The Theory and Methods for Measurement Errors and Missing Data Problems in Semiparametric Nonlinear Mixed-effects Models

by

WEI LIU

B.Sc., Northeast Normal University, 1990

M.Sc., Northeast Normal University, 1993

M.Sc., Memorial University of Newfoundland, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

The University of British Columbia

October 2006

Abstract

Semiparametric nonlinear mixed-effects (NLME) models are flexible for modelling complex longitudinal data. Covariates are usually introduced in the models to partially explain inter-individual variations. Some covariates, however, may be measured with substantial errors. Moreover, the responses may be missing and the missingness may be nonignorable. In this thesis, we develop approximate maximum likelihood inference in the following three problems: (1). semiparametric NLME models with measurement errors and missing data in time-varying covariates; (2). semiparametric NLME models with covariate measurement errors and *outcome-based informative* missing responses; (3). semiparametric NLME models with covariate measurement errors and *random-effect-based informative* missing responses. Measurement errors, dropouts, and missing data are addressed simultaneously in a unified way. For each problem, we propose two joint model methods to simultaneously obtain approximate maximum likelihood estimates (MLEs) of all model parameters. Some asymptotic properties of the estimates are discussed. The proposed methods are illustrated in a HIV data example. Simulation results show that all proposed methods perform better than the commonly used two-step method and the naive method.

Contents

List of Tables

List of Figures

Acknowledgements

First of all, I thank my supervisor, Dr. Lang Wu, for his support, both academic and moral. I feel honored to have had the chance to work with him. Very special thanks are due the other members of my supervisory committee, Dr. Paul Gustafson and Dr. Matias Salibian-Barrera, for their invaluable suggestions and encouragement. Also, I am very grateful to Dr. Nancy Heckmen, Dr. Harry Joe, and Dr. John Petkau for teaching me, academically and personally.

I am also indebted to the office staff, Christine Graham, Elaine Salameh, and Rhoda Morgan, for all their help with administrative matters. It is a pleasure to acknowledge helpful discussions with my fellow graduate students: Juxin Liu, Guohua Yan, Weiliang Qiu, Hui Shen, and all of the others.

Finally, I would like to thank my family, my husband, and my son for their tireless love, patience, understanding, and support.

WEI LIU

The University of British Columbia

September 2006

To my family

Chapter 1

Introduction

1.1 Longitudinal Studies

A longitudinal study is defined as a study in which the response for each individual in the study is observed on two or more occasions. Longitudinal studies are very common in health and life sciences, epidemiology, medical, and biomedical research. Longitudinal studies are also common in other areas including education, psychology, social sciences, and econometrics. A major advantage of longitudinal studies over cross-sectional studies is that in longitudinal studies one can model the individual response trajectory over time while in cross-sectional studies one cannot.

In longitudinal studies, covariates may be classified into two categories: time-varying covariates and time-independent covariates. Time-varying covariates represent variables which vary over time within individuals. Time itself may be viewed as a covariate in that often there is interest in testing whether there are any changes in the response variable over time. When one studies children's weight trajectories over time, the height may be a time-varying covariate which can change with time. Time-independent covariates, on the

other hand, may represent baseline factors which do not vary with time. Examples of time-independent covariates might include an individual's gender and race. One of the goals in longitudinal research is to investigate the effects of important covariates on individual response trajectories over time.

A defining feature of a longitudinal data set is repeated observations on a number of individuals. The repeated observations on the same individual tend to be correlated. It is important to explicitly recognize two sources of variability in a longitudinal data set: random variation among repeated measurements within a given individual and random variation between individuals. Moreover, the number of observations within individuals often varies from individual to individual (i.e., the data are often unbalanced). Therefore, longitudinal data require special statistical methods to draw valid statistical inferences.

There are three approaches to a longitudinal data analysis. The *marginal model* approach is to model the marginal expectation of a response as a function of covariates. The methods are designed to permit separate modelling of the regression of the response on covariates, and the association among repeated observations of the response for each individual. Marginal models are appropriate when inferences about the population average are the main interest. For example, in a clinical trial the average difference between control and treatment is most important, not the difference for any one individual.

The *random effects model* approach assumes that the response is a function of covariates with regression coefficients varying from one individual to the next. A random effects model is a reasonable description if the set of coefficients for a set of individuals can be thought of as a sample from a distribution. In random effects models, correlation arises among repeated responses because the regression coefficients vary across individuals, and regression coefficients represent the effects of the covariates on an individual, which is in contrast to the marginal model coefficients which describe the effect of covariates on the

population average. Random effects models are most useful when the objective is to make inference about individuals, such as in AIDS studies. They may focus on both population parameters and individuals characteristics.

The *transition model* approach describes the conditional distribution of each response on an individual as an explicit function of his past responses and covariates. Under transition models, correlation among the response observations on one individual exists because the past response observations explicitly influence the present response observation. The past response observations are treated as additional covariates.

In each of the three approaches, we consider both the dependence of the responses on covariates and the correlation among the responses. With cross-sectional data, only the dependence of the responses on covariates needs to be specified since there is no correlation of responses. In longitudinal studies, in which correlation usually exists among responses, there are at least two consequences of ignoring it as follows. First, incorrect inferences about regression coefficients. In particular, confidence intervals are too short based on assumption of independence when in fact there is positive dependence. Secondly, the estimation method may be inefficient, that is, less precise than possible.

1.2 Parametric Nonlinear Mixed-effects Models

Parametric nonlinear mixed-effects (NLME) models, or hierarchical nonlinear models, have been widely used in many longitudinal studies such as human immunodeficiency virus (HIV) viral dynamics, pharmacokinetic analyses, and studies of growth and decay (Davidian and Giltinan 1995; Vonesh and Chinchilli 1997). In these studies, the intra-individual variation and the inter-individual variation are typically modelled by a two-stage hierarchical model. The first stage specifies the mean and covariance structure for a given individual, whereas

the second stage characterizes the inter-individual variation. Understanding the nature of inter-individual systematic and random variation at the second stage often receives more emphasis. This inter-individual variation may be partially explained by some baseline or time-varying covariates.

Suppose that there are n individuals with measurements over time. Let y_{ij} and $\mathbf{z}_{ij}$ respectively be the response value and the $\nu \times 1$ covariate values for individual i at time t_{ij}, $i = 1, \ldots, n,\ j = 1, \ldots, n_i$. The covariates $\mathbf{z}_{ij}$ may incorporate variables such as time, dose, etc. A general parametric NLME model can be written as a hierarchical two-stage model as follows (Davidian and Giltinan, 1995)

$$y_{ij} = g(\mathbf{z}_{ij};\, \boldsymbol{\beta}_{ij}) + e_{ij}, \quad \mathbf{e}_i | \boldsymbol{\beta}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \delta^2 I), \tag{1.1}$$

$$\boldsymbol{\beta}_{ij} = \mathbf{d}(\mathbf{z}_{ij};\, \boldsymbol{\beta},\, \mathbf{b}_i), \quad \mathbf{b}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, B), \quad i = 1, \ldots, n,\ j = 1, \ldots, n_i, \tag{1.2}$$

where $g(\cdot)$ and $\mathbf{d}(\cdot)$ are known (possibly nonlinear) functions, $\boldsymbol{\beta}_{ij}$ are individual-specific parameters, $\boldsymbol{\beta}$ are population parameters (fixed effects), $\mathbf{b}_i$ are random effects, $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_{i1}^T, \ldots, \boldsymbol{\beta}_{in_i}^T)^T$, $\mathbf{e}_i = (e_{i1}, \ldots, e_{in_i})^T$ are within-individual random errors and are assumed to be independent of $\mathbf{b}_i$, δ^2 is the unknown within-individual variance, I is the identity matrix, and B is an unknown variance-covariance matrix.

In AIDS studies, for example, viral loads (Plasma HIV-1 RNA copies) and various covariates such as CD4 count are usually measured over time after initiation of treatments. The following parametric NLME model has been widely used to fit short-term (the first three months after treatments) HIV viral dynamics (Wu, 2002; Wu and Zhang, 2002)

$$y_{ij} = \log_{10}(P_{1i} e^{-\lambda_{1ij} t_{ij}} + P_{2i} e^{-\lambda_{2ij} t_{ij}}) + e_{ij}, \tag{1.3}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z_{ij} + b_{2i}, \tag{1.4}$$

$$\log(P_{2i}) = \beta_4 + b_{3i}, \qquad \lambda_{2ij} = \beta_5 + b_{4i}, \tag{1.5}$$

where y_{ij} and z_{ij} are the $\log_{10}$-transformation of the viral load measurement and CD4 count

for patient i at time t_{ij} respectively, $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$ are random effects, P_{1i} and P_{2i} are baseline values, and λ_{1ij} and λ_{2ij} are the first (initial) and the second phase viral decay rates respectively (they may be interpreted as the turnover rates of productively infected cells and long-lived and/or latently infected cells respectively).

Although NLME models are popular in practice, their use has been somewhat limited because of the complexity of the likelihood function. Estimation of model parameters based on maximum likelihood can be challenging since these models are typically nonlinear with respect to the random effects and thus have no closed-form expressions for the marginal likelihood. This has led to the development of some widely used approximate methods based on Taylor expansions or Laplace approximation of the likelihood function (Lindstrom and Bates 1990; Wolfinger 1993; Vonesh, Wangs, Nie, and Majumdar 2002). These approximate methods are computationally efficient in the sense that they may converge faster and have less computational problems than the "exact" likelihood method, which finds the maximum likelihood estimator (MLE) using numerical integration techniques or Monte Carlo methods. These approximate methods often perform well if the number of intra-individual measurements is not small, but their performance may be less satisfactory if the intra-individual data are sparse, especially when the inter-individual variability is large (Davidian and Giltinan 1995; Vonesh and Chinchilli 1997; Pinheiro and Bates 1995). Thus there is still a need for developing "exact" methods. "Exact" likelihood inference for generalized linear mixed models based on Monte Carlo EM algorithms has been investigated by McCulloch (1997) and Booth and Hobert (1999).

1.3 Semiparametric Nonlinear Mixed-effects Models

Parametric NLME models are powerful tools in many longitudinal analyses. In some cases, however, parametric NLME models may not be flexible enough in modelling complex longitudinal processes, since the underlying mechanism which generates the data may be complicated in practice. In these cases, semiparametric or nonparametric models may be more flexible in modelling the complex longitudinal process (Ke and Wang, 2001; Rice and Wu, 2001). In particular, semiparametric NLME models are very useful in characterizing both the intra-individual variation and the inter-individual variation, in which the intra-individual variation is modelled semiparametrically while the inter-individual variation is incorporated by random effects (Davidian and Giltinan, 1995; Ke and Wang, 2001; Wu and Zhang, 2002).

In AIDS studies, for instance, the parametric NLME model (1.3) – (1.5) is appropriate only for fitting short-term HIV viral dynamics. Due to long-term clinical factors, drug resistance, and other complications, the viral load trajectories can be very complex after the initial phase viral decay (see Figure 1.1 for long-term viral load trajectories of six randomly selected HIV patients). Grossman et al. (1999) pointed out that viral decay rates after the initial period may be complicated and may vary over time since they may depend on some phenomenological parameters which hide considerable microscopic complexity and change over time. Therefore, a nonparametric smooth curve modelling for the second phase viral decay rate may be more appropriate than parametric modelling (Wu and Zhang, 2002). This leads to the following semiparametric NLME model

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{1.6}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z_{ij} + b_{2i}, \tag{1.7}$$

$$\log(P_{2i}) = \beta_4 + b_{3i}, \qquad \lambda_{2ij} = w(t_{ij}) + h_i(t_{ij}), \tag{1.8}$$

where $w(\cdot)$ and $h_i(\cdot)$ in (1.8) are unknown nonparametric smooth fixed- and random-effects

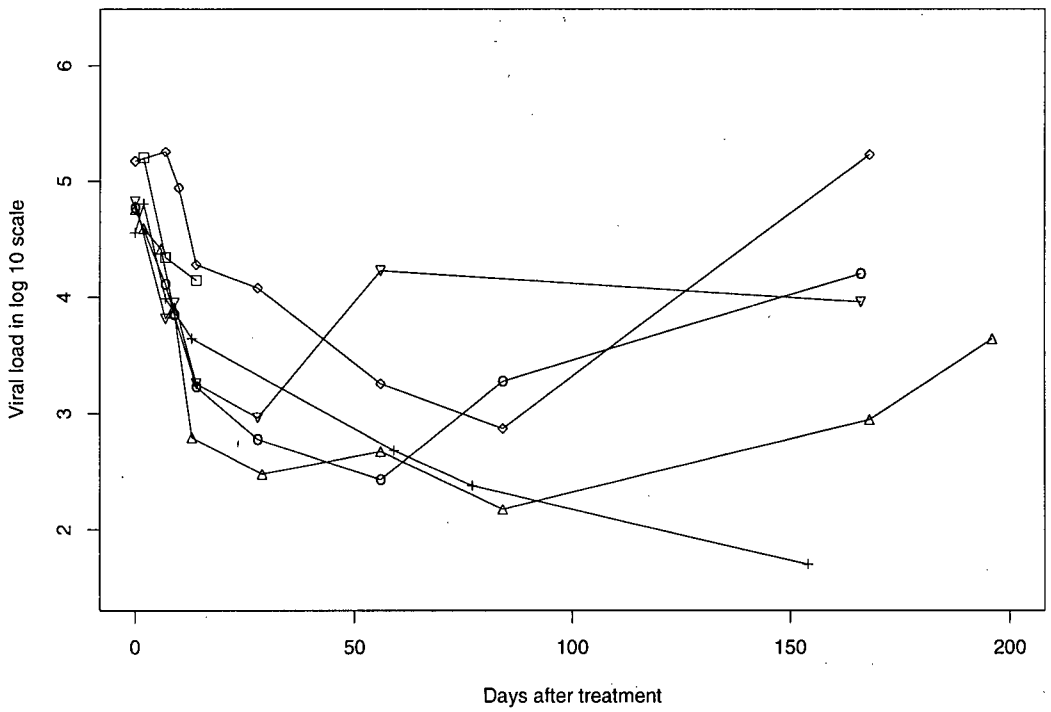

Figure 1.1: Viral loads (response) of six randomly selected HIV patients.

functions used to describe the complicated second phase decay rate λ_{2ij}.

Wu and Zhang (2002) introduced a class of semiparametric NLME models for longitudinal data. The standard parametric NLME models can be regarded as a special case of their models. Their models are more flexible than the semiparametric NLME models proposed by Ke and Wang (2001). Details of the semiparametric NLME models proposed by Wu and Zhang (2002) will be described in Chapter 2.

1.4 Measurement Errors and Dropouts

In many longitudinal studies, the inter-individual variation may be large and this variation may be partially explained by time-varying covariates. Some covariates, however, may be measured with substantial errors and may contain missing values as well. Ignoring measurement errors and missing data in covariates may lead to biased results (Carroll et al. 1995; Higgins et al. 1997; Wu, 2002). Moreover, some individuals may drop out of the study before the scheduled end for various reasons such as drug intolerance, which leads to missing data. Measurement errors and missing data make statistical analysis in longitudinal studies much more complicated, because standard complete-data methods are not directly applicable. Therefore, it is very important to find appropriate methods to deal with measurement errors and missing data.

In AIDS studies, for example, it is well known that CD4 counts, which may be used as covariates, are usually measured with substantial errors and are usually measured at time points different from the response (viral load) measurement schedule. In addition, it is very common that some patients drop out of the study early or miss scheduled visits due to drug intolerance or other problems. Visual inspection of the raw data seems to indicate that dropout patients may have slower viral decay, compared with the remaining patients. Thus,

the dropouts are likely to be informative or nonignorable.

