Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 27

5-1-2016

A Comparison of Estimation Methods for Nonlinear Mixed-Effects Models Under Model Misspecification and Data Sparseness: A Simulation Study

Jeffrey R. Harring
University of Maryland, harring@umd.edu

Junhui Liu
Educational Testing Service

A Comparison of Estimation Methods for Nonlinear Mixed-Effects Models Under Model Misspecification and Data Sparseness: A Simulation Study

Cover Page Footnote
This research was partially funded by the Institute of Education Sciences (R305A130042).

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 539-569.

ISSN 1538 − 9472

A Comparison of Estimation Methods For Nonlinear Mixed Effects Models Under Model Misspecification and Data Sparseness: A Simulation Study

Jeffrey R. Harring
University of Maryland
College Park, MD

Junhui Liu
Educational Testing Service
Princeton, NJ

A Monte Carlo simulation is employed to investigate the performance of five estimation methods of nonlinear mixed effects models in terms of parameter recovery and efficiency of both regression coefficients and variance/covariance parameters under varying levels of data sparseness and model misspecification.

Keywords: Random coefficient models, linearization, quadrature, Bayesian, nonlinear models, non-normality

Introduction

A common challenge for substantive researchers across numerous research domains is to make inferences on features underlying profiles of continuous repeated measures data for a sample of individuals from a population of interest. Nonlinear mixed effects (NLME) models (Davidian & Giltinian, 1995; Pinheiro & Bates, 2000; Vonesh & Chinchilli, 1997) have become the tools of choice for analyses in which the primary interest of researchers focuses on understanding the nature of systematic and random variation between and within individuals. The biomedical literature, for example, is replete with studies from areas like pharmacokinetics, which have developed NLME models to examine drug concentration and dispersion in patients (see e.g., Beal & Sheiner, 1985) or modeling markers of disease progression (Morrell, Pearson, Carter, & Bryant, 1995). In the social sciences, Burke, Shrout, and Bolger (2007) used NLME models to examine individual differences in adjustment to spousal loss; while Grimm and Ram (2009) investigated the effects of preschool instruction on

Dr. Harring is Associate Professor of Measurement, Statistics, and Evaluation in the Department of Human Development and Quantitative Methodology. Email him at harring@umd.edu. Dr. Liu is a Psychometrician at ETS.

academic gain using an individual-specific logistic growth model. There are many more examples across diverse research domains.

These applications share several common features. First, mean response for a particular individual is thought to follow a *scientifically-relevant* nonlinear function which characterizes intra-individual behavior in terms of meaningful parameters directly related to the underlying change process. Second, individuals' regression coefficients, in turn, are often formulated to be functions of fixed effects (parameters common to all individuals in the population), covariates (often treatment condition or other individual-level attributes), and individual-specific random effects (parameters representing individual variation). The distribution of random effects captures random variation of the parameters in the population of individuals and is frequently assumed to be multivariate normal.

Although the benefits of incorporating random effects into this framework are undeniable, for a NLME model there is one major drawback. Unlike its linear counterpart (the linear mixed effects model, Laird & Ware, 1982), one liability is that estimation of model parameters is no longer straightforward. The conditional (on the random effects) mean of the response for an individual depends on the random effects in a nonlinear fashion. This nonlinear dependence requires multidimensional integration over the random effects distribution to derive the needed marginal distribution of the data from which inferences can be made. This integral is almost always intractable having no closed form solution.

Several methods were proposed to overcome this problem. Davidian and Giltinan (1993) summarized these methods and classified them into four main categories: (1) methods based on individual estimates, (2) methods based on approximating the likelihood through linearizing the nonlinear function, (3) methods based on the exact likelihood which tackle the multidimensional integration directly, and (4) a Bayesian approach which uses both the likelihood based on the data and prior information about model parameters.

The methodological literature has suggested that these methods may not perform equally well under non-ideal data-analytic situations often encountered in practice, including, but not limited to, violation of distributional assumptions, existence of missing data, and small sample sizes. Although a few modest simulation studies were conducted wherein a small subset of these methods were compared for estimating parameters in NLME models, the primary objective of this study was to do a more comprehensive investigation of a broader set of methods across data analytic conditions found in practice presumed to directly impact the estimation methods themselves.

The Nonlinear Mixed-Effects Model

The basic version of the model is considered, although elaborations are possible (see, e.g., Davidian & Giltinan, 2003; Vonesh & Chinchilli, 1997). Following Davidian and Giltinan (1995), the formulation of the nonlinear mixed-effects model for a typical individual selected from the population can be specified in the general form as,

$$\mathbf{y}_i = \mathbf{f}_i(\mathbf{x}_i,\boldsymbol{\beta}_i) + \mathbf{e}_i, \quad \mathbf{e}_i \mid \boldsymbol{\beta}_i : [\mathbf{0}, \boldsymbol{\Lambda}_i(\boldsymbol{\lambda})] \tag{1}$$

$$\boldsymbol{\beta}_i = \mathbf{g}(\mathbf{z}_i,\boldsymbol{\beta},\mathbf{b}_i), \quad \mathbf{b}_i : [\mathbf{0}, \boldsymbol{\Phi}], \tag{2}$$

where $\mathbf{y}_i' = \left(y_{i1},\ldots,y_{in_i}\right)$ is a $n_i \times 1$ vector of responses, y_{ij}, for the i^{th} individual, $i = 1,\text{K},\text{N}$, at times t_{ij}, $j = 1,\text{K},n_i$. Note that the subscript, n_i, on the response implies that the number of measurements and/or the occasions of measurement could vary by individual. Unbalanced data-gathering designs, planned missingness, or data that are missing at random can all be handled by the NLME model in a straightforward fashion. $\mathbf{f}_i(\mathbf{x}_i,\boldsymbol{\beta}_i)$ is an $n_i \times 1$ vector of nonlinear functions with j^{th} element $f(x_{ij},\boldsymbol{\beta}_i)$, where f is a nonlinear function governing within-individual behavior and is dependent on individual-specific regression parameters $\boldsymbol{\beta}_i$ ($p \times 1$), and x_{ij} contains t_{ij} and other covariates specific to individual i. The $n_i \times 1$ vector of regression residuals, $\mathbf{e}_i$, reflects uncertainty in the response of the i^{th} individual and is assumed to satisfy $E(\mathbf{e}_i \mid \boldsymbol{\beta}_i) = \mathbf{0}$ for all i. Given the individual coefficients, $\mathbf{y}_i$ has covariance structure $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda})$ which is of dimension $n_i \times n_i$ with $q \times 1$ parameters, $\boldsymbol{\lambda}$, common to all subjects. While many different structures for $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda})$ are possible that reflect various data nuances, when coupled with the random effects covariance structure typically takes on a simple structure such as $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda}) = \sigma^2 \mathbf{I}_{n_i}$. This structure will be used in the forthcoming Monte Carlo simulation.

