth, G.J.; Hobert, P.J. Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *J. R. Stat. Soc. Ser. B* **1999**, *61*, 265–285.
35. Lai, T.L.; Shih, M.C. A Hybrid Estimator in Nonlinear and Generalized Linear Mixed Effects Models. *Biometrika* **2006**, *90*, 791–795.
36. Leonard, T.; Hsu, J.S.J.; Tsui, K.W. Bayesian Marginal Inference. *J. Am. Stat. Assoc.* **1989**, *84*, 1051–1058.

37. Bates, D.M.; Watts, D.G. Relative Curvature Measures of Nonlinearity. *J. R. Stat. Soc. Ser. B* **1980**, *42*, 1–25.
38. R Development Core Team. *R. A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
39. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with Discussion). *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
40. Wei, G.C.G.; Tanner, M.A. A Monte Carlo Implementation of the EM Algorithm and the Poor's Man's Data Augmentation Algorithms. *J. Am. Stat. Assoc.* **1990**, *85*, 699–704.
41. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.
42. Lederman, M.M.; Connick, E.; Landay, A.; Kuritzkes, D.R.; Spritzler, J.; Clair, M.S.; Kotzin, B.L.; Fox, L.; Chiozzi, M.H.; Leonard, J.M.; *et al*. Immunologic Responses Associated with 12 Weeks of Combination Antiretroviral Therapy Consisting of Zidovudine, Lamivudine, and Ritonavir: Results of AIDS Clinical Trials Group Protocol 315. *J. Infect. Dis.* **1998**, *178*, 70–79.
43. Connick, E.; Lederman, M.M.; Kotzin, B.L.; Spritzler, J.; Kuritzkes, D.R.; Clair, M.S. Sevin, A.D.; Fox, L.; Chiozzi, M.H.; Leonard, J.M.; *et al*. Immune Reconstitution in the First Year of Potent Antiretroviral Therapy and Its Relationship to Virologic Response. *J. Infect. Dis.* **2000**, *181*, 358–363.
44. 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.
45. Liang, H.; Wu, H.; Carroll, R.J. The Relationship between Virologic Responses in AIDS Clinical Research Using Mixed-Effects Varying-Coefficient Models with Measurement Error. *Biostatistics* **2003**, *4*, 297–312.
46. Wu, H.; Liang, H. Backfitting Random Varying-Coefficient Models with Timedependent Smoothing Covariates. *Scand. J. Stat.* **2004**, *31*, 3–19.
47. Lin, T.I.; Wang, W.L. Multivariate Skew-Normal Linear Mixed Models for Multi-outcome Longitudinal Data. *Stat. Model.* **2013**, *13*, 199–221.
48. Lin, T.I.; Lee, J.C. A Robust Approach to t Linear Mixed Models Applied to Multiple Sclerosis Data. *Statist. Med.* **2006**, *25*, 1397–1412.
49. Lin, T.I.; Lee, J.C. Bayesian Analysis of Hierarchical Linear Mixed Modeling Using Multivariate t Distributions. *J. Statist. Plan. Inf.* **2007**, *137*, 484–495.
50. Lin, T.I.; Lee, J.C. Estimation and Prediction in Linear Mixed Models with Skew Normal Random Effects for Longitudinal Data. *Statist. Med.* **2008**, *27*, 1490–1507.
51. Arellano-Valle, R.B.; Genton, M. On Fundamental Skew Distributions. *J. Multivar. Anal.* **2005**, *96*, 93–116.
52. Azzalini, A.; Capitaino, A. Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t-Distribution. *J. R. Stat. Soc. Ser. B* **2003**, *65*, 367–389.
53. Branco, M.; Dey, D. A General Class of Multivariate Skew-Elliptical Distribution. *J. Multivar. Anal.* **2001**, *79*, 93–113.

54. Bandyopadhyay, D.; Lachos, V.H.; Abanto-Vallec, C.A.; Ghosh, P. Linear Mixed Models for Skew-Normal/Independent Bivariate Responses with an Application to Periodontal Disease. *Statist. Med.* **2010**, *29*, 2643–2655.
55. Bandyopadhyay, D.; Castro, L.M.; Lachos, V.H.; Pinheiro, H.P. Robust Joint Non-linear Mixed-Effects Models and Diagnostics for Censored HIV Viral Loads with CD4 Measurement Error. *J. Agr. Biol. Environ. Stat.* **2015**, *20*, 121–139.