Commonly used measurement error models are reviewed in Carroll et al. (1995). For NLME models with covariate measurement errors, Higgins, et al. (1997) proposed a two-step method and a bootstrap method, and Wu (2002) considered censored response and covariate measurement errors based on a joint model. There is also extensive literature on dropouts in longitudinal studies (e.g., Diggle and Kenward, 1994; Little 1995; Ibrahim et al. 2001). However, there is little literature on addressing measurement errors, informative dropouts, and missing data in semiparametric NLME models.

In the presence of missing data, the missing data mechanism must be taken into account to obtain valid statistical inferences. Little and Rubin (1987) and Little (1995) discussed statistical analyses with missing values. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be a vector of repeated observations of a variable y on individual i. Write $\mathbf{y}_i = (\mathbf{y}_i^{(o)}, \mathbf{y}_i^{(m)})$, with $\mathbf{y}_i^{(o)}$ denoting the observed components of $\mathbf{y}_i$ and $\mathbf{y}_i^{(m)}$ denoting the missing components of $\mathbf{y}_i$. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{in_i})^T$ denote a set of indicator variables such that $r_{ij} = 1$ if y_{ij} is missing and $r_{ij} = 0$ otherwise. The probability distribution of $\mathbf{r}_i$ defines a probability model for the missing value mechanism. Little and Rubin (1987) classified the missing value mechanism as follows.

- Missing data are *missing completely at random* (MCAR) if the probability of missingness is independent of both observed and unobserved data. When missing data are caused by features of the study design, rather than the behavior of the study subjects, the MCAR mechanism may be plausible. For example, some values are missing because of reasons irrelevant to the treatment (e.g., the medical equipment is broken down on a certain day). So missingness is MCAR if $\mathbf{r}_i$ is independent of both $\mathbf{y}_i^{(o)}$ and $\mathbf{y}_i^{(m)}$

- Missing data are *missing at random* (MAR) if the probability of missingness depends

only on observed data, but not on unobserved data. For example, a patient may fail to visit the clinic because he/she is too old. In mathematical notation, missingness is MAR if $\mathbf{r}_i$ is independent of $\mathbf{y}_i^{(m)}$.

- Missing data are *nonignorable* or *informative* (NIM) if the probability of missingness depends on unobserved data. For random effects models, we consider the following two nonignorable response missing mechanisms. First, the probability of the missingness depends on unobserved responses. For example, a patient fails to visit the clinic because he/she is too sick. We call the missingness *outcome-based informative* (Little, 1995) if $\mathbf{r}_i$ is dependent on $\mathbf{y}_i^{(m)}$, but not on the random effects $\mathbf{b}_i$. Secondly, the probability of missingness depends on unobservable random effects. For example, an AIDS patient may drop out if his/her individual-specific viral decay is too slow. We call the missingness *random-effect-based informative* (Little, 1995) if $\mathbf{r}_i$ is dependent on random effects $\mathbf{b}_i$, but not on $\mathbf{y}_i^{(m)}$.

Both MCAR and MAR missing mechanisms are sometimes referred to without distinction as *ignorable*. Little and Rubin (1987) showed that, when missing data are nonignorable, likelihood inference must incorporate the missing data mechanism to avoid biased results.

1.5 A Motivating Example

Our research is motivated by HIV viral dynamic studies, which model the viral load trajectories after initiation of anti-HIV treatments. HIV viral dynamic models have received great attention in AIDS studies in recent years (Ho et al. 1995; Perelson et al. 1996; Wu and Ding, 1999; Wu, 2005). These viral dynamic models provide good understanding of the pathogenesis of HIV infection and evaluation of anti-HIV therapies. NLME models have been popular in modelling the initial period of HIV viral dynamics and in characterizing

the large inter-patient variation. It is shown that the initial viral decay rate may reflect the efficacy of the anti-HIV therapy (Ding and Wu, 2001). One of the major challenges in modelling long-term HIV viral dynamics is that, during late stages of an anti-HIV treatment, it is difficult to model the viral load trajectory parametrically, because drug resistance, non-compliance, and other long-term clinical factors may affect viral load trajectories. Therefore, semiparametric NLME models may be more suitable for modelling HIV viral dynamics (Wu and Zhang, 2002).

Understanding the large inter-patient variation in HIV viral dynamic studies often receives great attention, which may help to provide individualized treatments. It has been shown that covariates such as CD4 cell count (see Figure 1.2) may partially explain the large inter-patient variation (Wu et al. 1999; Wu, 2002). However, some covariates such as CD4 cell count may be measured with substantial errors and may be measured at time points different from the response measurement schedule (which leads to missing data in covariates). Ignoring these measurement errors and missing data in covariates may lead to biased results (Wu, 2002). In addition, it is very common that some patients may drop out of the study early or miss scheduled visits due to drug resistance/intolerance and other problems (although dropout patients may return to study later). It appears that dropout patients may have slower viral decay rates, compared with the remaining patients (see Figure 1.1). Thus the dropouts are likely to be informative or nonignorable. Therefore, it is important to address measurement errors, informative dropouts, and missing data in semiparametric NLME models in order to obtain reliable results, which may make significant contributions to HIV/AIDS studies.

The following AIDS dataset motivates our research. A more detailed data description can be found in Wu (2002). The dataset includes 53 HIV infected patients who were treated with a potent antiretroviral regimen. Viral loads (Plasma HIV-1 RNA copies) were measured

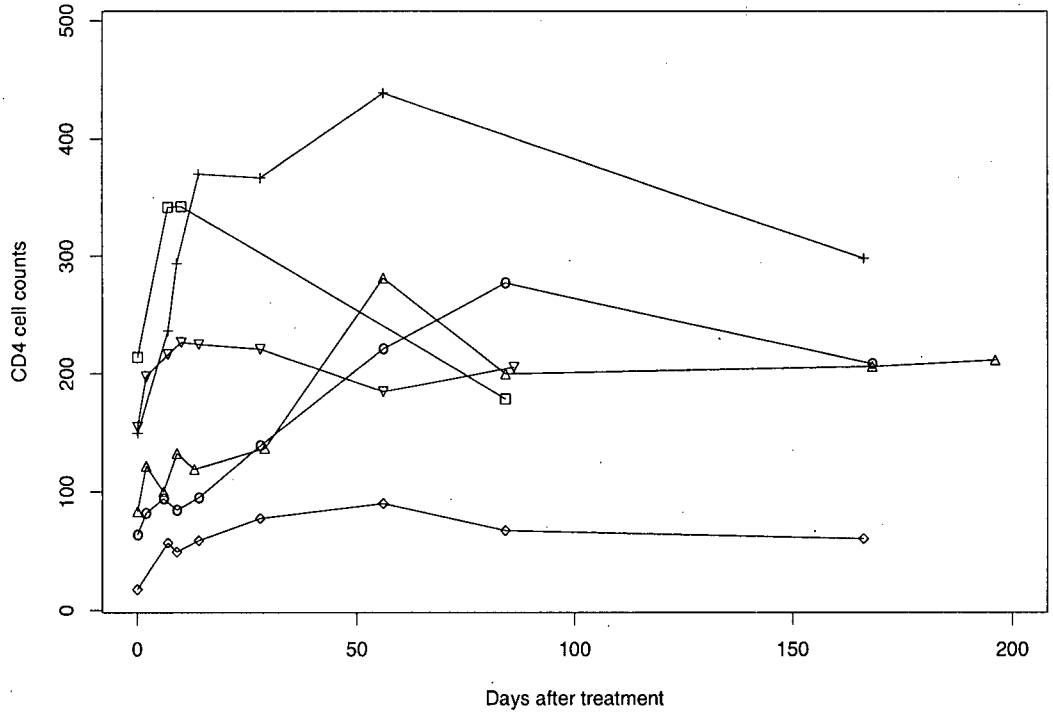

Figure 1.2: CD4 counts of six randomly selected HIV patients.

on days 0, 2, 7, 10, 14, 21, 28 and weeks 8, 12, 24, and 48 after initiation of treatments. After the antiretroviral treatment, the patients' viral loads will decay, and the decay rates may reflect the efficacy of the treatment. Throughout the time course, the viral load may continue to decay, fluctuate, or even start to rise (rebound). The data at the late stage of study are likely to be contaminated by long-term clinical factors, which leads to complex longitudinal trajectories. Various covariates such as CD4 count were also recorded throughout the study on similar schedules. It is well known that CD4 counts are usually measured with substantial errors. The number of response (viral load) measurements for each individual varies from 6 to 10. Five patients dropped out of the study due to drug intolerance or other problems and sixteen patients have missing viral loads at scheduled time points. There were 104 out of 403 CD4 measurements missing at viral load measurement times, due mainly to a somewhat different CD4 measurement schedule. Six patients are randomly selected and their viral loads are plotted in Figure 1.1.

In the presence of measurement errors in CD4 count, we consider the following semi-parametric NLME model, which corresponds model (1.6) – (1.8), to fit the viral dynamics

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{1.9}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z^*_{ij} + b_{2i}, \tag{1.10}$$

$$\log(P_{2i}) = \beta_4 + b_{3i}, \qquad \lambda_{2ij} = w(t_{ij}) + h_i(t_{ij}), \tag{1.11}$$

where z^*_{ij} is the unobservable true CD4 count, reflecting the belief that actual, not possibly corrupted, CD4 counts govern the initial phase viral decay rate λ_{1ij}. Model (1.9) – (1.11) will be used in our data analyses in later chapters. The CD4 count trajectories for six randomly selected patients are plotted in Figure 1.2. There exists large variability in CD4 count between patients. Most CD4 count trajectories appear to have roughly quadratic polynomial shapes. We will discuss covariate models in the next chapter.

1.6 Research Objectives and Thesis Organization

In this thesis, we consider approximate maximum likelihood inference in the following three problems: (1). semiparametric NLME models with measurement errors and missing data in time-varying covariates; (2). semiparametric NLME models with covariate measurement errors and *outcome-based informative* missing responses; (3). semiparametric NLME models with covariate measurement errors and *random-effect-based informative* missing responses. Measurement errors, dropouts, and missing data are addressed simultaneously in a unified way. Some asymptotic results are developed. For each problem, we propose two joint model methods to simultaneously obtain approximate maximum likelihood estimates (MLEs) of all model parameters. The first method, implemented by a Monte Carlo EM algorithm, is more accurate than the second method but it is computationally very intensive and may offer computational difficulties such as slow or non-convergence, especially when the dimensions of random effects are not small. The second method, which approximates joint log-likelihood functions, is always computationally feasible and is often computationally much more efficient, but it is usually less accurate than the first method. The second method may be used as a reasonable alternative when the first method has convergence problems or may be used to provide excellent parameter starting values for the first method.

The remainder of this thesis is organized as follows. In Chapter 2, we introduce general semiparametric NLME models with covariate measurement errors. Following Rice and Wu (2001) and Wu and Zhang (2002), we employ natural cubic spline bases with the percentile-based knots to transform semiparametric NLME models into a parametric NLME models. In Chapter 3, we address measurement errors and missing data in time-varying covariates in semiparametric NLME models and propose two joint model methods, implemented by a Monte Carlo EM algorithm and by a first-order Taylor approximation to log-likelihood func-

tions, respectively. We also compare the two joint model methods with the two-step method suggested by Higgins, et al. (1997) and discuss the asymptotic properties of approximate MLEs. We finally apply the two joint model methods to a HIV dataset. In Chapter 4, we address *outcome-based informative* dropouts and covariate measurement errors in semiparametric NLME models and propose two joint model methods, implemented by Monte Carlo EM algorithms. We illustrate our proposed methods in a HIV dataset and evaluate their performance via simulation studies. In Chapter 5, we consider *random-effect-based informative* missing responses in semiparametric NLME models with covariate measurement errors. We propose two joint model methods, implemented by a Monte Carlo EM algorithm and by a first-order Laplace approximation to log-likelihood functions respectively, to simultaneously obtain approximate MLEs of all model parameters. We also discuss some asymptotic properties of the approximate MLEs. We illustrate our methods in a HIV dataset and evaluate their performance by simulation studies. We conclude this thesis with some discussion and possible future work in Chapter 6.

Chapter 2

A Semiparametric Nonlinear Mixed-effects Model with Covariate Measurement Errors

2.1 Introduction

In this chapter we present the general form for semiparametric NLME models with covariate measurement errors. In Section 2.2, we describe a general semiparametric NLME model for the response process and incorporate possibly mis-measured time-varying covariates. We approximate the proposed semiparametric NLME model by a parametric NLME model, using linear combinations of natural cubic splines with percentile-based knots. Consistency of the estimates is discussed. In Section 2.3, the covariate process is modelled using a mixed-effects model to address measurement errors and missing data.

2.2 A Semiparametric NLME Model for the Response Process

2.2.1 A Semiparmetric NLME Model with Mis-measured Covariates

We describe a semiparametric NLME model in general form. Let y_{ij} be the response value for individual i at time t_{ij}, $i = 1, \ldots, n,\ j = 1, \ldots, n_i$. Let z_{ikl} be the observed value and let z^*_{ikl} be the unobservable "true" value of covariate k for individual i at time u_{il}, $i = 1, \ldots, n,\ k = 1, \ldots, \nu,\ l = 1, \ldots, m_i$. For simplicity, we focus on the case where z^*_{ikl} is the current true covariate value, but our method can be extended to the case where z^*_{ikl} is a summary of the true covariate values up to time u_{il}. Note that for each individual, we allow the covariate measurement times u_{il} to differ from the response measurement times t_{ij}. In other words, we allow missing data in the covariates. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ and $\mathbf{z}_i = (\mathbf{z}_{i1}^T, \ldots, \mathbf{z}_{im_i}^T)^T$, where $\mathbf{z}_{il} = (z_{i1l}, \ldots, z_{i\nu l})^T,\ l = 1, \ldots, m_i$.

For the response process, we consider a general semiparametric NLME model similar to Wu and Zhang (2002), but incorporate possibly mis-measured time-varying covariates

$$y_{ij} = g(t_{ij},\ \boldsymbol{\beta}^*_{ij},\ r_i(t_{ij})) + e_{ij}, \tag{2.1}$$

$$\boldsymbol{\beta}^*_{ij} = \mathbf{d}^*(\mathbf{z}^*_{ij},\ \boldsymbol{\beta}^*,\ \mathbf{b}^*_i), \tag{2.2}$$

$$r_i(t) = v(w(t),\ h_i(t)), \qquad i = 1, \ldots, n,\ j = 1, \ldots, n_i, \tag{2.3}$$

where $g(\cdot)$, $\mathbf{d}^*(\cdot)$, and $v(\cdot)$ are known (possible nonlinear) functions, $w(t)$ and $h_i(t)$ are unknown nonparametric smooth fixed-effects and random-effects functions respectively, $\boldsymbol{\beta}^*_{ij}$ are individual-specific parameters, $\boldsymbol{\beta}^*$ are population parameters, e_{ij} is the within-individual random error, and $\mathbf{b}^*_i$ are random effects. Let $\mathbf{e}_i = (e_{i1}, \ldots, e_{in_i})^T$. We assume that $\mathbf{e}_i \sim$

$N(0,\ \delta^2 I)$, where δ^2 is the unknown within-individual variance and I is the identity matrix, $\mathbf{b}_i^* \overset{i.i.d.}{\sim} N(\mathbf{0},\ B^*)$, $h_i(t)$'s are identical and independent realizations of a zero-mean stochastic process $h(t)$, and $\mathbf{b}_i^*$ and $h_i(t)$ are independent of $\mathbf{e}_i$. We can rewrite the semiparametric NLME models (2.1) – (2.3) in a compact way

$$y_{ij} = g(t_{ij},\ \mathbf{d}^*(\mathbf{z}_{ij}^*,\ \boldsymbol{\beta}^*,\ \mathbf{b}_i^*),\ v(w(t_{ij}),\ h_i(t_{ij}))) + e_{ij}. \tag{2.4}$$

Note that in (2.1) or (2.4), we assume that the individual-specific parameters $\boldsymbol{\beta}_{ij}^*$ depend on the true but unobservable covariates $\mathbf{z}_{ij}^*$ rather than the observed covariates $\mathbf{z}_{ij}$, which are measured with error.