In the model in Equation 1, variation occurring between individuals is captured through individual-specific parameters, $\boldsymbol{\beta}_i$. Dependence of $\boldsymbol{\beta}_i$ on individual-level covariates $\mathbf{z}_i$ is modeled through $\mathbf{g}(\mathbf{z}_i,\boldsymbol{\beta},\mathbf{b}_i)$, a p – dimensional function depending on a $r \times 1$ vector of population parameters $\boldsymbol{\beta}$ and a $k \times 1$ vector of unobservable random effects $\mathbf{b}_i$, associated with individual i. Here, function $\mathbf{g}(\cdot)$ characterizes how elements of $\boldsymbol{\beta}_i$ vary among subjects, due in part to the systematic association with individual attributes, $\mathbf{z}_i$, and unexplained variation in the population captured through $\mathbf{b}_i$. Specifications of $\mathbf{g}(\cdot)$ can be complicated

(see, e.g., Cudeck & Harring, 2007), but at least initially, $\mathbf{g}(\cdot)$ is typically specified as the sum of fixed and random effects such that, $\mathbf{g}(\mathbf{z}_i,\boldsymbol{\beta},\mathbf{b}_i) = \boldsymbol{\beta} + \mathbf{b}_i$. The variability of the random effects is captured through the $k \times k$ symmetric covariance matrix, $\boldsymbol{\Phi}$. The conventional assumption of normality of the random effects is routinely adopted, but as Hartford and Davidian (2000) state, "simply may be inappropriate." Numerous scenarios are possible. It may be, for example, that the distribution of the random effects $\mathbf{b}_i$ is skewed or not unimodal. In the latter case, this situation might arise if an important covariate is left out of the model with the resulting systematic variation that would have been attributed to the covariate relegated to the variation in $\mathbf{b}_i$. Consequently, a bimodal or multimodal distribution may be evident, which would not be well-approximated by a normal distribution. In other settings, the distribution of any of the k random effects (b_{ki}) may be symmetric but may be influenced by more cases in the tails of the distribution than would be expected under normality. This might occur because the sample does not accurately reflect the target population and too many individuals in the sample are on the fringe of the distribution resulting in a heavier-tailed distribution with greater dispersion than would be expected otherwise.

A variant of an exponential function will be used in the Monte Carlo simulation. In the social and behavioral sciences, variants of exponential functions are regularly used to summarize the change processes for many phenomena including the learning of a task (see, e.g., Blozis, 2004; Browne, 1993; Meredith & Tisak, 1990), development of language acquisition (Burchinal & Appelbaum, 1991), and growth characteristics (Browne, 1993). Let the individual-specific function, f, characterize the development on a learning task, for example, be an exponential function of the form

$$f(x_{ij},\boldsymbol{\beta}_i) = \beta_{2i} - (\beta_{2i} - \beta_{1i}) \exp(-\beta_{3i}\, t_{ij}), \tag{3}$$

which at time t_{ij} for individual i, may provide a suitable summary for intra-individual task performance. The parameters of the model correspond to interesting features of the change process. In Equation 3, β_{1i} represents initial performance when $t_{ij} = 0$, β_{2i} denotes the potential performance at later trials (i.e., $f(t_{ij}) \rightarrow \beta_{2i}$ as $t_{ij} \rightarrow \infty$), and β_{3i} governs the rate of change from initial to potential performance.

Estimation Methods

Much methodological work has been done in recent years for fitting NLME models. The need to derive different approaches may be appreciated by inspection of the form of the marginal distribution of $\mathbf{y}_i$ implied by Equations 1 and 2. Denote the conditional density of $\mathbf{y}_i$ given $\mathbf{b}_i$ as $p(\mathbf{y}_i \mid \mathbf{b}_i)$ and the density of $\mathbf{b}_i$ be denoted as $p(\mathbf{b}_i)$, then the marginal distribution of $\mathbf{y}_i$ is given by

$$p(\mathbf{y}_i) = \int p(\mathbf{y}_i \mid \mathbf{b}_i)\, p(\mathbf{b}_i)\, d\mathbf{b}_i. \tag{4}$$

Define the vector of unique elements in $\boldsymbol{\Phi}$ as

$$\begin{aligned} \boldsymbol{\varphi} &= vech(\boldsymbol{\Phi}) \\ &= (\varphi_{11}, \varphi_{21}, \ldots, \varphi_{rr})' \end{aligned}$$

where the $vech(\cdot)$ operator creates a column vector of a symmetric matrix by stacking the diagonal and lower diagonal elements below one another. Putting all relevant model parameters into vector, $\boldsymbol{\theta} : \boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\lambda}', \boldsymbol{\varphi}')$, the maximum likelihood estimates for $\boldsymbol{\theta}$ can be found by maximizing in $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int p(\mathbf{y}_i \mid \mathbf{b}_i)\, p(\mathbf{b}_i)\, d\mathbf{b}_i. \tag{5}$$

Note that, even if both $p(\mathbf{y}_i \mid \mathbf{b}_i)$ and $p(\mathbf{b}_i)$ are $n_i -$ and $k -$ dimensional normal densities, respectively, $p(\mathbf{y}_i)$ need not be normal. Furthermore, except in a few special cases, the integral will be analytically intractable. Finding a closed form solution is thwarted because $\mathbf{b}_i$ enters function f in a nonlinear manner. In short, inference based on the likelihood of the observed data will be complicated by an inability to express the likelihood in closed form. Therefore, it is crucial to find alternate ways to handle the integration.

Estimation approaches can be categorized into four main categories: (a) methods based on individual estimates, (b) methods based on approximating the likelihood through linearizing the nonlinear function, (c) methods based on the exact likelihood which tackle the multi-dimensional integration directly, and (d) a Bayesian approach which uses both the likelihood based on the data and prior information about model parameters. A thorough description of the aforementioned methods, including complete derivations, may be found in

Wolfinger and Lin (1997), Pinheiro and Bates (1995), Demidenko (2004), and Skrondal and Rabe-Hesketh (2004). A synopsis of each of the methods can also be found on the first author's website (http://www.education.umd.edu/EDMS/fac/Harring/webpage.html).

Software Considerations

A self-generated program written in SAS Interactive Matrix Language (IML) was used in the simulation for parameter estimation using the two-stage method based on individuals' estimates with calls to SAS **MIXED** procedure as warranted. Methods based on linearization use algorithms that are numerically simpler then integration methods. They can be found in popular software packages accessible to practitioners. SAS **NLMIXED** procedure was used, based on the First Order (**FIRO**) option (Wolfinger, 1999) for the first-order linearization method. The algorithm of Lindstrom and Bates (1990) conditional first-order method can be obtained by using the **EBLUP** option in the SAS macro **NLINMIX** (Littell et al., 1996). SAS **NLMIXED** was used to implement and execute the Gaussian-Hermite quadrature method using the **NOAD** argument to facilitate the non-adaptive quadrature. Lastly, the **R2WinBUGS** package (Sturtz, Ligges, & Gelman, 2005) in **R** was used to make calls to **WinBUGS** (Spiegelhalter, Thomas, Best, & Lunn, 2002) to facilitate the Bayesian estimation approach. Sample software code for each of these methods can be found in the Appendix.

Review of Previous Simulation Results

Previous simulation studies come from the statistical literature. A non-exhaustive list includes Davidian and Giltinan (1993); Pinheiro and Bates (1995); Roe et al. (1997); Wolfinger and Lin (1997), Hartford and Davidian (2000); Ge, Bickel, and Rice (2004), and Wu (2004).

Davidian and Giltinan (1993) examined the performance of a semiparametric method based on individual estimates and linearization when data had different structures for both inter- and intra-individual variability. They concluded that performance of both methods depended on the relative magnitude of the inter- and intra-individual variability. Misspecification of the intra-individual covariance structure may lead to deterioration in performance for both methods in terms of parameter bias. These methods performed equally well in estimating fixed effects, however, methods based on individual estimates had better estimation of variance and covariance components.

Pinheiro and Bates (1995) examined the performance of the conditional linearization method (Lindstrom & Bates, 1990), Laplace approximation, Gaussian-Hermite and Adaptive Gaussian-Hermite quadrature methods. Their results suggested that the conditional linearization method had the highest computational efficiency but did not provide the most accurate estimation of parameters in terms of bias. Gaussian-Hermite quadrature only provided accurate estimates for large number of quadrature points which made it, in their opinion, computationally inefficient. They concluded that Laplace approximation and Adaptive Gaussian-Hermite quadrature had the best combination of efficiency and accuracy. Pinheiro and Bates' study assumed all assumptions of nonlinear mixed models were met under intensively sampled data. They did not investigate how these methods would perform under distributional misspecification and data sparseness.

Wolfinger and Lin (1997) examined the first-order linearization method and Laplace's approximation method as they are implemented in the SAS macro NLINMIX and concluded that both methods produced reliable estimates, with Laplace's method slightly outperforming the former at the expense of longer computing times and greater instability of the algorithm.

Hartford and Davidian (2000) investigated the consequences for population inference using first-order linearization and Laplace's method when the distribution for the random effects was misspecified – not following a normal distribution. They encountered serious convergence difficulty using Laplace's method when distributions of random effects were far from normal or the population model was not correctly specified. Nevertheless, Laplace's approximation method was still superior to the first-order expansion in parameter accuracy and relative efficiency of estimation except when the random effects distribution was bimodal.

Very little in the NLME model methodological literature has been devoted to how these different estimation methods react to the existence of missing responses or covariates. Wu (2004) suggested that missing values for some of model covariates may have a deleterious effect on parameter recovery. Wu concluded that when the missing data mechanism is nonignorable, serious bias in the parameter estimates may occur.

There has been no simulation work done on the performance of the Bayesian approach. Table 1 provides a summary of the past simulation studies, the estimation methods that were used, the simulation factors that were manipulated, statistical software that was employed if known, and the major findings and limitations.

Research Questions

Specific research questions we address in this simulation study are:

1. Do differences exist between the five estimation methods in terms of parameter bias of fixed effects, variances of the random effects and residual variance? If so, which manipulated study conditions influence the accuracy of parameter recovery?
2. Do differences exist between the five estimation methods in terms of variability of parameter estimates as measured by parameter estimate variance? If so, which manipulated study conditions influence variability of the parameter estimates?

Simulation Design Overview

There are often numerous decision points in analyses involving NLME models. The choice of which method to use often depends on the analytic situation, hypothesis about covariance structures, software availability, sample size, and so on. In order to study the robustness of the five methods of estimating NLME models to the assumptions of normal random effects, conditional normality of the residuals, $\mathbf{e}_i$, data sparseness, and sample size, we carried out a Monte Carlo simulation in which several factors were varied. The data generation model follows Equations 1 and 2, with the exponential model in Equation 3 as the intra-individual function. Although other nonlinear functions could have fewer parameters, we chose this particular function, in part, because it has three coefficients which make the integration feasible, yet is complex enough to examine time to convergence for methods which tackle the integration directly as well as convergence rates for all methods.

Assume that inter-individual function g, is the sum of fixed and random effects

$$\beta_{pi} = \beta_p + b_{pi} \qquad p = 1,2,3$$

This simple model specification was chosen so that, hopefully, model identification and convergence issues would be less likely to confound interpretation of performance. Population values for the regression coefficients are $\beta_1 = 100$, $\beta_2 = 10$, and $\beta_3 = 1$. The covariance matrices describing within- and inter-individual variability in Equations 1 and 2, respectively are given as

Table 1. Summary of past simulation and empirical studies on NLME models

Author(s)	Estimation Method	Study Conditions	Summary of Key Findings
Davidian & Giltinan (2003)	• GTS • Other pooled and un-pooled procedures	Intra-individual variability	• Pooling information about intra-individual variability to obtain correct weighting results in improved efficiency • Pooling had little impact on estimation of parameters in β and Φ
Pinheiro & Bates (1995)	• CFO • Laplace • GHQ • Importance Sampling • AGHQ	• Computational efficiency • Parameter estimate comparison • No simulation study	• CFO provides good approximation and is computationally efficient • GHQ is accurate as number of quadrature points increases resulting in computational inefficiency • AGHQ was as accurate as other methods requiring fewer quadrature points and increased computational efficiency
Wolfinger & Lin (1997)	• FO • Laplace	• Normal random effects distribution and no missing data	• Laplace provided less biased estimates but at greater computational cost and instability in the estimation algorithm
Hartford & Davidian (2000)	• FO • Laplace	• Sampling mechanism • Random effects distribution • Population model misspecification	• Laplace converged to a suitable solution with less frequency when model or random effects distribution was misspecified • Estimates under Laplace were generally less biased than FO • No convergence problems under FO method
Ge, Bickel, & Rice (2004)	• CFO • Spline Approximation	• Model followed that of empirical example regularly found in Pharmacokinetics • Random effects distribution	• Inter-individual variability is small, CFO method is efficient and accurate in terms of parameter bias
Wu (2004)	• Exact method of integration • Approximate method of integration	• Response and covariate missingness • Random effects distribution • Sampling mechanism • Error distributions	• Missing data mechanism is non-ignorable, serious bias in the parameter estimates may occur

$$\Lambda_i = \sigma^2 \mathbf{I}_{n_i} \text{ where } \sigma = 2$$

$$\Phi = \begin{pmatrix} \varphi_{11} & & \\ \varphi_{21} & \varphi_{22} & \\ \varphi_{31} & \varphi_{32} & \varphi_{33} \end{pmatrix} = \begin{pmatrix} 25 & & \\ 3 & 4 & \\ 0.05 & 0.05 & 0.075 \end{pmatrix}$$

The empirical performance of each estimation method is evaluated with respect to bias, precision of estimation, and standard error ratios of the fixed parameters $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$, and σ^2. On the basis of ($\hat{\boldsymbol{\theta}}_b : b = 1,\ldots,500$) obtained from 500 replications, bias is calculated as the differences between the true population values and the means of the estimates obtained from the 500 replications. The variance of the estimates will be used to get some idea as to the precision with which parameters are estimated across study conditions. The variance is computed for the m^{th} element of parameter vector $\boldsymbol{\theta}$ as

$$\text{var}\left[\hat{\theta}(m)\right] = 500^{-1}\sum_{b=1}^{500}\left[\hat{\theta}_b(m) - \bar{\hat{\theta}}(m)\right]^2$$

where for a particular cell, $\bar{\hat{\theta}}(m)$ is the mean of the estimates across the 500 replications, and $\hat{\theta}(m)$ is the estimate obtained by the approach under consideration.

Sample Size and Sampling Scheme

In many applications using NLME models, the sample size is quite small. In a small simulation study, Pinheiro & Bates (1995) used $N = 10$ as the number. In practice, the sample sizes can of course be larger. The total number of subjects will be manipulated to be either: 50, 100, or 250 representing small, medium and large sample sizes, respectively. These correspond to sample sizes found in previous simulation studies (Hartford & Davidian, 2000) as well as empirical studies (see e.g., Cudeck, 1996).

Generated data had a maximum of $n_i = 8$ time points $t_{ij} = 0,\ldots,7$. For all cases, the intra-individual sampling scheme had five total conditions. Data contained either (*i*) no missingness ($n_i = 8$), (*ii*) 10% missing, or (*iii*) 20% missing. Because attrition and drop out seem to occur with some frequency in empirical studies, the missingness was implemented in two ways: (a) deleting the percentage of data for the corresponding time points at the end of the study, and (b) randomly selecting which times would be deleted using the *sample* function in R. R (R Core Team, 2014) was used as the data generation software. The *sample* function in R allows elements from a larger set of elements to be chosen at random.

Data were prohibited at the first time point to be deleted as we felt this was unrealistic in terms of practical data collection protocol – although each of the estimation methods could handle this nuance in a straightforward fashion.