Because of the nonparametric parts (i.e., $w(t)$ and $h_i(t)$) in the model, the semiparametric NLME model (2.4) is more flexible than parametric NLME models for modelling longitudinal data, and it reduces to a parametric NLME model when the nonparametric parts $w(t)$ and $h_i(t)$ are constants. Following Wu and Zhang (2002), model (2.4) is also more flexible than other semiparametric NLME models that have appeared in the literature, such as Ke and Wang (2001). The semiparametric NLME models in Ke and Wang (2001) can be considered as a special case of model (2.4). In particular, their model only put the random effects in $\boldsymbol{\beta}_{ij}^*$ as in (2.2) and considered $w_i(t_{ij}) = w(t_{ij};\ \boldsymbol{\beta}_{ij}^*)$ in (2.3). Therefore, model (2.4) is a very general and flexible semiparametric NLME model.

2.2.2 A Basis-based Approach to Nonparametric Functions

To do statistical inference for the semiparametric NLME model (2.4), a main difficulty is how to fit the nonparametric smooth fixed-effects function $w(t)$ and random-effects function $h_i(t)$. Following Rice and Wu (2001) and Wu and Zhang (2002), we use a basis-based approach which transforms a general semiparametric NLME model into a set of parametric NLME models indexed by a smoothing parameter (the number of basis functions). We use the

fixed-effects function $w(t)$ to illustrate the basis-based approach.

Let χ be the support of t and $L^2(\chi)$ be the inner product space of all square integrable functions with norm $\|\cdot\|$ and inner product $<\cdot,\cdot>$, where for any $\psi_1,\ \psi_2 \in L^2(\chi)$, we define

$$\|\psi_1\|^2 = \int_\chi \psi_1^2(t)dt, \qquad <\psi_1,\ \psi_2> = \int_\chi \psi_1(t)\psi_2(t)dt.$$

Assume that $w(t)$ is an element of a smooth function space $S_w(\chi)$, a subspace of $L^2(\chi)$. An example of $S_w(\chi)$ is the Sobolev space

$$W_2^m(\chi) = \{\psi \mid \psi, \psi', \ldots, \psi^{(m-1)} \text{ absolutely continuous, } \psi^{(m)} \in L^2(\chi)\}.$$

Denote a complete orthonormal basis of $S_w(\chi)$ by $\Psi(t) = [\psi_0(t),\ \psi_1(t), \psi_2(t), \ldots]^T$ where $\psi_0(t) \equiv 1$. Then $w(t)$ can be expanded as

$$w(t) = \sum_{k=0}^{\infty} \mu_k\, \psi_k(t),$$

where the coefficients

$$\mu_k = \int_\chi w(t)\psi_k(t)dt = <w,\ \psi_k>.$$

Let $\Psi_p(t) = [\psi_0(t),\ \psi_1(t), \ldots, \psi_{p-1}(t)]^T$ and $\boldsymbol{\mu}_p = (\mu_0, \mu_1, \ldots, \mu_p)^T$. Since $w(t)$ are square integrable, the truncations of $w(t)$ at term p

$$w_p(t) = \sum_{k=0}^{p-1} \mu_k\, \psi_k(t) = \Psi_p(t)^T\, \boldsymbol{\mu}_p,$$

will converge to $w(t)$ in L^2-norm as p tends to infinity. It follows that when p is large enough, $w_p(t)$ can approximate $w(t)$ very well, i.e., $w(t) \approx w_p(t)$.

Similarly, if we assume that $h_i(t)$ is an element of a smooth function space $S_h(\chi)(\subset L^2(\chi))$ with a complete orthonormal basis $\Phi(t) = [\phi_0(t),\ \phi_1(t), \phi_2(t), \ldots]^T$ where $\phi_0(t) \equiv 1$, the truncations of $h_i(t)$ at term q

$$h_{iq}(t) = \sum_{k=0}^{q-1} \xi_{ik}\, \phi_k(t) = \Phi_q(t)^T\, \boldsymbol{\xi}_{iq},$$

will converge to $h_i(t)$ in L^2-norm as q tends to infinity, where $\Phi_q(t) = [\phi_0(t), \phi_1(t), \ldots, \phi_{q-1}(t)]^T$ and $\boldsymbol{\xi}_{iq} = (\xi_{i0}, \ldots, \xi_{iq})^T$. It follows that when q is large enough, $h_{iq}(t)$ can approximate $h_i(t)$ very well, i.e., $h_i(t) \approx h_{iq}(t)$.

The function $w_p(t)$ can be considered as the projection of $w(t)$ on the linear space $S(\chi, \Psi_p) = \{\psi \,|\, \psi = \Psi_p(t)^T \boldsymbol{\mu}_p, \boldsymbol{\mu}_p \in R^p\} \subset S_w(\chi)$, spanned by basis functions $\Psi_p(t)$, and the function $h_{iq}(t)$ can be considered as the projection of $h_i(t)$ on the linear space $S(\chi, \Phi_q) = \{\phi \,|\, \phi = \Phi_p(t)^T \boldsymbol{\xi}_{iq}, \boldsymbol{\xi}_{iq} \in R^q\} \subset S_h(\chi)$, spanned by basis functions $\Phi_q(t)$. With p and q increasing, $w_p(t)$ and $h_{iq}(t)$ approach to $w(t)$ and $h_i(t)$, respectively. Parameters $\boldsymbol{\mu}_p$ and $\boldsymbol{\xi}_{iq}$ are unknown vectors of fixed- and random-effects coefficients, respectively. Since $h_i(t)$'s are assumed to be identical and independent realizations of a zero-mean stochastic process, we can regard $\boldsymbol{\xi}_{iq}$ as identical and independent realizations of a zero-mean random vector with unknown covariance matrix K.

There are many bases available in the literature for curve fitting. Among global bases are Legendre polynomials and Fourier series, and among local bases are regression splines (Eubank, 1988), B-splines (de Boor, 1978) and natural splines (Green and Sliverman, 1994). A *B-spline* of degree d on χ with knots $t_0 < t_1 < \cdots < t_M < t_{M+1}$ is a piecewise polynomial with polynomial pieces of degree d joining together smoothly at the interior knots $t_1 < \cdots < t_M$ while satisfying some boundary conditions. In other words, a B-spline is a polynomial of degree d within each of the intervals $[t_k, t_{k+1})$, $0 \leq k \leq M-1$, and $[x_M, x_{M+1}]$, which globally has $(d-1)$-continuous derivatives. All such B-splines form a linear space with $M + d + 1$ basis functions which are mainly determined by three factors: the degree d; the location of the knots, and the number of interior knots M. When the degree $d = 3$, the corresponding B-splines are called *cubic splines*. When the cubic splines have zero second and third derivatives at the two extreme knots x_0 and x_{M+1}, they are called *natural cubic splines*. Without loss of generality, throughout this thesis, we assume that the nonparametric

fixed- and random-effects functions $w(t)$ and $h_i(t)$ are elements of the Sobolev space $W_2^2(\chi)$ and we use natural cubic spline bases (Green and Sliverman, 1994) due to their many good properties, for example, easy construction, good smoothness, and flexibility to model the underlying curves of various shapes (de Boor, 1978).

2.2.3 Percentile-based Knot Placing for Splines

The placing of knots is an important issue for splines in which we attempt to use a few knots to represent a sample of design time points. We use sample percentiles of the design time points as knots so that there are more (fewer) knots in the area where more (fewer) design time points are available, as suggested by Wu and Zhang (2002). They indicated that the percentile-based knot placing rule should work better for longitudinal data than the equally-spaced knot placing rule used by Rice and Wu (2001), since the design time points of longitudinal data are usually sparse and often not uniformly spaced. Moreover, the percentile-based knot placing rule guarantees that the locations of the knots (and also the resulting basis functions) are sample-dependent and design-adaptive. These properties are not shared by the equally-spaced knot placing rule. After the degree d and the knot placing rule are determined, we need to choose the numbers of the interior knots, or equivalently to choose the numbers p and q of the basis functions, which are called *smoothing parameters*.

2.2.4 Selection of Smoothing Parameters

Using natural cubic spline bases with percentile-based knots to fit the nonparametric fixed- and random-effects functions $w(t)$ and $h_i(t)$, we can transform the semiparametric NLME model into a parametric NLME model. To assess how well the resulting parametric NLME model approximates the original semiparametric NLME model, we need to consider two factors: the goodness-of-fit and the model complexity. Goodness-of-fit usually indicates how

well the model fits the data (or how small the biases of the associated estimators are). It can be improved by increasing p and q or equivalently, enlarging the linear spaces $S_w(\chi, \Psi_p(t))$ and $S_h(\chi, \Phi_q(t))$. However, the model complexity represents how complex the model is (or how large the variances of the associated estimators are). The model usually becomes more complicated with increasing p and q. Thus, there is a trade-off between the goodness-of-fit and the model complexity. To balance the two components, it is natural to employ some model selection rules such as the Akaike Information Criterion (AIC) or the Schwarz's Bayesian Information Criterion (BIC) (Davidian and Giltinan, 1995). This is because the transformed parametric NLME models are indexed by p and q and choosing different p and q is equivalent to choosing different parametric NLME models.

Let φ be the number of independent parameters in a parametric NLME model, say, model (1.1) and (1.2). Then the AIC and the BIC are defined as

$$\begin{aligned} \text{AIC} &= -2\text{Loglik} + 2\varphi, \\ \text{BIC} &= -2\text{Loglik} + \left[\log\left(\sum_{i=1}^{n} n_i\right)\right]\varphi, \end{aligned}$$

where Loglik is the log-likelihood of the fitted the parametric NLME model (see Davidian and Giltinan, 1995, p156). Since a parametric NLME model with a larger number of parameters will always produce a larger value for the log-likelihood (a smaller value for –2 Loglik), the penalty terms 2φ in AIC and $\left[\log(\sum_{i=1}^{n} n_i)\right]\varphi$ in BIC are needed to offset this advantage. Since the penalty term in BIC is usually much larger than that in AIC, BIC is a conservative rule and generally favors a parsimonious model. Since both the AIC and the BIC of a parametric NLME model are defined as twice the negative log-likelihood of the model (representing the goodness-of-fit) plus a penalty term related to the number of parameters used in the model (representing the model complexity), we will choose $\Psi_p(t)$ and $\Phi_q(t)$ so that the AIC or the BIC are minimized over a series of $\Psi_p(t)$ and $\Phi_q(t)$, which leads to the best approximate parametric NLME model to the original semiparametric NLME model in terms of the AIC

and BIC criteria. Liang et al. (2003) noted that the model obtained this way often provides good approximation in practice. We will evaluate the performance of the AIC and the BIC in the current setting (see Section 4.6).

2.2.5 Transformation of the Semiparametric NLME Model

After determining the smoothing parameters p and q via the AIC and the BIC criteria, we replace $w(t)$ and $h_i(t)$ in the nonparametric function $r_i(t)$ in (2.3) by their approximations $w_p(t)$ and $h_{iq}(t)$. Thus, we obtain an approximation to the nonparametric function $r_i(t)$, and approximate the semiparametric NLME model (2.4) as follows

$$\begin{aligned} y_{ij} &\approx g(t_{ij},\, \mathbf{d}^*(\mathbf{z}^*_{ij},\, \boldsymbol{\beta}^*,\, \mathbf{b}^*_i),\, v(\Psi_p(t)^T\, \boldsymbol{\mu}_p,\, \Phi_q(t)^T\, \boldsymbol{\xi}_{iq})) + e_{ij} \\ &= g(t_{ij},\, \mathbf{d}(\mathbf{z}^*_{ij},\, \boldsymbol{\beta},\, \mathbf{b}_i)) + e_{ij} \end{aligned} \tag{2.5}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}^*,\, \boldsymbol{\mu}_p)$ are fixed effects, $\mathbf{b}_i = (\mathbf{b}^*_i,\, \boldsymbol{\xi}_{iq})$ are random effects, and $\mathbf{d}(\cdot)$ is a known but possible nonlinear function. Then, we can approximate the semiparametric NLME model (2.1) – (2.3) by the following parametric NLME model

$$y_{ij} = g(t_{ij},\, \boldsymbol{\beta}_{ij}) + e_{ij}, \qquad \mathbf{e}_i|\boldsymbol{\beta}_i \overset{i.i.d.}{\sim} N(0,\, \delta^2\, I), \tag{2.6}$$

$$\boldsymbol{\beta}_{ij} = \mathbf{d}(\mathbf{z}^*_{ij},\, \boldsymbol{\beta},\, \mathbf{b}_i), \qquad \mathbf{b}_i \overset{i.i.d.}{\sim} N(0, B), \tag{2.7}$$

where B is an unstructured covariance matrix. Note that $\mathbf{e}_i$ and $\mathbf{b}_i$ are independent of each other. Approximate statistical inference can then be based on the approximate model (2.6) and (2.7), as shown in Chapters 3 – 5.

2.2.6 Consistency of the Estimate of $w(t)$

After we obtain estimates $\hat{\boldsymbol{\mu}}_p$ and $\hat{\boldsymbol{\xi}}_{iq}$ based on the parametric NLME model (2.6) and (2.7), we can then estimate the nonparametric functions $w(t)$ and $h_i(t)$ in the semiparametric

NLME model (2.1) – (2.3) as follows

$$\begin{aligned} \hat{w}(t) &= \hat{w}_p(t) = \Psi_p(t)^T \hat{\boldsymbol{\mu}}_p, \\ \hat{h}_i(t) &= \hat{h}_{iq}(t) = \Phi_q(t)^T \hat{\boldsymbol{\xi}}_{iq}, \end{aligned}$$

Therefore, the consistency of the estimates in the semiparametric NLME model (2.1) – (2.3) is strongly related to the consistency of estimates in the parametric NLME model (2.6) and (2.7). Under some mild conditions, the following Theorem 2.1 guarantees that we can obtain a consistent estimate $\hat{w}(t)$ of the nonparametric fixed-effects function $w(t)$ in the semiparametric NLME model (2.1) – (2.3) if we can find $\sqrt{n}$-consistent estimates $\hat{\boldsymbol{\mu}}_p$ of the fixed-effects coefficients $\boldsymbol{\mu}_p$.

Following Wu and Zhang (2002), we prove the consistency of the estimate $\hat{w}(t)$ of the nonparametric fixed-effects function $w(t)$ in the semiparametric NLME model (2.1) – (2.3) based on the following conditions:

(a). $\Psi(t)$ is a complete orthonormal basis of $S(\chi)$, a subspace of $L^2(\chi)$.

(b). The nonparametric fixed-effects function $w(t) \in S(\chi)$ so that $w(t) = \sum_{k=0}^{\infty} \mu_k \psi_k(t)$.