Violation of Normality on Random Effects and Error Distributions

Several different distributions for $\mathbf{b}_i$ were used to generate random effects,

N. A normal distribution, $\mathbf{b}_i : N(\mathbf{0},\mathbf{\Phi})$

NN. A non-normal distribution with skew = 2 and kurtosis = 7. The non-normal condition was implemented using the procedure outlined in Headrick and Sawilowsky (1999).

M1. A mildly contaminated normal distribution, $\mathbf{b}_i : (1-\pi)N(\mathbf{0},\mathbf{\Phi}) + \pi N(\mathbf{0},\mathbf{\Phi}^*)$, with contamination fraction $\pi = 0.05$ and $\mathbf{\Phi}^*$ chosen as described below.

M2. A moderately contaminated normal distribution, $\mathbf{b}_i : (1-\pi)N(\mathbf{0},\mathbf{\Phi}) + \pi N(\mathbf{0},\mathbf{\Phi}^*)$, with contamination fraction $\pi = 0.10$ and $\mathbf{\Phi}^*$ chosen as described below.

Distribution N denotes the case where the usual assumption of normality on the random effects is applicable. Distribution NN represents a situation where the true distribution of the random effects is positively skewed and heavy-tailed than expected from a normal distribution but with the same variability in the population. Distributions M1 and M2 are meant to characterize the

$$\mathbf{\Phi}^* = \begin{pmatrix} 40 & & \\ 4.5 & 5.5 & \\ 0.07 & 0.07 & 0.095 \end{pmatrix}$$

chosen so that variability is larger but the correlation between effects in approximately the same as those in $\mathbf{\Phi}$. Conceptually, this represents the situation where the apparent inter-individual variation is greater than that in the target population of interest attributable to errors in sampling.

Two distributions for the intra-individual errors, $\mathbf{e}_i$ were used to generate the regression errors

NE. A normal distribution, $\mathbf{e}_i : N\left(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}\right)$

NNE. A non-normal condition with skew = 3 and kurtosis = 21, respectively. Similarly to the random effects generation, the non-normal condition was implemented using the procedure outlined in Headrick and Sawilowsky (1999).

Distribution NE represents the typical specification of a normal distribution with a simple independence structure. Distribution NNE represents a situation where the true distribution of the regression errors is positively skewed and heavy-tailed than expected from a normal distribution but with the same variability in the population.

A simulation scenario thus consisted of a particular choice of random effects distribution, choice of intra-individual error distribution, sampling scheme, and sample size. The full factorial of $4 \times 2 \times 5 \times 3 = 120$ possible combinations was investigated, where for each scenario, 500 Monte Carlo data sets were generated. For each data set in each scenario, fitting was carried out using each of the five estimation methods as described above. A summary of the manipulated conditions can be found in Table 2.

Table 2. Simulation conditions and levels

Manipulated Condition	# Levels	Levels
Sample Size	3	50, 100, 250
Random Effects Distribution	4	N, NN, M1, M2
Error Distribution	2	NE, NNE
Missingness	5	C, E-10, E-20, R-10, R-20

Note: Levels of the random effects distribution (N = normal, NN = non-normal, M1 = Contaminated 5%, M2 = Contaminated 10%). Levels of the error distribution (NE = normal, NNE = non-normal). Levels of missingness (C = complete cases, E-10 = 10% missing at the end, E-20 = 20% missing at the end, R-10 = 10% randomly missing, R-20 = 20% randomly missing)

Results

The simulations were conducted on several different platforms. The majority of the simulations were completed in a Windows environment on Dell Latitude and Dell Vostro workstations with duo-core processors. Consistency of results was examined across platforms to ensure that conclusions were the results of properties of the methods rather than numerical irregularities. Considering the simulation design, there were 120 fully-crossed conditions for each estimation method, and 500 data sets per scenario. As is often the case in fitting nonlinear mixed effects models by any estimation method, there were some convergence issues and other numerical problems. When numerical problems were encountered, the replicate was repeated with efforts to identify and correct the problem. Despite these efforts several nonconvergent data sets were still present. These trials were categorized as nonconvergent.

In all of the simulation trials, starting values were taken as the true values generating the data to allow the greatest possibility of automation of this large number of simulations. Of course, even in the most optimal condition combinations it may happen that universal convergence can never be achieved. This may be due to poor starting values, practical lack of identifiability with the specific available data, or other unknown factors. Several sets of starting values can be tried to address the first of these issues. However, because of the large number of replications, only limited attempts were made to emulate this "real" practice for initially nonconvergent data sets, which unfortunately did not improve the rate of convergence. The number of data sets (out of 500) for which satisfactory convergence was not achieved for each condition combination and estimation method are shown in Table 3.

There was no convergence problems encountered with the FO (First-order linearization), GHQ (Gaussian-Hermite quadrature), and BAY (Bayesian) methods, although a substantial amount of time was spent preliminarily to examine these methods under worst-case scenario conditions that were thought to influence the successful estimation of the model (i.e., number of quadrature points for the GHQ method, sensitivity of results and convergence to different prior distribution of the parameters for the Bayesian analysis, etc.). The FO method which linearizes the nonlinear function, making it the least computationally intensive method, exhibited no convergence problems what so ever. This is not to say that problems did not occur with these other methods.

The GHQ and FO methods, for example, did not demonstrate lack of convergence based on the default convergence criteria and settings in SAS PROC NLMIXED. Some strange behavior was noticed for several replicate data sets in the Bayesian analysis for the variance components of the model. The reasons behind the odd estimates appears to be that the Bayesian approach is quite sensitive to departures from the assumptions dictated by the prior and data distributions. That is, sensible estimates are not guaranteed for variance-covariance parameters using Bayesian estimation when the underlying distribution is far from the distributions that are presumed in the model set up.

Both the CFO (Conditional first-order) and GTS (Global two-stage) methods showed varying amounts of convergence issues although the number overall was not that significant. It should be noted that unlike the *nlme*( ) procedure in R, which uses the profiled loglikelihood to stabilize the optimization

Table 3. Rate of nonconvergence out of 500 trials for each distribution, sample size, missingness across estimation method.

			N					NN					M1					M2				
SS	ED	Meth	C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10
50	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	1	0	1	2	0	0	0	1	1	0	1	0	1	1	0	0	0	0
		GTS	0	0	0	2	2	0	0	2	4	4	0	0	2	3	2	0	0	0	3	2
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	1	1	2	1	0	0	0	0	5	0	2	0	2	2	3	0	0
		GTS	3	0	0	4	7	0	0	0	3	3	0	2	0	2	3	2	0	0	3	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	3	0	0	0	0	3	2	0	0	0	4	2	0	0	0	2	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	2	3	0	0	0	1	0	0	0	0	1	4	0	0	0	0	0
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
250	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	2	4	0	0	0	0	3	0	0	0	1	5	0	0	0	0	4
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	0	2	0	2	0	4	3	0	0	0	1	1	0	0	0	2	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: 50, 100, and 250. Missingness levels: C = Complete, E-10 = 10% missing at the end, E-20 = 20% missing at the end, R-10 = 10% randomly missing, R-20 = 20% randomly missing.

algorithm, the *nlinmix* macro in SAS, which was used to estimate the CFO method, does not use profiling. This appears to have some bearing on the stability of the algorithm to estimate parameters under non-ideal conditions. The GTS method uses both nonlinear least squares estimation (which is not affected by distributional assumptions) and PROC MIXED in SAS, which assumes normality in the random effects as well as the data distribution, and therefore could be

susceptible to convergence issues. Surprisingly, this method showed the greatest number of nonconvergent cases among the competitors.