(c). The design time points $\{t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$ are identically and independently distributed such that when the number n of individuals tends to infinity, the number of distinct time points will tend to infinity. In this case, we can truncate $w(t)$ in the semiparametric NLME model (2.1) – (2.3) in such a way that $w_p(t) = \sum_{k=0}^{p-1} \mu_k \psi_k(t) = \Psi_p(t)^T \boldsymbol{\mu}_p$ so that $p \to \infty$, $p/n \to 0$ as $n \to \infty$.

(d). For any fixed p, we assume that we can obtain $\sqrt{n}$-consistent estimates $\hat{\boldsymbol{\mu}}_p$ of the fixed-effects coefficients $\boldsymbol{\mu}_p$ so that as $n \to \infty$, $E(\hat{\boldsymbol{\mu}}_p) \to \boldsymbol{\mu}_p$ and $\mathrm{Cov}(\sqrt{n}\hat{\boldsymbol{\mu}}_p) \to \Sigma_p$ for some semidefinite positive matrix Σ_p with $p^{-1}\mathrm{tr}(\Sigma_p)$ bounded.

Theorem 2.1. Under Conditions (a) – (d), as $n \to \infty$, we have $\|\hat{w} - w\| \to 0$ in probability.

Proof. First we consider $E\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2$.

$$
\begin{aligned}
E\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2 &= E\left[\sum_{k=0}^{p-1}(\hat{\mu}_k - \mu_k)^2\right] = \sum_{k=1}^{p-1} E(\hat{\mu}_k - \mu_k)^2 \\
&= \sum_{k=1}^{p-1} E[\hat{\mu}_k - E(\hat{\mu}_k) + E(\hat{\mu}_k) - \mu_k]^2 \\
&= \sum_{k=1}^{p-1} E\{[\hat{\mu}_k - E(\hat{\mu}_k)]^2 + [E(\hat{\mu}_k) - \mu_k]^2 + 2[\hat{\mu}_k - E(\hat{\mu}_k)][E(\hat{\mu}_k) - \mu_k]\} \\
&= \sum_{k=1}^{p-1} \{E[\hat{\mu}_k - E(\hat{\mu}_k)]^2 + [E(\hat{\mu}_k) - \mu_k]^2\} \\
&= \sum_{k=1}^{p-1} \mathrm{Var}(\hat{\mu}_k) + \sum_{k=1}^{p-1} [E(\hat{\mu}_k) - \mu_k]^2 \\
&= \mathrm{tr}\,[\mathrm{Cov}(\hat{\boldsymbol{\mu}}_p)] + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2.
\end{aligned}
$$

Under Conditions (a) and (b), and $\mathrm{Cov}(\sqrt{n}\hat{\boldsymbol{\mu}}_p) \to \Sigma_p$ in Condition (d), we have

$$
\begin{aligned}
E\|\hat{w} - w\|^2 = E\|\hat{w}_p - w\|^2 &\le 2\{E\|\hat{w}_p - w_p\|^2 + \|w_p - w\|^2\} \\
&= 2\{E\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2 + \|w_p - w\|^2\} \\
&= 2\{\mathrm{tr}[\mathrm{Cov}(\hat{\boldsymbol{\mu}}_p)] + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2 + \|w_p - w\|^2\} \\
&= 2\{n^{-1}\mathrm{tr}[\mathrm{Cov}(\sqrt{n}\hat{\boldsymbol{\mu}}_p)] + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2 + \|w_p - w\|^2\} \\
&= 2\left\{n^{-1}\mathrm{tr}[\Sigma_p + o(1)] + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2 + \sum_{k=p}^{\infty}\mu_k^2\right\} \\
&= 2\left\{n^{-1}[\mathrm{tr}(\Sigma_p) + p\,o(1)] + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2 + \sum_{k=p}^{\infty}\mu_k^2\right\} \\
&= 2\left\{\frac{p}{n}\cdot\frac{\mathrm{tr}[\Sigma_p]}{p} + \|E(\hat{\boldsymbol{\mu}}_p) - \boldsymbol{\mu}_p\|^2 + \sum_{k=p}^{\infty}\mu_k^2\right\} + o\left(\frac{p}{n}\right).
\end{aligned}
$$

Under Conditions (b)–(d), it is easy to show that the three terms in parentheses $\{\cdot\}$ of the right-hand side tend to 0 as $n \to \infty$. Under Condition (d), as $n \to \infty$, $E\|\hat{w} - w\|^2 \to 0$ implies $\|\hat{w} - w\| \to 0$ in probability. □

2.3 Measurement Errors and Missing Data in Covariates

At the presence of measurement errors and missing data in the time-varying covariates $\mathbf{z}_{il} = (z_{i1l}, \ldots, z_{ivl})^T$, we need to model the covariate processes. We consider the following multivariate linear mixed-effects (LME) model (Shah et al., 1997) to empirically describe the covariate process

$$\mathbf{z}_{il} = U_{il}\,\boldsymbol{\alpha} + V_{il}\,\mathbf{a}_i + \boldsymbol{\epsilon}_{il} \;\; (\equiv \mathbf{z}^*_{il} + \boldsymbol{\epsilon}_{il}), \qquad i = 1, \ldots, n,\; l = 1, \ldots, m_i, \tag{2.8}$$

where U_{il} and V_{il} are design matrices, $\boldsymbol{\alpha}$ and $\mathbf{a}_i$ are unknown population (fixed-effects) and individual-specific (random-effects) parameter vectors, and $\boldsymbol{\epsilon}_{il}$ are the random measurement errors for individual i at time u_{il}. For example, we may model the covariate processes parametrically based on empirical polynomial models with random coefficients, as in Higgins et al. (1997) and Wu (2002). Alternatively, we may model the covariate processes nonparametrically, and approximate the nonparametric fixed- and random-effects functions by linear combination of some basis functions, as in Section 2.2. For either parametric or nonparametric covariate models, we may convert the covariate models to the LME model (2.8). Note that the covariate model (2.8) incorporates both the correlation of the repeated measurements on each individual and the correlation among different covariates.

Note that the parameters in the covariate model (2.8) may be viewed as nuisance parameters because they are often not of main interest. We assume that the true (unobservable) covariate values are

$$\mathbf{z}^*_{il} = U_{il}\,\boldsymbol{\alpha} + V_{il}\,\mathbf{a}_i.$$

We also assume that $\mathbf{a}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, A)$, $\boldsymbol{\epsilon}_{il} \overset{i.i.d.}{\sim} N(0, R)$, and $\mathbf{a}_i$ and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}^T_{i1}, \ldots, \boldsymbol{\epsilon}^T_{im_i})^T$ are independent, where A is an unrestricted covariance matrix and R is an unknown within-

individual covariance matrix. We further assume that $\boldsymbol{\epsilon}_i$ and $\mathbf{a}_i$ are independent of $\mathbf{e}_i$ and $\mathbf{b}_i$. Models (2.8) may be interpreted as a covariate measurement error model (Carroll et al., 1995; Higgins et al., 1997).

To allow for missing data in the time-varying covariates (or different measurement schedules for the time-varying covariates), we recast model (2.8) in continuous time:

$$\mathbf{z}_i(t) = U_i(t)\,\boldsymbol{\alpha} + V_i(t)\,\mathbf{a}_i + \boldsymbol{\epsilon}_i(t), \quad i = 1, \ldots, n,$$

where $\mathbf{z}_i(t)$, $U_i(t)$, $V_i(t)$, and $\boldsymbol{\epsilon}_i(t)$ are the covariate values, design matrices, and measurement errors at time t respectively. At the response measurement time t_{ij}, which may be different from the covariate measurement times u_{il}, the possibly unobserved "true" covariate values can be viewed as $\mathbf{z}^*_{ij} = U_{ij}\,\boldsymbol{\alpha} + V_{ij}\,\mathbf{a}_i$, where $U_{ij} = U_i(t_{ij})$ and $V_{ij} = V_i(t_{ij})$. In other words, missing covariates at time t_{ij} may be imputed by their estimated true values $\mathbf{z}^*_{ij}$.

Chapter 3

A Joint Model for Semiparametric NLME Models with Covariate Measurement Errors and Missing Data

3.1 Introduction

In this chapter, we address measurement errors and missing data in time-varying covariates for semiparametric NLME models. In Section 3.2, we review the two-step method proposed by Higgins et al. (1997). We derive some analytic and asymptotic results for the two-step estimates for parametric NLME models with mis-measured covariates, and analytically show that the variances of the main parameter estimates based on the two-step method are underestimated.

To address measurement errors and missing data in time-varying covariates in semi-

parametric NLME models based on models (2.6) – (2.8), in Sections 3.3 and 3.4 we propose two joint model methods, implemented by a Monte Carlo EM algorithm and by a first-order Taylor approximation to the log-likelihood function respectively, to find approximate MLEs of model parameters. We also discuss asymptotic properties of these approximate MLEs.

In Section 3.5, we apply the two joint model methods to a real dataset. We evaluate the proposed methods and compare them with the two-step method via simulation studies. We conclude this chapter with some discussion in Section 3.6. Proofs of the asymptotic properties of approximate MLEs are presented in Section 3.7.

3.2 A Two-step Method

For covariate measurement error problems, a commonly used method is the so-called *two-step method* (Higgins, et al., 1997; Liang, et al., 2003): in the first step the "true" covariate values are estimated based on an assumed covariate model, and then in the second step, the possibly mis-measured covariates in the response model are simply replaced by the estimated covariates from the first step. The estimation of the main parameters in the response model proceeds as if the estimated covariate values are the true covariate values without measurement error. Intuitively, the resulting estimates of the main parameters may be approximately unbiased if the covariate estimates from the first step are unbiased, but the variances of the main parameter estimates may be underestimated because the variability of the covariate estimation in the first step is ignored in the estimation of the main parameters in the second step. Higgins et al. (1997) and Ogden and Tarpey (2005) realized this problem and proposed bootstrap methods which incorporate the variability from estimating the covariates in the first step. Wu (2002) considered an approximate joint model approach which also incorporates the variability in the covariate estimation. In this section, we derive some

analytic results for the two-step estimates for parametric NLME models with mis-measured covariates such as the models (2.6) – (2.8), analytically show that the variances of the main parameter estimates based on the two-step method are underestimated, and derive some asymptotic results for the two-step estimates.

Let $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)^T$, and define $\mathbf{z}$, $\mathbf{z}^*$, and $\boldsymbol{\epsilon}$ similarly. If $\mathbf{z}^*$ is known, an estimate $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}$ can be expressed as $\hat{\boldsymbol{\beta}}^* = \mathbf{s}(\mathbf{y}, \mathbf{z}^*)$, where $\mathbf{s}$ is a vector function. Since $\mathbf{z}$ is recorded with errors and $\mathbf{z}^*$ is unobservable, we assume $\mathbf{z} = \mathbf{z}^* + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Let $\hat{\mathbf{z}}$ be an unbiased estimate of the true covariate value $\mathbf{z}^*$ based on the observed covariate value $\mathbf{z}$ (i.e., $E(\hat{\mathbf{z}}) = \mathbf{z}^*$). The two-step method estimates $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = \mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})$, which depends on realizations of two random variables $\mathbf{y}$ and $\hat{\mathbf{z}}$. If we assume that $E[\mathbf{s}(\mathbf{y}, \mathbf{z}^*)] \approx \boldsymbol{\beta}$, and $\mathbf{y}$ and $\mathbf{z}$ are roughly independent so that

$$E\Big\{\Big[\frac{\partial \mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})}{\partial \hat{\mathbf{z}}}\,|_{\hat{\mathbf{z}}=\mathbf{z}^*}\Big](\hat{\mathbf{z}} - \mathbf{z}^*)\Big\} \approx E\Big[\frac{\partial \mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})}{\partial \hat{\mathbf{z}}}\,|_{\hat{\mathbf{z}}=\mathbf{z}^*}\Big]\, E(\hat{\mathbf{z}} - \mathbf{z}^*),$$

and that the function $\mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})$ is well approximated by a first-order Taylor series expansion around $\mathbf{z}^*$, then we show next that the estimate $\hat{\boldsymbol{\beta}}$ is also approximately unbiased, following Ogden and Tarpey (2005). Taking a first-order Taylor expansion of $\mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})$ around the "true" covariate value $\mathbf{z}^*$, we have

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E(\mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})) \approx E\Big\{\mathbf{s}(\mathbf{y}, \mathbf{z}^*) + \Big[\tfrac{\partial \mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})}{\partial \hat{\mathbf{z}}}\,|_{\hat{\mathbf{z}}=\mathbf{z}^*}\Big](\hat{\mathbf{z}} - \mathbf{z}^*)\Big\} \\ &\approx \boldsymbol{\beta} + E\Big[\tfrac{\partial \mathbf{s}(\mathbf{y}, \hat{\mathbf{z}})}{\partial \hat{\mathbf{z}}}\,|_{\hat{\mathbf{z}}=\mathbf{z}^*}\Big]\, E(\hat{\mathbf{z}} - \mathbf{z}^*) = \boldsymbol{\beta}, \end{aligned}$$

provided that the expectations exist (the expectation operator in the above expression is to be taken as the expectation with respect to both $\mathbf{y}$ and $\hat{\mathbf{z}}$). However, the variances of the two-step estimates will be underestimated, as shown below. When $\hat{\mathbf{z}}$ is plugged in the estimation of $\boldsymbol{\beta}$, the resulting variance-covariance matrix is actually an estimate of $\mathrm{Cov}(\hat{\boldsymbol{\beta}}|\hat{\mathbf{z}})$. By the well-known variance decomposition formula

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathrm{Cov}[E(\hat{\boldsymbol{\beta}}|\hat{\mathbf{z}})] + E[\mathrm{Cov}(\hat{\boldsymbol{\beta}}|\hat{\mathbf{z}})],$$

we know that $\text{Cov}(\hat{\boldsymbol{\beta}}) - E[\text{Cov}(\hat{\boldsymbol{\beta}}|\hat{\mathbf{z}})]$ is nonnegative definite, since $\text{Cov}[\text{E}(\hat{\boldsymbol{\beta}}|\hat{\mathbf{z}})]$ is nonnegative definite. Thus, on average, the two-step approach underestimates the true variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

Following the general approach taken by Amemiya (1983), under the suitable regularity conditions on the log-likelihood function of $\boldsymbol{\beta}$ (e.g., the third order derivatives exist and are continuous in an open neighborhood about $\boldsymbol{\beta}$), we derive the asymptotic distribution for the MLE of $\boldsymbol{\beta}$ when the unobservable "true" covariates $\mathbf{z}_i^*$ are imputed by their estimated values $\hat{\mathbf{z}}_i$. Suppose that $\hat{\mathbf{z}}_i$ are $\sqrt{m_i}$-consistent estimates of $\mathbf{z}_i^*$ and that

$$\sqrt{m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*) \xrightarrow{d} N(\mathbf{0}, \Omega_{\mathbf{z}_i}), \qquad i = 1, \ldots, n,$$

where m_i is the number of observations for covariates on individual i. We assume that $m_i = O(m)$ uniformly for $i = 1, \ldots, n$, where $m = \min_i(m_i)$, and that m_i and n go to infinity at the same rate with $n/m_i \to c_i$, where $0 < c_i < \infty$. Let

$$l(\boldsymbol{\beta}; \mathbf{y}, \hat{\mathbf{z}}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}; \mathbf{y}_i, \hat{\mathbf{z}}_i)$$

be the log-likelihood function of $\boldsymbol{\beta}$ based on the observed data $\mathbf{y}$ and $\mathbf{z}^*$ with the unobservable "true" covariates $\mathbf{z}^*$ are imputed by their estimated values $\hat{\mathbf{z}}$. The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfies a set of equations

$$\frac{\partial l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \hat{\mathbf{z}})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \hat{\mathbf{z}}_i)}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

where

$$\frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \hat{\mathbf{z}}_i)}{\partial \boldsymbol{\beta}} = \frac{\partial l_i(\boldsymbol{\beta}; \mathbf{y}_i, \hat{\mathbf{z}}_i)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

Taking a first-order Taylor expansion of $\partial l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \hat{\mathbf{z}})/\partial \boldsymbol{\beta}$ around the "true" covariates $\mathbf{z}^*$, we

have

$$
\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \Big[\frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O(\|\hat{\mathbf{z}}_i - \mathbf{z}_i^*\|^2) \Big] \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \Big[\frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p\Big(\frac{1}{m_i}\Big) \Big] \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + \sum_{i=1}^{n} O_p\Big(\frac{1}{m_i}\Big) \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p\Big[\max_i\Big(\frac{1}{m_i}\Big)\Big] \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p\Big[\Big(\min_i m_i\Big)^{-1}\Big] \\
&= \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p(m^{-1}),
\end{aligned}
$$

since $\hat{\mathbf{z}}_i$ are $\sqrt{m_i}$-consistent estimates of $\mathbf{z}_i^*$. Next, carrying out a first-order Taylor expansion of $\sum_{i=1}^{n} \partial l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)/\partial \boldsymbol{\beta}$ in the above expression around $\boldsymbol{\beta}$, we can obtain

$$
\begin{aligned}
& \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + O(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) \\
& \qquad + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p(m^{-1}) = \mathbf{0} \\
\Longleftrightarrow \quad & \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + O_p(n^{-1}) \\
& \qquad + \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} (\hat{\mathbf{z}}_i - \mathbf{z}_i^*) + O_p(m^{-1}) = \mathbf{0},
\end{aligned}
$$

since $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and thus it is $\sqrt{n}$-consistent under the necessary regularity conditions on the log-likelihood function of $\boldsymbol{\beta}$. The above expression can be written as

$$
\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \Big[-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big]^{-1} \Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} \\
& \qquad + \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \times \sqrt{n}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*) \Big] + \Big(1 + \frac{n}{m}\Big) O_p(n^{-3/2}) \Big\}.
\end{aligned}
$$

Assume that the following limits exist

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} c_i \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \Omega_{\mathbf{z}_i} \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \hat{\mathbf{z}}_i \partial \boldsymbol{\beta}^T} = \Omega_{\mathbf{z}},$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_i(\boldsymbol{\beta}) = \bar{I}(\boldsymbol{\beta}),$$

where $I_i(\boldsymbol{\beta}) = -E[\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T]$ is the Fisher information matrix for individual i. It follows from Lemma 3.2 in Section 3.7 that

$$\frac{1}{n^2} \sum_{i=1}^{n} c_i \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \Omega_{\mathbf{z}_i} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \hat{\mathbf{z}}_i \partial \boldsymbol{\beta}^T} \xrightarrow{p} \Omega_{\mathbf{z}}.$$

Note that $\sqrt{n}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*) \approx \sqrt{c_i m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*)$ for large n and m_i. Using the asymptotic normality of $\hat{\mathbf{z}}_i$, we know that for large m_i, $E[\sqrt{m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*)] \approx \mathbf{0}$ and $\mathrm{Cov}[\sqrt{m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*)] \approx \Omega_{\mathbf{z}_i}$. By Lindeberg's central limit theorem, we have

$$\Big[\frac{1}{n^2} \sum_{i=1}^{n} c_i \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \Omega_{\mathbf{z}_i} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \hat{\mathbf{z}}_i \partial \boldsymbol{\beta}^T}\Big]^{-1/2} \Big[\frac{1}{n} \sum_{i=1}^{n} \sqrt{c_i} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \times \sqrt{m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*)\Big] \xrightarrow{d} N(\mathbf{0}, I).$$

It follows from Slutsky's theorem that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \times \sqrt{n}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*) \approx \frac{1}{n} \sum_{i=1}^{n} \sqrt{c_i} \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \hat{\mathbf{z}}_i^T} \times \sqrt{m_i}(\hat{\mathbf{z}}_i - \mathbf{z}_i^*) \xrightarrow{d} N(\mathbf{0}, \Omega_{\mathbf{z}}).$$

Under the necessary regularity conditions on the log-likelihood function of $\boldsymbol{\beta}$, based on the standard arguments for showing asymptotic normality of MLEs, we can obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta}} \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\beta})),$$

and

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{y}_i, \mathbf{z}_i^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \xrightarrow{p} \bar{I}(\boldsymbol{\beta}).$$

Since $\mathbf{y}$ and $\mathbf{z}$ are roughly independent, the two limit random variables with distributions $N(\mathbf{0}, \Omega_{\mathbf{z}})$ and $N(\mathbf{0}, \bar{I}(\boldsymbol{\beta}))$ are roughly independent. Note that $n/m = O(1)$. Putting these pieces together, we have asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ as follows:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \bar{I}(\boldsymbol{\beta})^{-1} + \bar{I}(\boldsymbol{\beta})^{-1} \Omega_{\mathbf{z}} \bar{I}(\boldsymbol{\beta})^{-1}]. \tag{3.1}$$

Note that the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ based on the asymptotic distribution (3.1) has two components. The first component $\bar{I}(\boldsymbol{\beta})^{-1}$ is the "naive" estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ in the two-step method, in which the variability of $\hat{\mathbf{z}}_i$ is neglected. The second component $\bar{I}(\boldsymbol{\beta})^{-1}\,\Omega_{\mathbf{z}}\,\bar{I}(\boldsymbol{\beta})^{-1}$ arises from the variability in estimating $\hat{\mathbf{z}}_i$ and summarizes the extra uncertainty of $\hat{\boldsymbol{\beta}}$ due to estimation of $\mathbf{z}_i$. Since the second component $\bar{I}(\boldsymbol{\beta})^{-1}\,\Omega_{\mathbf{z}}\,\bar{I}(\boldsymbol{\beta})^{-1}$ is nonnegative definite, the variances of the main parameter estimates based on the two-step method are underestimated.

3.3 A Joint Model Method for Likelihood Inference

3.3.1 The Likelihood for the Joint Model

We consider likelihood inference for semiparametric NLME models with measurement errors and missing data in time-varying covariates, based on the approximate parametric NLME models (2.6) – (2.8). The observed data are $\{(\mathbf{y}_i,\ \mathbf{z}_i),\ i = 1,\ldots,n\}$. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2,\ R,\ A,\ B)$ be the collection of all unknown parameters in models (2.6) – (2.8). We assume that the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, δ^2, R, A, and B are distinct. Let $f(\cdot)$ be a generic density function, and let $[X|Y]$ denote a conditional distribution of X given Y. The approximate log-likelihood for the observed data $\{(\mathbf{y}_i,\ \mathbf{z}_i),\ i = 1,\ldots,n\}$ can be written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\int \int f_Y(\mathbf{y}_i|\mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2)\, f_Z(\mathbf{z}_i|\mathbf{a}_i;\ \boldsymbol{\alpha},\ R)\, f(\mathbf{a}_i;\ A)\, f(\mathbf{b}_i;\ B)\, d\mathbf{a}_i\, d\mathbf{b}_i \right],$$

where

$$
\begin{aligned}
f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) &= \textstyle\prod_{j=1}^{n_i} f_Y(y_{ij}|\mathbf{z}_{ij},\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \\
&= \textstyle\prod_{j=1}^{n_i} (2\pi\delta^2)^{-1/2} \exp\{-[y_{ij} - g(t_{ij},\, d(\mathbf{u}_{ij}^T\boldsymbol{\alpha} + \mathbf{v}_{ij}^T\mathbf{a}_i,\, \boldsymbol{\beta},\, \mathbf{b}_i))]^2/2\delta^2\}, \\
f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) &= \textstyle\prod_{k=1}^{m_i} f_Z(\mathbf{z}_{ik}|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
&= \textstyle\prod_{k=1}^{m_i} |2\pi R|^{-1/2} \exp\{-(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)^T\, R^{-1} \\
&\qquad \times(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)/2\}, \\
f(\mathbf{a}_i;\, A) &= |2\pi A|^{-1/2} \exp\{-\mathbf{a}_i^T A^{-1}\mathbf{a}_i/2\}, \\
f(\mathbf{b}_i;\, B) &= |2\pi B|^{-1/2} \exp\{-\mathbf{b}_i^T B^{-1}\mathbf{b}_i/2\}.
\end{aligned}
$$

This approximate log-likelihood function generally does not have a closed-form expression since the functions in the integral can be nonlinear in the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$. Exact likelihood calculations therefore require numerical evaluation of an integral whose dimension is equal to the dimension of the random effects $(\mathbf{a}_i,\, \mathbf{b}_i)$. This is straightforward to do by direct numerical integration such as Gaussian quadrature when the dimension of $(\mathbf{a}_i,\, \mathbf{b}_i)$ is very small (say, 1 or 2). However, when $(\mathbf{a}_i,\, \mathbf{b}_i)$ has a dimension of 3 or more as is often the case in practice, one needs to consider alternative methods such as computationally intensive Monte Carlo methods.

Laird and Ware (1982) obtained MLEs in LME models using the EM algorithm. Here we use a Monte Carlo EM (MCEM) algorithm to find the approximate MLEs of all parameters $\boldsymbol{\theta}$. By treating the unobservable random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ as additional "missing" data, we have "complete data" $\{(\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i),\, i = 1, \ldots, n\}$. The complete-data log-likelihood function for all individuals can be expressed as

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} l_c^{(i)}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \{\log f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
&\qquad + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B)\},
\end{aligned}
\tag{3.2}
$$

where $l_c^{(i)}$ is the complete-data log-likelihood for individual i.

3.3.2 A MCEM Method

The EM algorithm (Dempster, Laird, and Rubin, 1977) is a very useful and powerful algorithm to compute MLEs in a wide variety of situations, such as missing data and random-effects models. The EM algorithm iterates between an E-step, which computes the conditional expectation of the complete-data log-likelihood given the observed data and previous parameter estimates, and a M-step, which maximizes this conditional expectation to update parameter estimates. The computation iterates between the E-step and M-step until convergence leads to the MLEs (or local maximizers). When there are several modes in the conditional expectation, the MLEs can be determined by trying different parameter starting values. For our models, the E-step is quite intractable due to nonlinearity, so we use Monte Carlo methods to approximate the intractable conditional expectations. In the M-step, we use standard complete-data optimization procedures to update parameter estimates.

Let $\boldsymbol{\theta}^{(t)}$ be the parameter estimates from the t-th EM iteration. The E-step for individual i at the $(t+1)$th EM iteration can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E(l_c^{(i)}(\boldsymbol{\theta})|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= \int\int \Big[\log f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\ &\qquad + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B)\Big] \times f(\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)})\, d\mathbf{a}_i\, d\mathbf{b}_i \\ &\equiv I_1^{(i)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + I_2^{(i)}(\boldsymbol{\alpha},\, R) + I_3^{(i)}(A) + I_4^{(i)}(B). \end{aligned} \tag{3.3}$$

The above integral generally does not have a closed form, and evaluation of the integral by numerical quadrature is usually infeasible, except for simple cases. However, note that expression (3.3) is an expectation with respect to the conditional distribution $f(\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)})$, and it may be evaluated using the MCEM algorithm of Wei and Tanner (1990), as in Ibrahim et al. (1999, 2001). Specifically, for individual i, let $\{(\tilde{\mathbf{a}}_i^{(1)},\, \tilde{\mathbf{b}}_i^{(1)}), \ldots, (\tilde{\mathbf{a}}_i^{(k_t)},\, \tilde{\mathbf{b}}_i^{(k_t)})\}$ denote a random sample of size k_t generated from $[\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}]$. Note that each $(\tilde{\mathbf{a}}_i^{(k)},\, \tilde{\mathbf{b}}_i^{(k)})$ de-

pends on the EM iteration number t, which is suppressed throughout. Then we approximate the conditional expectation $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in the E-step by its empirical mean, with missing data replaced by simulated values, as follows

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &\approx \left\{ \tfrac{1}{k_t} \sum_{k=1}^{k_t} l_c^{(i)}(\boldsymbol{\theta};\, \mathbf{y}_i,\, \mathbf{z}_i,\, \tilde{\mathbf{a}}_i^{(k)},\, \tilde{\mathbf{b}}_i^{(k)}) \right\} \\ &= \tfrac{1}{k_t} \sum_{k=1}^{k_t} \log\, f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \tilde{\mathbf{a}}_i^{(k)},\, \tilde{\mathbf{b}}_i^{(k)};\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \tfrac{1}{k_t} \sum_{k=1}^{k_t} \log\, f_Z(\mathbf{z}_i|\tilde{\mathbf{a}}_i^{(k)};\, \boldsymbol{\alpha},\, R) \\ &\quad + \tfrac{1}{k_t} \sum_{k=1}^{k_t} \log\, f(\tilde{\mathbf{a}}_i^{(k)};\, A) + \tfrac{1}{k_t} \sum_{k=1}^{k_t} \log\, f(\tilde{\mathbf{b}}_i^{(k)};\, B). \end{aligned}$$

We may choose k_0 as a large number and $k_t = k_{t-1} + k_{t-1}/c$, $t = 1, 2, 3, \ldots$, for some positive constant c, in the t-th iteration. Increasing k_t with each EM iteration may speed up the EM convergence (Booth and Hobert, 1999). The E-step at the $(t+1)$th EM iteration can then be expressed as

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{i=1}^{n} \left\{ \tfrac{1}{k_t} \sum_{k=1}^{k_t} l_c^{(i)}(\boldsymbol{\theta};\, \mathbf{y}_i,\, \mathbf{z}_i,\, \tilde{\mathbf{a}}_i^{(k)},\, \tilde{\mathbf{b}}_i^{(k)}) \right\} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{k_t} \tfrac{1}{k_t} \log\, f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \tilde{\mathbf{a}}_i^{(k)},\, \tilde{\mathbf{b}}_i^{(k)};\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \tfrac{1}{k_t} \log\, f_Z(\mathbf{z}_i|\tilde{\mathbf{a}}_i^{(k)};\, \boldsymbol{\alpha},\, R) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \tfrac{1}{k_t} \log\, f(\tilde{\mathbf{a}}_i^{(k)};\, A) + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \tfrac{1}{k_t} \log\, f(\tilde{\mathbf{b}}_i^{(k)};\, B) \\ &\equiv Q^{(1)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2|\boldsymbol{\theta}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha},\, R|\boldsymbol{\theta}^{(t)}) + Q^{(3)}(A|\boldsymbol{\theta}^{(t)}) + Q^{(4)}(B|\boldsymbol{\theta}^{(t)}). \end{aligned}$$

To generate independent samples from $[\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}]$, we use the Gibbs sampler (Gelfand and Smith, 1990) by sampling from the two full conditionals $[\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}]$ and $[\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}]$ as follows.

$$\begin{aligned} f(\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f(\mathbf{a}_i,\, \mathbf{y}_i|\mathbf{z}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) = f(\mathbf{a}_i|\mathbf{z}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{a}_i|\mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}), \qquad (3.4) \\ f(\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f(\mathbf{b}_i,\, \mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) = f(\mathbf{b}_i|\mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}), \end{aligned}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are independent each other. Monte Carlo samples from the above full conditionals can be generated using rejection sampling methods, as in Wu (2004). Alternatively,

integral (3.3) may be evaluated using the importance sampling method. Details of these sampling methods and convergence issues will be investigated in the next section.