Time to convergence was not an issue for either linearization method (i.e., FO or CFO) as both converged quickly for each replicate with average convergence time of 1.02 and 6.24 seconds, respectfully across all study conditions. Computational speed notwithstanding, time to convergence for these two methods increased as the sample size increased and with random effects distributions that departed from normality. The GTS method was slower to convergence than expected with an average replicate time to convergence of 55 seconds (range of 12.7 seconds under sample size of 50, no missing data, and normal distributions compared with 150.4 seconds per replicate under the most severe study conditions). This may be due to the stage 2 computation using PROC MIXED which utilizes the individual estimates in stage 1 iteratively to compute the variance components of the model. Surprisingly, the GHQ method was faster than expected overall (average time to convergence of 2 minutes per replicate), but suffered a lack of computational speed as the sample size increased and random effects distributions departed from normality. Under these severe conditions, the GHQ method took over 5 minutes to converge. Due to the preliminary investigative analyses, time to convergence for the BAY method was as expected with an average time to convergence of 75 seconds.

ANOVA and Classification Trees

Because of the large number of cells in the design coupled with the numerous parameters and outcomes to evaluate, it is instructive, if not necessary, to use quantitative procedures like analysis of variance (ANOVA) or classification trees as an initial filter of the results – to inform where real effects and "interesting" results occur. Factorial ANOVA was performed on each outcome variable (i.e., bias and parameter estimate variance) for each of the model parameters in –$\boldsymbol{\beta}$, $\boldsymbol{\Phi}$, and σ^2 modeling only main effects as well as two- and three-way interactions. Partial eta-squared, defined as the proportion of total variation attributable to the factor, excluding other factors from the total non-error variation (Pierce, Block & Aguinis, 2004), was used as the arbitrator in deciding which effects to examine more closely, using Cohen's (1988) heuristic value of (0.14 – large effect) as the cut point.

In conjunction with the ANOVA results, classification trees (Breiman, Friedman, Olshen, & Stone, 1984) were used to aid in determining which factors were most related to each of the outcomes while at the same time establishing

which levels were different from one another. The Chi-squared Automatic Interaction Detection (CHAID) method of constructing each tree (as implemented in SPSS version 20) is an exploratory tool that chooses the independent variable (factor) that has the strongest relation with the dependent variable. Categories of each factor are subsequently merged if they are not significantly different with respect to the dependent variable and the procedure stops when factors (independent variables) no longer affect the outcome. For illustrative purposes, the classification tree for the bias of the estimate of β_2 is shown in Figure 1 below.

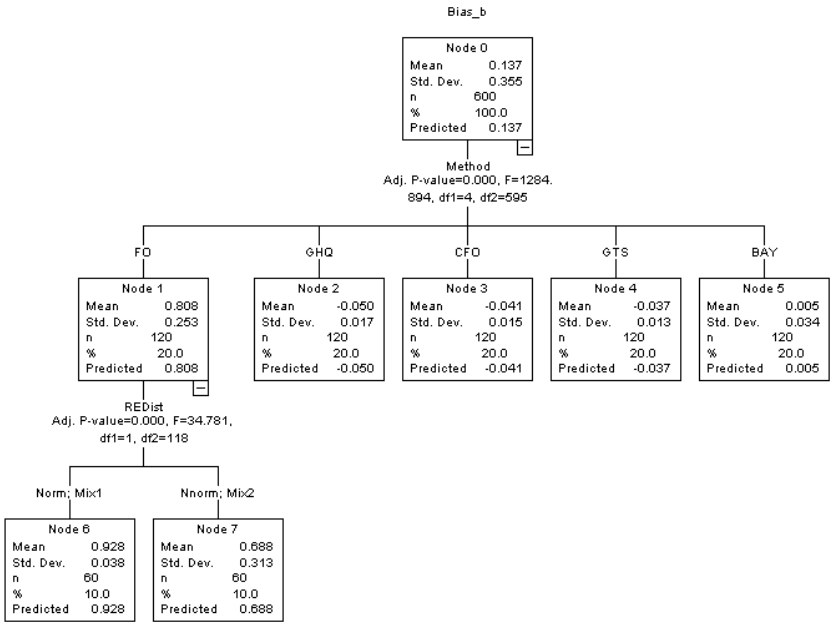

Figure 1. Classification tree for bias in β_2.

The first set of boxes below the initial node represents the method factor as being most related to differences in bias; and the procedure has determined that each method has mean bias that is statistically different from one another with the BAY method showing the least average parameter bias (0.005); the GTS, CFO, and GHQ methods showing comparable values (−0.037, −0.041, and −0.050, respectfully); and the FO clearly exhibiting larger average bias than its competitors (0.808). For the FO method it appears that the random effects

distributions which were non-normal and more severely contaminated as a mixture did not seem to have impacted the bias as much as the other distributional conditions. Evidently, no other factors contributed to delineating bias of β_2 further. Nodes representing factors that appear at subsequent levels in this hierarchical structure can be thought of as a type of interaction between itself and the node (or factor) above it. This interaction, however, is specific to particular levels of the factors involved. The entire set of ANOVA and classification tree results as well as tabulated mean bias and variance estimates can be found at the first author's website (http://www.education.umd.edu/EDMS/fac/Harring/webpage.html).

Table 4 and Table 5 summarize the results from the ANOVA and classification tree procedures. The ANOVA results for bias in the fixed regression coefficients are displayed in Table 4, which includes the variance components of the random effects associated with these fixed effects, and residual variance. The classification tree results corresponding to the parameters in Table 4 are compiled in Table 5.

Table 4. Main effects, two- and three-way interaction results from a factorial ANOVA for bias of parameters in $\boldsymbol{\beta}$, Φ, and σ^2.

	Bias β_1	**Bias β_2**	**Bias β_3**	**Bias φ_{11}**	**Bias φ_{22}**	**Bias φ_{33}**	**Bias σ^2**
Factor Combi-nations	M R M*R	M R MI S M*MI M*R M*S M*MI*S R*MI*S	M M*R		M R MI S M*MI M*R M*S M*R*MI M*MI*S M*R*S	R M*R*MI M*MI*S	

Note: M = Method, R = Random Effects Distribution, E = Error Distribution, MI = Missingness, S = Sample Size. The symbol '*' represents the interaction between effects present. To be included, the partial eta-squared for each effect was larger than 0.14 and the effect was significant at the 0.05 level.

Parameter Bias

No main effect or interaction effect was found for the bias in intercept or residual variance, φ_{11} and σ^2, respectively. Clearly, there were differences in bias across the five methods (M) for each of the regression coefficients; however, method of estimation only influenced the variance of β_{2i} among the variance parameters. This result coincides with the first column (node 1) in Table 5 from the classification tree analysis. Overall, the mean bias values for β_1, β_2, and β_3 were

negligible (−0.038, 0.137, −0.016), yet there were differences between the methods. For all regression parameters the FO method showed the greatest bias with the GHQ, CFO, and GTS methods producing less bias estimates. The BAY method constantly generated the least biased estimates (by a factor of 10) compared to the other methods excluding FO. From Table 5, it is clear that for β_1 and β_2, the random effects distribution significantly impacted the FO method with

Table 5. Results from a classification tree analysis for bias of parameters in β, Φ, and σ^2.