The M-step then maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to produce an updated estimate $\boldsymbol{\theta}^{(t+1)}$ at the $(t+1)$-th iteration. Note that the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, δ^2, R, A, and B are all different, so we can update the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, R)$, A, and B by maximizing $Q^{(1)} + Q^{(2)}$, $Q^{(3)}$, and $Q^{(4)}$ separately in the M-step.

The maximizer $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \delta^{2(t+1)}, R^{(t+1)})$ for $Q^{(1)} + Q^{(2)}$ may be computed via iteratively re-weighted least squares where the random effects are replaced by their simulated values $\{(\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)})\}$:

$$\begin{aligned}(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \delta^{2(t+1)}, R^{(t+1)}) &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2,R} \{Q^{(1)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2|\boldsymbol{\theta}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha}, R|\boldsymbol{\theta}^{(t)})\} \\ &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2,R} \left\{ \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f_Y(\mathbf{y}_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) \right. \\ &\qquad \left. + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f_Z(\mathbf{z}_i|\tilde{\mathbf{a}}_i^{(k)}; \boldsymbol{\alpha}, R) \right\}. \qquad (3.5)\end{aligned}$$

In general, the function in (3.5) is nonlinear in parameters and thus, the maximizers have no closed-form expressions. The maximizers could be obtained via standard optimization procedures for complete-data nonlinear models, such as the Newton-Raphson method. Note that optimization procedures for nonlinear models may be iterative as well.

We can use the following Lemma to obtain analytic expressions of the maximizer $A^{(t+1)}$ for $Q^{(3)}$ and the maximizer $B^{(t+1)}$ for $Q^{(4)}$.

Lemma 3.1. (Seber, 1984). Consider the matrix function

$$h(\Sigma) = \log|\Sigma| + \mathrm{tr}[\Sigma^{-1}\Omega].$$

If Ω is positive definite, then, subject to positive definite Σ, $h(\Sigma)$ is minimized uniquely at $\Sigma = \Omega$.

By Lemma 3.1, the maximizer $A^{(t+1)}$ for $Q^{(3)}$ can be written as

$$\begin{aligned}
A^{(t+1)} &= \arg\max_A \{Q^{(3)}(A|\boldsymbol{\theta}^{(t)})\} \\
&= \arg\max_A \sum_{i=1}^{n}\sum_{k=1}^{k_t} \frac{1}{k_t} \log f(\tilde{\mathbf{a}}_i^{(k)}; A) \\
&= \arg\max_A \sum_{i=1}^{n}\sum_{k=1}^{k_t} \frac{1}{k_t}\Big\{ -\frac{1}{2}\log|2\pi A| - \frac{1}{2}[\tilde{\mathbf{a}}_i^{(k)}]^T A^{-1}[\tilde{\mathbf{a}}_i^{(k)}]\Big\} \\
&= \arg\min_A \sum_{i=1}^{n}\sum_{k=1}^{k_t} \frac{1}{k_t}\Big\{ \log|A| + [\tilde{\mathbf{a}}_i^{(k)}]^T A^{-1}[\tilde{\mathbf{a}}_i^{(k)}]\Big\} \\
&= \arg\min_A \Big\{ n\log|A| + \frac{1}{k_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t} [\tilde{\mathbf{a}}_i^{(k)}]^T A^{-1}[\tilde{\mathbf{a}}_i^{(k)}]\Big\} \\
&= \arg\min_A \Big\{ n\log|A| + \frac{1}{k_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t} \mathrm{tr}([\tilde{\mathbf{a}}_i^{(k)}]^T A^{-1}[\tilde{\mathbf{a}}_i^{(k)}])\Big\} \\
&= \arg\min_A \Big\{ n\log|A| + \frac{1}{k_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t} \mathrm{tr}(A^{-1}[\tilde{\mathbf{a}}_i^{(k)}][\tilde{\mathbf{a}}_i^{(k)}]^T)\Big\} \\
&= \arg\min_A \Big\{ n\log|A| + \frac{1}{k_t}\mathrm{tr}\Big\{ A^{-1}\sum_{i=1}^{n}\sum_{k=1}^{k_t}[\tilde{\mathbf{a}}_i^{(k)}][\tilde{\mathbf{a}}_i^{(k)}]^T\Big\}\Big\} \\
&= \arg\min_A \Big\{ \log|A| + \mathrm{tr}\Big\{ A^{-1}\frac{1}{nk_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t}[\tilde{\mathbf{a}}_i^{(k)}][\tilde{\mathbf{a}}_i^{(k)}]^T\Big\}\Big\} \\
&= \frac{1}{nk_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t}[\tilde{\mathbf{a}}_i^{(k)}][\tilde{\mathbf{a}}_i^{(k)}]^T.
\end{aligned}$$

Similarly, the maximizer $B^{(t+1)}$ for $Q^{(4)}$ can be obtained by

$$\begin{aligned}
B^{(t+1)} &= \arg\max_B \{Q^{(4)}(B|\boldsymbol{\theta}^{(t)})\} \\
&= \arg\max_B \sum_{i=1}^{n}\sum_{k=1}^{k_t} \frac{1}{k_t} \log f(\tilde{\mathbf{b}}_i^{(k)}; B) \\
&= \frac{1}{nk_t}\sum_{i=1}^{n}\sum_{k=1}^{k_t}[\tilde{\mathbf{b}}_i^{(k)}][\tilde{\mathbf{b}}_i^{(k)}]^T.
\end{aligned}$$

To obtain the asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$, we can use the formula of Louis (1982), which involves evaluating the second-order derivative of the complete-

data log-likelihood function. Alternatively, we may consider the following approximate formula (McLachlan and Krishnan, 1997). Let $\mathbf{s}_c^{(i)} = \partial l_c^{(i)}/\partial \boldsymbol{\theta}$, where $l_c^{(i)}$ is the complete-data log-likelihood for individual i. Then an approximate formula for the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{n} E(\mathbf{s}_c^{(i)}|\mathbf{y}_i,\ \mathbf{z}_i;\ \hat{\boldsymbol{\theta}})\ E(\mathbf{s}_c^{(i)}|\mathbf{y}_i,\ \mathbf{z}_i;\ \hat{\boldsymbol{\theta}})^T\right]^{-1}.$$

where the expectations can be approximated by Monte-Carlo empirical means, as above.

In summary, the foregoing MCEM algorithm proceeds as follows.

Step 1. Obtain an initial estimate of $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ based on a naive method such as the two-step method, and set $\mathbf{a}_i^{(0)} = \mathbf{0}$ and $\mathbf{b}_i^{(0)} = \mathbf{0}$.

Step 2. At the $(t+1)$th $(t \geq 0)$ iteration, obtain Monte Carlo samples of the "missing data" $(\mathbf{a}_i,\ \mathbf{b}_i)$ using the Gibbs sampler along with rejection sampling methods by sampling from the full conditionals $[\mathbf{a}_i|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{b}_i;\ \boldsymbol{\theta}^{(t)}]$ and $[\mathbf{b}_i|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{a}_i;\ \boldsymbol{\theta}^{(t)}]$, or using importance sampling methods to approximate the conditional expectation in the E-step.

Step 3. Obtain updated estimates $\boldsymbol{\theta}^{(t+1)}$ using standard complete-data optimization procedures.

Step 4. Iterate between Step 2 and Step 3 until convergence.

3.3.3 Sampling Methods

Gibbs Sampler

For the proposed Monte Carlo EM algorithm, we can see that generating samples from the conditional distribution $[\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_i,\ \mathbf{z}_i;\ \boldsymbol{\theta}^{(t)}]$ is an important step for implementing the E-step of the Monte Carlo EM algorithm. The Gibbs sampler (Gelfand and Smith, 1990) is a popular method to generate samples from a complicated multi-dimensional distribution by sampling from full conditionals in turn, until convergence after a burn-in period. Here, we

use the Gibbs sampler to simulate the "missing" random effects $\mathbf{a}_i$ and $\mathbf{b}_i$. Set initial values $(\tilde{\mathbf{a}}_i^{(0)}, \tilde{\mathbf{b}}_i^{(0)})$. If the current generated values are $(\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)})$, we can obtain $(\tilde{\mathbf{a}}_i^{(k+1)}, \tilde{\mathbf{b}}_i^{(k+1)})$ as follows:

Step 1. Draw a sample for the "missing" random effects $\tilde{\mathbf{a}}_i^{(k+1)}$ from the full conditional

$$f(\mathbf{a}_i|\mathbf{y}_i, \mathbf{z}_i, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\theta}^{(t)}).$$

Step 2. Draw a sample for the "missing" random effects $\tilde{\mathbf{b}}_i^{(k+1)}$ from the full conditional

$$f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i, \tilde{\mathbf{a}}_i^{(k+1)}; \boldsymbol{\theta}^{(t)}).$$

We assess the convergence of the Gibbs sampler by examining time series plots and sample autocorrelation function plots. After a sufficiently large burn-in of r iterations, the sampled values will achieve a steady state as reflected by the time series plots. Then, $\{(\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)})\}$ can be treated as a sample from the multidimensional density function

$$f(\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)}).$$

If we choose a reasonably large gap r' (say $r' = 10$), we can treat the sample series $\{(\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}), k = r + r', r + 2r', \dots\}$ as an independent sample from the multidimensional density function. The simplest choice for initial values $(\tilde{\mathbf{a}}_i^{(0)}, \tilde{\mathbf{b}}_i^{(0)})$ is $(\mathbf{0}, \mathbf{0})$.

Multivariate Rejection Algorithm

Sampling from the two full conditionals can be accomplished by rejection sampling methods as follows. If the density functions are log-concave in the appropriate parameters, the adaptive rejection algorithm of Gilks and Wild (1992) may be used, as in Ibrahim et al. (1999). However, for arbitrary NLME models, some densities may not be log-concave. In such cases, the multivariate rejection sampling method (see Section 3.2 in Geweke, 1996) may be used to obtain the desirable samples. Booth and Hobert (1999) discussed such a

method in the context of complete-data generalized linear models, which can be extended to our models. For example, consider sampling from $f(\mathbf{a}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ in (3.4). Let $f^*(\mathbf{a}_i) = f(\mathbf{z}_i|\mathbf{a}_i; \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ and $\varsigma = \sup_{\mathbf{u}}\{f^*(\mathbf{u})\}$. We assume $\varsigma < \infty$. A random sample from $f(\mathbf{a}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ can then be obtained as follows by multivariate rejection sampling:

Step 1. Sample $\mathbf{a}_i^*$ from $f(\mathbf{a}_i; \boldsymbol{\theta}^{(t)})$, and independently, sample w from the uniform (0, 1) distribution.

Step 2. If $w \leq f^*(\mathbf{a}_i^*)/\varsigma$, then accept $\mathbf{a}_i^*$, otherwise, go back to step 1.

Samples from $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{a}_i; \boldsymbol{\theta}^{(t)})$ can be obtained in a similar way. Therefore, the Gibbs sampler in conjunction with the multivariate rejection sampling can be used to obtain samples from $[\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)}]$. Booth and Hobert (1999) noted that, when it is easy to simulate from the assumed densities, the multivariate rejection sampling method can be very fast even if the acceptance rate is quite low.

Importance Sampling

When the dimensions of $\mathbf{a}_i$ or $\mathbf{b}_i$ are not small, however, the foregoing rejection sampling methods may be slow. In this case, we may consider importance sampling methods where the importance function can be chosen to be a multivariate Student t density whose mean and variance match the mode and curvature of $f(\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)})$. Note that a multivariate t distribution, which has heavier tails than a multivariate normal distribution, will produce a more robust approximation since underestimating the tails can have serious consequences such as unstable behavior that may be difficult to diagnose. Booth and Hobert (1999) discussed an importance sampling method for complete-data generalized linear models. Here, we may extend their method to our models and use importance sampling methods to approximate the integral in the E-step. Specifically, we write

$$f(\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)}) = s \exp[h(\mathbf{a}_i, \mathbf{b}_i)],$$

where s is an unknown normalizing constant. Let $\dot{h}(\mathbf{a}_i, \mathbf{b}_i)$ and $\ddot{h}(\mathbf{a}_i, \mathbf{b}_i)$ be the first and second derivatives of $h(\mathbf{a}_i, \mathbf{b}_i)$ respectively, and let $(\mathbf{a}_i^*, \mathbf{b}_i^*)$ be the solution of $\dot{h}(\mathbf{a}_i, \mathbf{b}_i) = 0$, which is the maximizer of $h(\mathbf{a}_i, \mathbf{b}_i)$. Then, the Laplace approximations of the mean and variance of $f(\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)})$ are $(\mathbf{a}_i^*, \mathbf{b}_i^*)$ and $-(\ddot{h}(\mathbf{a}_i^*, \mathbf{b}_i^*))^{-1}$ respectively. Suppose that $\{(\tilde{\mathbf{a}}_i^{*(1)}, \tilde{\mathbf{b}}_i^{*(1)}), \ldots, (\tilde{\mathbf{a}}_i^{*(k_t)}, \tilde{\mathbf{b}}_i^{*(k_t)})\}$ is a random sample of size k_t generated from an importance function $h^*(\mathbf{a}_i, \mathbf{b}_i)$, which is assumed to have the same support as $f(\mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)})$. Then we have

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{i=1}^{n} \left\{ \frac{1}{k_t} \sum_{k=1}^{k_t} w_{ik}^{(t)} \, l_c^{(i)}(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{z}_i, \tilde{\mathbf{a}}_i^{*(k)}, \tilde{\mathbf{b}}_i^{*(k)}) \right\},$$

where

$$w_{ik}^{(t)} = \frac{f(\tilde{\mathbf{a}}_i^{*(k)}, \tilde{\mathbf{b}}_i^{*(k)}|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)})}{h^*(\tilde{\mathbf{a}}_i^{*(k)}, \tilde{\mathbf{b}}_i^{*(k)})}$$

are importance weights. Other sampling methods have also been proposed (e.g. McCulloch, 1997).

For the above sampling methods, the adaptive rejection method is applicable only when the appropriate densities are log-concave, while the multivariate rejection sampling method and the importance sampling method are applicable in general. Adaptive and multivariate rejection sampling methods may be efficient when the dimensions of the random effects and the sample sizes are small. When the dimension of the integral in the E-step is high, however, rejection sampling methods can be inefficient due to low acceptance rate. If the sample size is not small, importance sampling methods may be more efficient than rejection sampling methods since in this case the importance function may closely resemble the true conditional distribution.

3.3.4 Convergence

For Monte Carlo EM algorithms, the incomplete-data log-likelihood is not guaranteed to increase at each iteration due to the Monte Carlo error at the E-step. However, under suitable regularity conditions, Monte Carlo EM algorithms still converge to the MLEs (Chan and Ledolter, 1995). When applying the Monte Carlo EM algorithm, Monte Carlo samples for the "missing" random effects are drawn at each EM iteration. Consequently, Monte Carlo errors are introduced. The Monte Carlo errors are affected by the Monte Carlo sample size. It is obvious that larger values of the Monte Carlo sample size k_t will result in more precise but slower computation. A common strategy is to increase k_t as the number t of EM iterations increases (Booth and Hobert, 1999). For sufficiently large values of k_t, the Monte Carlo EM algorithm would inherit the properties of the exact versions, such as the likelihood increasing properties of EM, but this would substantially increase the computational work load. Thus, we usually use a relatively small k_t at initial iterations, and then increase k_t with the iteration number t.

If the Monte Carlo error associated with $\boldsymbol{\theta}^{(t+1)}$ is large, the $(t+1)$th iteration of the Monte Carlo EM algorithm is wasted because the EM step is swamped by the Monte Carlo error. Booth and Hobert (1999) proposed an automated method for choosing k_t in the context of complete-data generalized linear models. Their method can be extended to our case in a straightforward way as follows.

Let

$$\begin{aligned} Q^{(1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}, \\ Q^{(2)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}, \end{aligned}$$

and let $\boldsymbol{\theta}^{*(t+1)}$ be the solution to $Q^{(1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0}$. When the simulated samples are independent, it can be seen that the conditional distribution of $[\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}]$ is approximately normal

with mean $\boldsymbol{\theta}^{*(t+1)}$ and a covariance matrix that can be estimated by

$$\widehat{\mathrm{Cov}}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = Q^{(2)}(\boldsymbol{\theta}^{*(t+1)}|\boldsymbol{\theta}^{(t)})^{-1}\,\widehat{\mathrm{Cov}}(Q^{(1)}(\boldsymbol{\theta}^{*(t+1)}|\boldsymbol{\theta}^{(t)}))\,Q^{(2)}(\boldsymbol{\theta}^{*(t+1)}|\boldsymbol{\theta}^{(t)})^{-1},$$

where

$$\widehat{\mathrm{Cov}}\left(Q_i^{(1)}(\boldsymbol{\theta}^{*(t+1)}|\boldsymbol{\theta}^{(t)})\right) = \frac{1}{k_t}\sum_{k=1}^{k_t}\left\{\left[w_{ik}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\mathbf{y}_i,\,\mathbf{z}_i,\,\tilde{\mathbf{a}}_i^{(k)},\,\tilde{\mathbf{b}}_i^{(k)};\,\boldsymbol{\theta}^{*(t+1)}\right)\right]\right.$$
$$\left.\left[w_{ik}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\mathbf{y}_i,\,\mathbf{z}_i,\,\tilde{\mathbf{a}}_i^{(k)},\,\tilde{\mathbf{b}}_i^{(k)};\,\boldsymbol{\theta}^{*(t+1)}\right)\right]^T\right\},$$

$(\tilde{\mathbf{a}}_i^{(k)},\,\tilde{\mathbf{b}}_i^{(k)})$ are simulated samples, and w_{ik} are the importance weights when the importance sampling is used and are all set to 1 when rejection sampling methods are used. After the $(t+1)$th iteration, we may construct an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\theta}^{*(t+1)}$ based on the above normal approximation. The EM step is swamped by the Monte Carlo error if the previous value $\boldsymbol{\theta}^{(t)}$ lies in the confidence ellipsoid, and in that case we need to increase k_t. For example, we may set k_t to be $k_{t-1} + k_{t-1}/c$ for some positive constant c and appropriate k_0. Increasing k_t with each iteration may speed up the EM convergence (Booth and Hobert, 1999). Note that this method of choosing k_t is completely automated.

The proposed Monte Carlo EM algorithm often works well for models with a small dimension of random effects. When the dimension of random effects is not small, however, the proposed MCEM algorithm and Gibbs sampler may converge very slowly or even may not converge. Therefore, in the next section, we propose an alternative approximate inference method which may avoid these convergence difficulties and may be more efficient in computation.

3.4 A Computationally More Efficient Approximate Method

3.4.1 The Need for an Alternative Method

The Monte Carlo EM method in the previous section may be computationally very intensive and may offer potential computational problems such as slow or non-convergence, especially when the dimensions of the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ are not small. When the dimensions of the random effects are not small, sampling the random effects in the E-step may lead to inefficient and computationally unstable Gibbs sampler, and may lead to a high degree of auto-correlation and lack of convergence. To overcome these difficulties, in this section we propose an alternative approximate method by iteratively using a first-order Taylor approximation to the nonlinear models. The proposed method avoids sampling the random effects and provides analytic expressions for parameter estimates at each iteration, So it may be preferable when the Monte Carlo EM method exhibits computational difficulties. Alternatively, the proposed method in this section can be used to obtain excellent parameter starting values for the Monte Carlo EM method.

For complete-data NLME models, approximate methods have been widely used, and these approximate methods perform reasonably well in most cases (Lindstrom and Bates, 1990; Pinheiro and Bates, 1995; Vonesh et al., 2002). These approximate methods are typically obtained via Taylor expansions or Laplace approximations to the nonlinear models. One particularly popular approximate method for complete-data NLME models is that of Lindstrom and Bates (1990), which is equivalent to *iteratively* carrying out maximum likelihood based on certain LME models (Wolfinger, 1993). Following Lindstrom and Bates (1990), we propose to further approximate model (2.5) by taking a first-order Taylor expansion around the current parameter and random effects estimates, which leads to a LME response model. For the resulting LME response model, with the covariate model (2.8), we

update parameter estimates based on the distribution of observations and an EM algorithm. In each iteration, analytic expressions for parameter estimates are available. Therefore, the proposed approximate method may provide substantial computational advantages over the Monte Carlo EM method.

We rewrite the NLME model (2.6) and (2.7) as a single equation

$$y_{ij} = g_{ij}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \mathbf{a}_i,\, \mathbf{b}_i) + e_{ij}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, n_i, \tag{3.6}$$

where $g_{ij}(\cdot)$ is a nonlinear function. Let $\mathbf{g}_i = (g_{i1}, \ldots, g_{in_i})^T$. Denote the current estimates of $(\boldsymbol{\theta},\, \mathbf{a}_i,\, \mathbf{b}_i)$ by $(\tilde{\boldsymbol{\theta}},\, \tilde{\mathbf{a}}_i,\, \tilde{\mathbf{b}}_i)$. Taking a first-order Taylor expansion of g_{ij} around the current parameter estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ and random effects estimates $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{b}}_i$, we obtain the following LME response model

$$\tilde{\mathbf{y}}_i = W_i\, \boldsymbol{\alpha} + X_i\, \boldsymbol{\beta} + H_i\, \mathbf{a}_i + T_i\, \mathbf{b}_i + \mathbf{e}_i, \tag{3.7}$$

where

$$\begin{aligned}
W_i &= (\mathbf{w}_{i1}, \ldots, \mathbf{w}_{in_i})^T \text{ with } \mathbf{w}_{ij} = \frac{\partial g_{ij}}{\partial \boldsymbol{\alpha}} \\
X_i &= (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^T \text{ with } \mathbf{x}_{ij} = \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}} \\
H_i &= (\mathbf{h}_{i1}, \ldots, \mathbf{h}_{in_i})^T \text{ with } \mathbf{h}_{ij} = \frac{\partial g_{ij}}{\partial \mathbf{a}_i} \\
T_i &= (\mathbf{t}_{i1}, \ldots, \mathbf{t}_{in_i})^T \text{ with } \mathbf{t}_{ij} = \frac{\partial g_{ij}}{\partial \mathbf{b}_i} \\
\tilde{\mathbf{y}}_i &= \mathbf{y}_i - \mathbf{g}_i(\tilde{\boldsymbol{\alpha}},\, \tilde{\boldsymbol{\beta}},\, \tilde{\mathbf{a}}_i,\, \tilde{\mathbf{b}}_i) + W_i\, \tilde{\boldsymbol{\alpha}} + X_i\, \tilde{\boldsymbol{\beta}} + H_i\, \tilde{\mathbf{a}}_i + T_i\, \tilde{\mathbf{b}}_i,
\end{aligned}$$

with all the partial derivatives being evaluated at $(\tilde{\boldsymbol{\alpha}},\, \tilde{\boldsymbol{\beta}},\, \tilde{\mathbf{a}}_i,\, \tilde{\mathbf{b}}_i)$.

The proposed method in this section is to *iteratively* carry out maximum likelihood based on the LME response model (3.7) and the covariate model (2.8). Let $\boldsymbol{\gamma} = (\boldsymbol{\alpha},\, \boldsymbol{\beta})$ be the mean parameters and $\boldsymbol{\lambda} = (\delta^2, R, A, B)$ be the variance-covariance parameters. The algorithm consists of alternately obtaining approximate estimates of $\boldsymbol{\gamma}$ given the current

estimates of variance-covariance parameters $\boldsymbol{\lambda}$ based on the distribution of $\tilde{\boldsymbol{r}}_i = (\tilde{\mathbf{y}}_i^T, \mathbf{z}_i^T)^T$ and then updating the variance-covariance parameter estimates via an EM step using the posterior curvatures of $(\mathbf{e}_i, \boldsymbol{\epsilon}_{ij}, \mathbf{a}_i, \mathbf{b}_i)$, as in Laird and Ware (1982).

3.4.2 Analytic Expressions of Estimates

We can combine the LME response model (3.7) with the covariate model (2.8) to form a unified LME model

$$\tilde{\boldsymbol{r}}_i = Q_i \boldsymbol{\gamma} + Z_i \boldsymbol{\omega}_i + \mathbf{v}_i, \qquad i = 1, \ldots, n, \tag{3.8}$$

where $\tilde{\boldsymbol{r}}_i = (\tilde{\mathbf{y}}_i^T, \mathbf{z}_i^T)^T$, $\boldsymbol{\omega}_i = (\mathbf{a}_i^T, \mathbf{b}_i^T)^T$, $\mathbf{v}_i = (\mathbf{e}_i^T, \boldsymbol{\epsilon}_i^T)^T$, $U_i = (U_{i1}^T, \ldots, U_{im_i}^T)^T$, $V_i = (V_{i1}^T, \ldots, V_{im_i}^T)^T$, and

$$Q_i = \begin{pmatrix} W_i & X_i \\ U_i & 0 \end{pmatrix}, \quad Z_i = \begin{pmatrix} H_i & T_i \\ V_i & 0 \end{pmatrix},$$

with 0's being appropriate zero matrices.

For the unified LME model (3.8), by standard arguments, we have

$$[\tilde{\boldsymbol{r}}_i | \boldsymbol{\omega}_i; \boldsymbol{\gamma}, \tilde{\boldsymbol{\lambda}}] \sim N(Q_i \boldsymbol{\gamma} + Z_i \boldsymbol{\omega}_i, \Lambda_i), \qquad [\tilde{\boldsymbol{r}}_i; \boldsymbol{\gamma}, \tilde{\boldsymbol{\lambda}}] \sim N(Q_i \boldsymbol{\gamma}, \Sigma_i), \tag{3.9}$$

where $\Sigma_i = Z_i D Z_i^T + \Lambda_i$,

$$\Lambda_i = \begin{pmatrix} \tilde{\delta}^2 I & 0 \\ 0 & I \otimes \tilde{R} \end{pmatrix}, \quad D = \begin{pmatrix} \tilde{A} & 0 \\ 0 & \tilde{B} \end{pmatrix},$$

and the Kronecker product $I \otimes \tilde{R}$ is a $\nu m_i \times \nu m_i$ matrix with the $\nu \times \nu$ submatrix $\tilde{R}$ on the diagonals and zeros elsewhere. Using known results for LME models (e.g., Vonesh and Chinchilli, 1997), we can update the estimate of $\boldsymbol{\gamma}$ given the current estimates $\tilde{\boldsymbol{\lambda}}$ by

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} Q_i^T \Sigma_i^{-1} Q_i \right)^{-1} \left(\sum_{i=1}^{n} Q_i^T \Sigma_i^{-1} \tilde{\boldsymbol{r}}_i \right). \tag{3.10}$$

Based on the unified LME model (3.8), we can also obtain the joint distribution of $(\tilde{\boldsymbol{r}}_i, \boldsymbol{\omega}_i)$ at $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$

$$\left[\begin{pmatrix} \tilde{\boldsymbol{r}}_i \\ \boldsymbol{\omega}_i \end{pmatrix}; \hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\lambda}}\right] \sim N\left[\begin{pmatrix} Q_i\hat{\boldsymbol{\gamma}} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_i & Z_i\, D \\ D\, Z_i^T & D \end{pmatrix}\right],$$

from which we obtain the conditional distribution of $\boldsymbol{\omega}_i$ given the observed data $\tilde{\boldsymbol{r}}_i$ at $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$:

$$[\boldsymbol{\omega}_i | \tilde{\boldsymbol{r}}_i; \hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\lambda}}] \sim N[D\, Z_i^T\, \Sigma_i^{-1}(\tilde{\boldsymbol{r}}_i - Q_i\, \hat{\boldsymbol{\gamma}}),\ D - D\, Z_i^T\, \Sigma_i^{-1}\, Z_i\, D]. \tag{3.11}$$

An estimate of the random effects $\boldsymbol{\omega}_i$ is thus given by

$$\hat{\boldsymbol{\omega}}_i = D\, Z_i^T\, \Sigma_i^{-1}(\tilde{\boldsymbol{r}}_i - Q_i\, \hat{\boldsymbol{\gamma}}). \tag{3.12}$$

Finally, following Laird and Ware (1982), we can update the variance-covariance parameter estimates as follows. Note that if we were to observe $\mathbf{a}_i$, $\mathbf{b}_i$, $\mathbf{e}_i$, and $\boldsymbol{\epsilon}_i$, in addition to $\mathbf{y}_i$ and $\mathbf{z}_i$, we would have the following estimates

$$\hat{\delta}^2 = \sum_{i=1}^{n} \mathbf{e}_i^T\, \mathbf{e}_i / \sum_{i=1}^{n} n_i, \qquad \hat{B} = \sum_{i=1}^{n} \mathbf{b}_i\, \mathbf{b}_i^T / n,$$
$$\hat{R} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \boldsymbol{\epsilon}_{ij}\, \boldsymbol{\epsilon}_{ij}^T / \sum_{i=1}^{n} m_i \qquad \hat{A} = \sum_{i=1}^{n} \mathbf{a}_i\, \mathbf{a}_i^T / n,$$

where $\sum_{i=1}^{n} \mathbf{e}_i^T\, \mathbf{e}_i$, $\sum_{i=1}^{n} \sum_{j=1}^{m_i} \boldsymbol{\epsilon}_{ij}^T\, \boldsymbol{\epsilon}_{ij}$, $\sum_{i=1}^{n} \mathbf{a}_i\, \mathbf{a}_i^T$, and $\sum_{i=1}^{n} \mathbf{b}_i\, \mathbf{b}_i^T$ are the "sufficient" statistics for δ^2, R, A, and B, respectively. Since $\mathbf{a}_i$, $\mathbf{b}_i$, $\mathbf{e}_i$, and $\boldsymbol{\epsilon}_{ij}$ are unobservable, we can "estimate" them by their conditional expectations given the observed data $\mathbf{y}_i$ and $\mathbf{z}_i$ at the current parameter estimates $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$. Based on standard results for multivariate normal distributions, we know that

$$\left[\begin{pmatrix} \tilde{\boldsymbol{r}}_i \\ \mathbf{e}_i \end{pmatrix}; \boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}, \boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}\right] \sim N\left[\begin{pmatrix} Q_i\, \hat{\boldsymbol{\gamma}} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_i & L_i \\ L_i^T & \tilde{\delta}^2\, I \end{pmatrix}\right],$$

and

$$\left[\begin{pmatrix} \tilde{\boldsymbol{r}}_i \\ \boldsymbol{\epsilon}_{ij} \end{pmatrix}; \boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}, \boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}\right] \sim N\left[\begin{pmatrix} Q_i\hat{\boldsymbol{\gamma}} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_i & R_{ij} \\ R_{ij}^T & \tilde{R} \end{pmatrix}\right],$$

where L_i is a $(n_i + \nu m_i) \times n_i$ matrix with the first $n_i \times n_i$ submatrix $\tilde{\delta}^2 I$ and zeros elsewhere, R_{ij} is a $(n_i + \nu m_i) \times \nu$ matrix consisting of the first $n_i \times \nu$ submatrix 0 and the remaining m_i $(\nu \times \nu)$ square submatrices with the jth square submatrix $\tilde{R}$ and zeros elsewhere. By the definition of conditional distributions, it can be shown that

$$[\mathbf{e}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}}] \sim N[L_i^T\,\Sigma_i^{-1}(\tilde{\boldsymbol{r}}_i - Q_i\,\hat{\boldsymbol{\gamma}}),\ \tilde{\delta}^2\,I - L_i^T\,\Sigma_i^{-1}\,L_i],$$
$$[\boldsymbol{\epsilon}_{ij}|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}}] \sim N[R_{ij}^T\,\Sigma_i^{-1}(\tilde{\boldsymbol{r}}_i - Q_i\hat{\boldsymbol{\gamma}}),\ \tilde{R} - R_{ij}^T\,\Sigma_i^{-1}\,R_{ij}].$$