	Nodes		
	Level 1	**Level 2**	**Level 3**
Bias β_1	FO	N/M1 NN/M2	
	GHQ		
	CFO		
	GTS/ BAY	100 50/250	
Bias β_2	FO	N/M1 NN/M2	
	GHQ		
	CFO		
	GTS		
	BAY		
Bias β_3	FO		
	GHQ		
	CFO		
	GTS		
	BAY		
Bias φ_{11}	-	-	-
Bias φ_{22}	FO	N/M1 NN/M2	
	GHQ	N/NN M1/M2	
	CFO/ GTS	M2 M1 N NN	
	BAY		
Bias φ_{33}	N/M1/M2	CFO FO/GHQ/GTS/BAY	
	NN	FO/GTS GHQ/CFO/BAY	
Bias σ^2	-	-	-

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: N = 50, N = 100, and N = 250. The symbol '/' represents levels that are considered the same while levels on different lines are different. Levels are listed from top to bottom in order of magnitude of the bias (greatest to least).

the non-normal and more contaminated mixture distribution producing less bias estimates than the other distributions.

As for the parameters in $\mathbf{\Phi}$ and σ^2, in terms of bias, the estimation method, random effects distribution, and combinations of missingness and sample size were consequential. Also, as can be seem in Table 4, the error distribution factor did not influence bias of any parameter in the model including σ^2. Interestingly, no condition had an effect on the bias of φ_{11} (the variance for β_1), but many conditions, including the amount of missingness, impacted parameter bias for φ_{22} (1. 095 overall) with the GHQ method producing less biased estimates on average than the other methods (−0.015). Bias in φ_{33} was negligible (−0.003 overall) even though there were statistical differences across combinations of random effects distributions and methods.

Parameter Variance

In addition to evaluating the accuracy in terms of bias with which these methods produce parameter estimates, precision of estimation is also an important consideration. Table 6 and Table 7 display the summary of results of the factorial ANOVA and classification tree analyses for the variability outcome measure.

Expectedly, sample size was a primary factor in explaining differences in estimate variance with parameter variance decreasing as sample size increased from N = 50 to N = 250 (0.656 to 0.167). This pattern was evident for all the parameters in which factors impacted variance magnitude. For regression parameters, β_2 and β_3, precision was also impacted by method and random effects distribution with the GTS and BAY methods producing slightly smaller variance than GHQ with larger discrepancies found in the CFO and FO methods. The ANOVA results coincide with the classification tree results remarkably well, although with slightly different interaction effects. The only variance parameter that showed difference in precision across study conditions was φ_{22}. For this parameter, method seemed to have the most impact with the GHQ and BAY methods producing estimates with the greatest precision (1.43) followed by the CFO and GTS methods (7.72), and lastly the FO method (97.34). When the random effects distribution factor influenced precision, the non-normal distribution frequently produced more precise estimates (less variability) than either of the mixture distributions or normal distribution condition.

Table 6. Main effects, two- and three-way interaction results from a factorial ANOVA for variance of parameter estimates in β, Φ, and σ^2.

	Var β_1	Var β_2	Var β_3	Var φ_{11}	Var φ_{22}	Var φ_{33}	Var σ^2
Factor Combinations	S	M	M		M		
		R	R		R		
		S	S		S		
		M*R	M*R		M*R		
		M*S	M*S		M*S		
		R*S	R*S		R*S		
		M*R*S	M*R*S		M*R*S		

Note: M = Method, R = Random Effects Distribution, E = Error Distribution, MI = Missingness, S = Sample Size. The symbol '*' represents the interaction between effects present. To be included, the partial eta-squared for each effect was larger than 0.14 and the effect was significant at the 0.05 level.

Table 7.

	Nodes		
	Level 1	**Level 2**	**Level 3**
Var β_1	50	M2 NN/M1 N	
	100	M2 NN/M1 N	
	250		
Var β_2	50	FO/GHQ/CFO GTS/BAY	
	100	FO/CFO GHQ/GTS/BAY	
	250	N/M1/M2	FO/BAY GHQ/CFO/GTS
		NN	
Var β_3	50	N/M1/M2	FO/GHQ CFO/GTS/BAY
		NN	
	100	N/M1/M2	FO/GHQ CFO/GTS/BAY
		NN	
	250	FO/GHQ/GTS CFO/BAY	
Var φ_{11}	-	-	-
Var φ_{22}	FO GHQ/BAY	50 100 250	
	CFO/GTS	N/M1/M2	50/250 100
		NN	
Var φ_{33}	-	-	-
Var σ^2	-	-	-

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: 50, 100, and 250. The symbol '/' represents levels that are considered the same while levels on different lines are different.

Results from large simulation studies are often hard to digest simply by examining tables of values and trying to extract important trends and patterns. The following are the main conclusions from this simulation:

1. Data missingness and error variance distributions seemed to have little if any effect on parameter recovery or estimation precision across the five estimation methods – at least at the levels we investigated.
2. Although the quickest method to converge to a solution and the method least sensitive to starting values, the first-order (FO) linearization method showed the greatest bias across both fixed effects and variance/covariance parameters compared to its competitors.
3. For the other four methods, the GHQ and BAY methods produced the least biased fixed effects although four were comparable for the linear effects.
4. Although slowest time to convergence, the GHQ and BAY methods produced the least biased estimates of the parameters in $\mathbf{\Phi}$, while the CFO and GTS methods produced the least biased residual variance. Bias was greatest in these estimates when the sample size was small and/or the random effects distribution was non-normal.
5. Fixed effects were estimated more precisely by the GHQ and BAY methods. For these parameters, precision was affected most by small sample size and non-normal and mixture random effects distributions.
6. Again, the GHQ and BAY methods produced more precise estimates of the variance components of $\mathbf{\Phi}$. Expectedly, sample size was also a significant factor variability of the estimates decreasing as the sample size increases.
7. Fixed parameters estimates based on the CFO, BAY, and GTS are fairly robust to mild deviations from normality of both the random effects and error distributions even though these methods sometimes had convergence problems.

These results point to the following recommendations:

1. The FO approach is not recommended for nonlinear mixed effects models as it is the least accurate method for fixed parameter and variance components estimates.
2. The GHQ and BAY methods appear to produce the least biased parameter estimates with the GTS and CFO methods showing comparable results. The GTS, CFO, and BAY methods were more robust to modest departures from normality of the random effects distribution. Thus when the random effects distribution is approximately normal and the sample sizes small to modest, then the GHQ or BAY estimation methods are recommended. For larger sample sizes and deviations from normality, the CFO or the BAY methods are recommended.

The efficacy of the Bayesian approach should be investigated on its own merits and not necessarily compared to likelihood-based methods for estimating nonlinear mixed effects models. This stems from having set up the simulation somewhat unfairly. Apart from the philosophical differences that exist between frequentist and Bayesian approaches, the obvious advantages that the Bayesian framework offers was not exploited. For example, as was previously mentioned, in a Bayesian approach prior knowledge about model parameters including their distributional assumptions can be incorporated into the model formulation. In this simulation, non-informative conjugate priors were used, which put the preponderance of weight in estimating the posterior distribution on the data (or the likelihood). It would be expected that the Bayesian method under this scenario to behave very similarly to the marginal maximum likelihood method, which in this set of simulations it did so unsurprisingly. Further exploration into the methodological underpinnings and extensions of the Bayesian approach that were not investigated here are warranted.

The results of a Monte Carlo simulation study undertaken to gain insight into the consequences of violation of distributional assumptions, sample size, and data sparseness underlying five popular approximations used in fitting nonlinear mixed effects models. Although it is not appropriate to draw general conclusions from a single simulation study, the findings are suggestive and highlight several interesting features that may be worthy of future investigation. It appears that estimation of fixed regression parameters based on the CFO, BAY, and GTS – and to a lesser extent the GHQ approximation – methods is fairly robust to mild

deviations from normality of both the random effects and error distributions, although the GTS method did show difficulty in achieving convergence in a small number of replications. Overall, the FO method showed greater bias than the other methods for the fixed parameters and even more so for the variance components of the model. While it has the least computational burden of any of the methods, it is least accurate and therefore its usefulness in practice is questionable.

Of course, a single simulation cannot possibly examine all of the interesting facets of a model – even if the facility to carry out the computations was limitless. The same could be said of the levels within the manipulated factors that were investigated. Some rationale was provided for the choices knowing that there are infinitely many levels that ultimately could have been chosen. For example, Hartford and Davidian (2000) examined misspecification of the inter-individual model in Equation 2 looking at the performance of both the likelihood ratio test as well as the Wald test to test a single additive component. The current focus was on the estimation of fixed parameters, most of which ($\boldsymbol{\beta}$,$\boldsymbol{\Phi}$) characterize the population. Individual regression coefficients, predicted random effects, were not addressed, even though the NLME model is individual-specific. It is not unreasonable to expect that distributional assumptions or other model misspecifications would have more profound effects. Interestingly, methods of carrying out this prediction are markedly different for each of the estimation methods inspected in this study.

Through methodological advances in estimation algorithms and by the sheer speed of today's computing environments, the number of applications using the NLME model has steadily increased – particularly in the social and behavioral sciences. NLME models are important tools for practitioners interested modeling nonlinear change with functions that have at least one regression parameter that enters the function in a nonlinear manner. Much of the methodological and computational techniques for these models were developed in late 1980s through the early 2000s, although some work in the area still exists (Lai & Shih, 2003; Kuhn & Lavielle, 2005; Wu, 2008). As such, many of the estimation methods and optimization schemes for these models have been implemented in popular commercial software. Still a choice for a particular method is required, and often, that choice is made predicated on the research situation and on the specific software being used not necessarily on the merits of the method's performance under sub-optimal, but realistic, data analytic conditions. Overall the results highlight the inherent difficulty in specifying any type of complex model with latent unobservable components; a problem that suggests that caution is in order

in interpreting both the nature of computational issues and results in the event convergence is achieved.

Acknowledgements

This research was partially funded by the Institute of Education Sciences (R305A130042).

References

Beal, S. L., & Sheiner, L. B. (1982). Estimating population pharmacokinetics. *CRC Critical Reviews in Biomedical Engineering, 8*, 195-222.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures data. *Psychological Methods, 9*(3), 334-353. doi: 10.1037/1082-989X.9.3.334

Blozis, S. A., & Cudeck, R. (1999). Conditionally linear mixed-effects models with latent variable covariates. *Journal of Educational & Behavioral Statistics, 24*(3), 245-270. doi: 10.3102/10769986024003245

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall/CRC.

Brooks, S. P., & Gelman A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455. doi: 10.1080/10618600.1998.10474787

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171-197). Amsterdam: Elsevier Science.

Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development, 62*(1), 23-43. doi: 10.2307/1130702

Burke, C. T., Shrout, P. E., & Bolger, N. (2007). Individual differences in adjustment to spousal loss: A nonlinear mixed model analysis. *The International Journal of Behavioral Development, 31*(4), 405-415. doi: 10.1177/0165025407077758

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, *31*(3), 371-403. doi: 10.1207/s15327906mbr3103_6

Cudeck, R., & Harring, J. R. (2007). The analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, *58*, 615-637. doi: 10.1146/annurev.psych.58.110405.085520

Davidian, M., & Giltinan, D. M. (1993). Some general estimation methods for nonlinear mixed-effects models. *Journal of Biopharmaceutical Statistics*, *3*(1), 23-55. doi: 10.1080/10543409308835047

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. London: Chapman & Hall.

Davidian, M**.,** & Giltinan, D. M. (2003). Nonlinear models for repeated measurements: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*, 387-419. doi: 10.1198/1085711032697

Demidenko, E. (2004). *Mixed models*. New York: Wiley

Ge, Z., Bickel, P. J., & Rice, J. A. (2004). An approximate likelihood approach to nonlinear mixed effects models via spline approximation. *Computational Statistics & Data Analysis*, *46*(4), 747-776. doi: 10.1016/j.csda.2003.10.011

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in M*plus* and SAS. *Structural Equation Modeling, A Multidisciplinary Journal*, *16*(4), 676-701. doi: 10.1080/10705510903206055

Harring, J. R., Cudeck, R., & du Toit, S. H. C. (2006). Fitting partially nonlinear random coefficient models as SEMs. *Multivariate Behavioral Research*, *41*, 579-596. doi: 10.1207/s15327906mbr4104_7

Hartford, A., & Davidian, M. (2000). Consequences of misspecifying distributional assumptions in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, *34*(2), 139–164. doi: 10.1016/S0167-9473(99)00076-6

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*(1), 25-35. doi: 10.1007/BF02294317

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, *42*(4), 805-820. doi: 10.2307/2530695

Kuhn, E., & Lavielle, M. (2005) Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, *49*(4), 1020-1038. doi: 10.1016/j.csda.2004.07.002

Lai, T. L., & Shih, M. C. (2003). Nonparametric estimation in nonlinear mixed effects models. *Biometrika*, *90*(1), 1-13. doi: 10.1093/biomet/90.1.1

Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, *38*(4), 963-974. doi: 10.2307/2529876

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*(3), 673-687. doi: 10.2307/2532087

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New Jersey: Springer.

Martins, C. R., Oulhaj, A. A., de Jager, C. A., & Williams, J. H. (2005). APOE alleles predict the rate of cognitive decline in Alzheimer disease: A nonlinear model. *Neurology*, *65*(12), 1888-1893. doi: 10.1212/01.wnl.0000188871.74093.12

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122. doi 10.1007/BF02294746

Paap, R. (2002). What are the advantages of MCMC based inference in latent variable models? *Statistica Neerlandica*, *56*(1), 2-22. doi: 10.1111/1467-9574.00060

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational & Psychological Measurement*, *64*(6), 916–924. doi: 10.1177/0013164404264848

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12-35. Doi: 10.1080/10618600.1995.10474663

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Roe, D. J., Vonesh, E. F., Wolfinger, R. D., Mesnil, F., & Mallet, A. (1997). Comparison of population pharmacokinetic modeling methods using simulated data: results from the Population Modeling Workgroup. *Statistics in Medicine*,

16(11), 1241-1262. doi: 10.1002/(SICI)1097-0258(19970615)16:11<1241::AID-SIM527>3.0.CO;2-C

Segawa, E. (1998, November). Application of hierarchical nonlinear models to children's growth in vocabulary and height. *Dissertation Abstracts International*, 59, Retrieved from EBSCO*host*.

Sheiner, L. B., & Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters, I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, *8*(6), 553-571. doi: 10.1007/BF01060053

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. London: Chapman & Hall.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2002). *WinBugs User Manual (Version 1.4)*. Cambridge, UK: MRC Biostatistics Unit 26.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*(3), 1-16. doi: 10.18637/jss.v012.i03

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wolfinger, R. D. (1993). Laplace's approximation for nonlinear mixed models, *Biometrika*, *80*(4), 791-795. doi: 10.1093/biomet/80.4.791

Wolfinger, R. D. (1999). *Fitting Nonlinear Mixed Models with the New NLMIXED Procedure, Paper 287* [Technical Report]. Cary, NC: SAS Institute Inc.

Wolfinger, R. D., & Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed effects models. *Computational Statistics and Data Analysis*, *25*(4), 465-490. doi: 10.1016/S0167-9473(97)00012-1

Wu, L. (2004). Nonlinear mixed effects models with nonignorably missing covariates. *Canadian Journal of Statistics*, *32*(1), 27-37. doi: 10.2307/3315997

Wu, L. (2008). An approximate method for nonlinear mixed-effects models with nonignorably missing covariates. *Statistics & Probability Letters*, *78*(4), 384-38. doi: 10.1016/j.spl.2007.07.011

Appendix A

Data for the simulation was generated in R (V 3.0.1). The following input statements were used to run each of the methods in the various statistical software programs.

First-Order Linearization (FIRO) Using SAS PROC NLMIXED

```
proc nlmixed data=aera  method=firo tech=quanew lis=2 lsp=.005 maxfu=5000
maxit=2000;
parms  au=100 bu=10 cu=1 sa=25 sb=1, sc=0.075 sab=3 sac=0.05 sbc=0.05 se=4;
a=au+ai;
b=bu+bi;
c=cu+ci;
mod= b-(b-a)*exp(-c*(time-1));
model aera ~ normal(mod,se);
random ai bi ci ~ normal([0,0,0],[sa,sab,sb,sac,sbc,sc]) subject=id;
run;
```

Global Two-Stage (GTS) Using SAS Macro

```
%macro GTS(size);
proc iml;
print &size;

*first stage estimate of individual person parameter;
%do k=1 %to 100;
proc iml;
use aera.aera;
read all;
dat=time||y||id;
uid=t(unique(id));
m=nrow(uid);
n=nrow(id);
p=3;
dati=J(8,3,0);
create indivdat from dati [colname={'time' 'y' 'subj'}] ;
do i=1 to 8;
dati[i,]=dat[i+(&k-1)*8,];
end;

append from dati;
quit;

proc nlin data=indivdat noprint save outest=test ;
  parms  b1=100, b2=10, b3=1;
  model y=b2-(b2-b1)*exp(-b3*(time-1));
  output out=nlinout predicted=pred residual=res ;
  run;
data par; set test; if _type_ ne "FINAL" then delete; subj=&k; keep subj
_status_ b1 b2 b3; run;
```

```
proc append base=stage1par data=par force;
proc append base=stage1pred data=nlinout force;
%end;
proc iml;
* read in nlin estimated results;
use stage1par;
read all into bols [colname=name];
use stage1pred;
read all;
n=nrow(pred);
m=nrow(bols);
p=3;

*pooled ols estimate of sigma;
sigma=sum(res#res)/(n-m*p);
create var_e from sigma [colname={'error variance'}];
append from sigma;

*get covariance matrix for each bols;
*prepare data for proc mixed analysis;
thisdati=J(3,8,0);
create mixdat from thisdati [colname={'id' 'y' 'x1' 'x2' 'x3' 'z1' 'z2' 'z3'}];
prednew=J(m,8,0);
grd1=J(1,8,0);
grd2=J(1,8,0);
grd3=J(1,8,0);
x={1,2,3,4,5,6,7,8};
do l=1 to 100;
  do j=1 to 8;
prednew[l,j]=bols[l,2] -(bols[l,2]-bols[l,1])*exp(-bols[l,3]*(x[j]-1));
grd1[j]=exp(-bols[l,3]*(x[j]-1));
grd2[j]=1-exp(-bols[l,3]*(x[j]-1));
grd3[j]=(bols[l,2]-bols[l,1])*(x[j]-1)*(exp(-bols[l,3]*(x[j]-1)));
grd=t(grd1)||t(grd2)||t(grd3);
  end;
thisid=J(p,1,l);
bi=I(3);
A=sigma*solve(t(grd)* grd,bi);
chalf=root(solve(A,bi));
respi=chalf*t(bols [l,1:3]) ;
thisxi=chalf;
thisdati=thisid||respi||thisxi||thisxi;
append from thisdati;
end;
quit;

*final population parameter estimate;
proc mixed data=mixdat method=ml covtest;
  class id;
  model y = x1 x2 x3  / noint solution chisq;
  random z1 z2 z3/ subject=id type=un  g gcorr gc;
  parms (25) (3) (1) (0.06) (0.06) (0.075) (1) / eqcons=7;
  ods output solutionf=fixedparms;
  ods output CovParms=covparms;
```

```
run;
%mend GTS;
```

Conditional First-Order Linearization (CFO) Using SAS Macro NLINMIX

```
%nlinmix(data=dat,
     model=%str(
        a=au+ai;
        b=bu+bi;
        c=cu+ci;
        predv= b-(b-a)*exp(-c*(time-1));
   ),
   parms=%str(au=100 bu=10 cu=.75),
   stmts=%str(
      class id;
      model pseudo_y = d_au d_bu d_cu  / noint notest solution cl;
      random d_ai d_bi d_ci / type=un subject=id solution;
   ),
   expand=eblup
   ),
run;
```

Gaussian-Hermite Quadrature (GHQ) Using SAS PROC NLMIXED

```
proc nlmixed data=aera method=gauss noad tech=quanew lis=2 lsp=.005 maxfu=5000
maxit=2000 qpoints=20;
parms  au=100 bu=10 cu=1 sa=25 sb=1, sc=0.075 sab=3 sac=0.05 sbc=0.05 se=4;
a=au+ai;
b=bu+bi;
c=cu+ci;
mod= b-(b-a)*exp(-c*(time-1));
model aera ~ normal(mod,se);
random ai bi ci ~ normal([0,0,0],[sa,sab,sb,sac,sbc,sc]) subject=id;
run;
```

Bayesian (BAY) Using R and WinBUGS

The Bayesian approach used the R2WinBUGS library and bugs function in R. R was utilized as the platform to call WinBUGS and collate results upon convergence of the program. There is a debugging option in the bugs function that allows monitoring of the iteration history and mixing. We used this extensively in the beginning to identify problematic code. The bugs function requires three files to call the WinBUGS program:

```
nlme.sim <- bugs(data, inits, parameters, "C:/ /programs/
   quadwin.txt",
   n.chains=3, n.iter=9000, n.burnin=7000,
   bugs.directory="C:/Program Files/WinBUGS14",
```

```
   n.thin = 1, debug=T)
```

File 1: Initial Values (init)

```
inits = function(){
          list(mub=c(100,10,1),
         tau=matrix(c(.05,0,0,0,.25,0,0,0,20),nrow=3,byrow=F),
          tauC=.5)
}
```

File 2: Parameters to Monitor (parameters)

```
parameters = c("mub", "sig", "sige")
```

File 3: Model Statement (quadwin.txt)

```
model {

for (i in 1:K) {
       for (j in 1:n) {
z[i, j] ~ dnorm(mnb[i, j], tauC)
mnb[i, j] <- b[i, 2] - ((b[i,2] - b[i,1])*exp(-b[i,3] * x[j]))
                    }
       b[i, 1:3] ~ dmnorm(mub[1:3], tau[1:3,1:3])
}

       mub[1:3] ~ dmnorm(mean[1:3], S2[1:3,1:3])
       tau[1:3, 1:3] ~ dwish(S3[1:3,1:3], 3)
       sigma2[1:3, 1:3] <- inverse(tau[1:3,1:3])
       sig[1,1] <- sigma2[1,1]
       sig[1,2] <- sigma2[2,1]
       sig[2,2] <- sigma2[2,2]
       sig[1,3] <- sigma2[3,1]
       sig[2,3] <- sigma2[3,2]
       sig[3,3] <- sigma2[3,3]
       tauC ~ dgamma(1.0E-3, 1.0E-3)
       sige <- 1 / tauC
}
```