Using the expectation and covariance properties for multivariate random variables and some matrix algebra, we update the estimates of the variance-covariance parameters (δ^2, R, A, B) as follows:

$$\begin{aligned}
\hat{\delta}^2 &= \sum_{i=1}^{n} E(\mathbf{e}_i^T\,\mathbf{e}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})/\sum_{i=1}^{n} n_i \\
&= \sum_{i=1}^{n}[\mathrm{tr}(\mathrm{Cov}(\mathbf{e}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})) + E(\mathbf{e}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})^T E(\mathbf{e}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})]/\sum_{i=1}^{n} n_i, \\
\hat{R} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} E(\boldsymbol{\epsilon}_{ij}\,\boldsymbol{\epsilon}_{ij}^T|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})/\sum_{i=1}^{n} m_i \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i}[\mathrm{Cov}(\boldsymbol{\epsilon}_{ij}|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}}) + E(\boldsymbol{\epsilon}_{ij}|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})\,E(\boldsymbol{\epsilon}_{ij}|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})^T]/\sum_{i=1}^{n} m_i, \\
\hat{A} &= \sum_{i=1}^{n} E(\mathbf{a}_i\,\mathbf{a}_i^T|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})/n \\
&= \sum_{i=1}^{n}[\mathrm{Cov}(\mathbf{a}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}}) + E(\mathbf{a}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})\,E(\mathbf{a}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})^T]/n, \\
\hat{B} &= \sum_{i=1}^{n} E(\mathbf{b}_i\,\mathbf{b}_i^T|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})/n \\
&= \sum_{i=1}^{n}[\mathrm{Cov}(\mathbf{b}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}}) + E(\mathbf{b}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})\,E(\mathbf{b}_i|\tilde{\boldsymbol{r}}_i;\ \hat{\boldsymbol{\gamma}},\ \tilde{\boldsymbol{\lambda}})^T]/n.
\end{aligned} \tag{3.13}$$

The foregoing results show the closed-form expressions of the parameter estimates at each iteration for LME model (3.8). Iteratively solving LME model (3.8) until convergence leads to an approximate MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\lambda}})$ of $\boldsymbol{\theta}$.

3.4.3 Asymptotic Properties

Following Vonesh et al. (2002), we show in Section 3.7 that, under fairly mild regularity conditions, $\hat{\gamma}$ satisfies the following properties

$$\begin{aligned}\hat{\gamma} &= \hat{\gamma}_{MLE} + O_p\Big\{\Big(\min_i N_i\Big)^{-1/2}\Big\} \\ &= \gamma_0 + O_p\Big\{\max\Big[n^{-1/2},\ \Big(\min_i N_i\Big)^{-1/2}\Big]\Big\},\end{aligned}$$

where $N_i = n_i + m_i$, and γ_0 and $\hat{\gamma}_{MLE}$ are the true value and exact MLE of γ, respectively. Thus the approximate MLE $\hat{\gamma}$ is not only consistent but also asymptotically equivalent to the exact MLE. The rate of convergence is shown to depend on both the number n_i of observations per individual and the number n of individuals.

Moreover, the estimate $\hat{\gamma}$ asymptotically follows a normal distribution:

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(\mathbf{0},\ \Omega(\gamma_0)),$$

where the asymptotic variance-covariance matrix, $\Omega(\gamma_0)$, can be consistently estimated by $\Big[\frac{1}{n}\sum_{i=1}^{n} Q_i^T\, \Sigma_i^{-1}\, Q_i\Big]^{-1}\Big|_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}$.

The proofs of the above results are given in Section 3.7.

3.5 Example and Simulation

3.5.1 An Application in AIDS Studies

We apply the foregoing proposed methods to a HIV dataset for illustration. We also compare the proposed methods with the commonly-used *naive method* which ignores covariate measurement errors and the two-step method, which is described in Section 3.2.

Data Description

The study consists of 46 HIV infected patients who were treated with a potent antiretroviral regimen consisting of protease inhibitor and reverse transcriptase inhibitor drugs. Viral loads (Plasma HIV-1 RNA copies) were measured on days 0, 2, 7, 10, 14, 21, 28 and weeks 8, 12, 24, and 48 after initiation of treatments. After the antiretroviral treatment, the patients' viral loads will typically decay, and the decay rates may reflect the efficacy of the treatment. Throughout the time course, the viral load may continue to decay, fluctuate, or even start to rise (rebound). The data at the late stage of study are likely to be contaminated by long-term clinical factors, which leads to complex longitudinal trajectories. Various covariates such as CD4 count were also recorded throughout the study on similar schedules. The viral load has a detectable limit of 100 RNA copies/mL. For simplicity, we imputed the censored viral loads, which are below the detection limit, by half the detection limit 50, as in Wu and Zhang (2002). The number of measurements for each individual varies from 4 to 10. There were 72 out of 361 CD4 measurements missing at viral load measurement times, due mainly to a somewhat different CD4 measurement schedule. The detailed data description can be found in Wu and Ding (1999) and Wu (2002).

The Response and Covariate Models

Modelling HIV viral dynamics after anti-HIV treatments has received a great deal of attention in recent years (Ho et al., 1995; Perelson et al., 1996; Wu and Ding, 1999; Wu, 2005). The following HIV viral dynamic model (first stage model) has been widely used (Wu 2002; Wu and Zhang, 2002)

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{3.14}$$

where y_{ij} is the $\log_{10}$-transformation of the viral load measurement for patient i at time t_{ij}, P_{1i} and P_{2i} are baseline values, and λ_{1ij} and λ_{2ij} are the first (initial) and the second phase viral decay rates, respectively, and they may be interpreted as the turnover rates of

productively infected cells and long-lived and/or latently infected cells respectively. The $\log_{10}$-transformation of the viral load is used to make the data more normally distributed and to stabilize the variance.

Wu (2002) noted that variation in the dynamic parameters such as the first phase decay rate λ_{1ij} may be partially associated with variation in CD4 counts. In AIDS studies, it is known that covariates such as CD4 count are often measured with substantial errors. Thus we assume that the dynamic parameters are related to the *true* covariate values, reflecting the belief that actual, not possibly corrupted, covariate values govern the model parameters, as in Higgins et al. (1997) and Wu (2002).

Due to long-term clinical factors, drug resistance, and other complications, the viral load trajectories can be very complex after the initial phase viral decay (see Figure 1.1). Grossman et al. (1999) pointed out that the viral decay rate after the initial period may be complicated and may vary over time since they may depend on some phenomenological parameters which hide considerable microscopic complexity and change over time. Therefore, a nonparametric smooth curve modelling for the second phase viral decay rate may be more appropriate than parametric modelling (Wu and Zhang, 2002). Based on the reasons noted above, we consider the following second stage model, which corresponds to the first stage model (3.14),

$$\begin{aligned} \log(P_{1i}) &= \beta_1 + b_{1i}, \qquad & \lambda_{1ij} &= \beta_2 + \beta_3 z^*_{ij} + b_{2i}, \\ \log(P_{2i}) &= \beta_4 + b_{3i}, \qquad & \lambda_{2ij} &= w(t_{ij}) + h_i(t_{ij}), \end{aligned} \tag{3.15}$$

where z^*_{ij} is the true (but unobserved) CD4 count, and $w(t_{ij})$ and $h_i(t_{ij})$ are nonparametric smooth fixed- and random-effects functions defined in Section 2.1. To avoid very small (large) estimates, which may be unstable, we standardize the CD4 counts and rescale the original time t (in days) so that the new time scale is between 0 and 1.

As discussed in Section 2.1, we employ the linear combinations of natural cubic splines

Table 3.1: AIC and BIC values for the viral load model (3.14) and (3.15), with $q \leq p = 1, 2, 3$.

Model	p=1,q=1	p=2,q=2	p=2,q=1	p=3,q=3	p=3,q=2	p=3,q=1
AIC	615.96	583.54	585.39	577.37	586.45	576.43
BIC	678.18	669.09	656.43	670.71	665.90	651.50

with the percentile-based knots to approximate the nonparametric smooth functions $w(t)$ and $h_i(t)$. Following Wu and Zhang (2002), we take the same natural cubic splines with $q \leq p$ in order to decrease the dimension of the random effects $\mathbf{b}_i$, i.e., more basis functions used to approximate the fixed-effects function than the random-effects functions. AIC and BIC criteria are used to determine the values of p and q. We use the observed CD4 counts for the unobservable true CD4 counts in the response model (3.14) and (3.15), and use SPLUS functions *nlme*() and *anova*() to obtain the values of AIC and BIC. Table 3.1 displays AIC and BIC values for various plausible models. Based on these AIC and BIC values, the model with $p = 3$ and $q = 1$, i.e.,

$$\lambda_{2ij} \approx \beta_5 + \beta_6 \psi_1(t_{ij}) + \beta_7 \psi_2(t_{ij}) + b_{4i}, \tag{3.16}$$

seems to be the best, and thus it is selected for our analysis.

For the CD4 process, in the absence of a theoretical rationale, we consider empirical polynomial LME models, and choose the best fitted model based again on AIC/BIC values for each possible model. This is done based on the observed CD4 values, and is done separately from the response model for simplicity. Specifically, since the inter-patient variation is large, we consider model (2.8) with $U_{il} = V_{il} = (1, u_{il}, \ldots, u_{il}^{a-1})$ and linear ($a = 2$), quadratic ($a = 3$), and cubic ($a = 4$) polynomials. Table 3.2 presents AIC and BIC values for these three models. The following quadratic polynomial LME model best fits the observed CD4

Table 3.2: AIC and BIC values for the linear, quadratic, and cubic CD4 models.

Model	a=2	a=3	a=4
AIC	796.17	703.19	742.12
BIC	819.50	761.52	781.01

process:

$$\text{CD4}_{il} = (\alpha_1 + a_1) + (\alpha_2 + a_2)\, u_{il} + (\alpha_3 + a_3)\, u_{il}^2 + \epsilon_{il}, \tag{3.17}$$

where u_{il} is the time and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ are the population parameters and $\mathbf{a}_i = (a_{i1}, a_{i2}, a_{i3})^T$ are the random effects.

Estimation Methods and Computation Issues

We estimate the parameters in the response and covariate models using the *naive method* which ignores measurement errors, the two-step method in Section 3.2, and the two proposed "joint" model methods discussed in Sections 3.3 and 3.4. We denote the method in Section 3.3 by MCEM and the method in Section 3.4 by APPR.

The two proposed joint model methods need starting values for model parameters. We respectively use the parameter estimates obtained by the naive method and by the two-step method as parameter starting values for the two joint model methods.

For the naive method and the two-step method, we use SPLUS functions *lme*() and *nlme*() to obtain parameter estimates and their default standard errors. For the MCEM method, we assess the convergence of the Gibbs sampler by examining time series plots and sample autocorrelation function plots. For example, Figures 3.1 and 3.2 show the time series and the autocorrelation function plots for b_2 associated with patient 10. From these figures, we notice that the Gibbs sampler converges quickly and the autocorrelations between successive generated samples are negligible after lag 15. Time series and autocorrelation

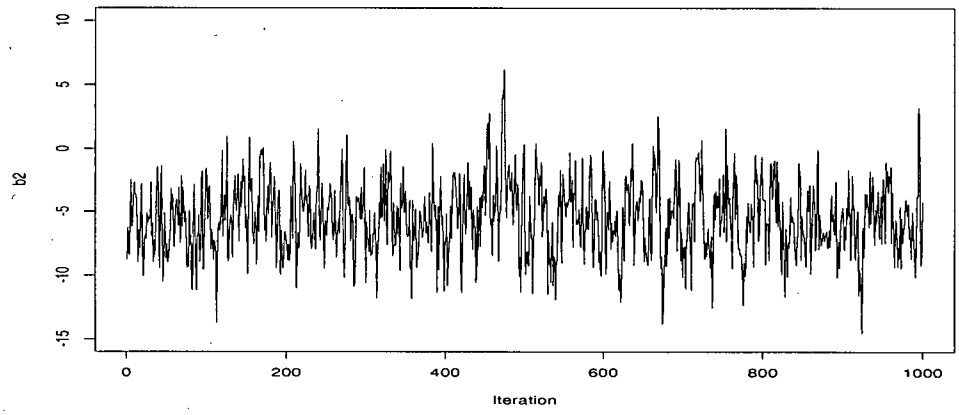

Figure 3.1: The time series plot for b_2 associated with patient 10.

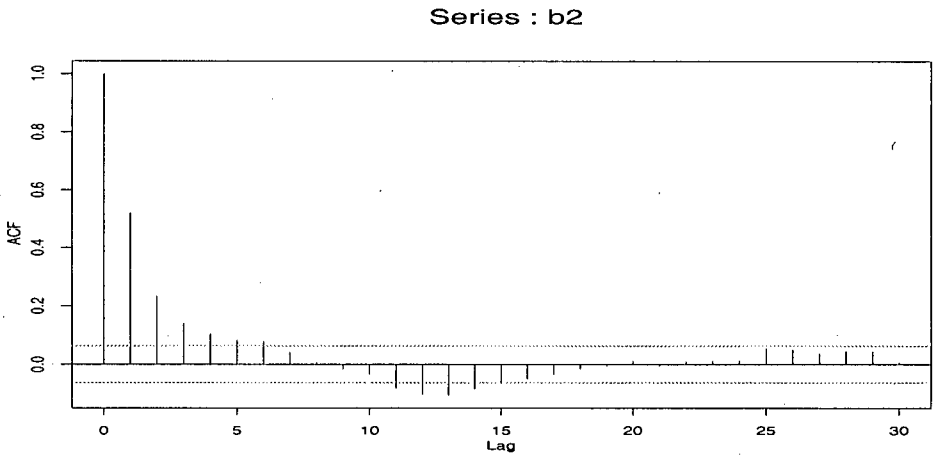

Figure 3.2: The autocorrelation function plot for b_2 associated with patient 10.

function plots for other random effects show similar behaviors. Therefore, we discard the first 500 samples as the burn-in, and then we take one sample from every 20 simulated samples to obtain "independent" samples (see sampling methods in Section 3.3.3).

For the Monte-Carlo EM algorithm, we start with $k_0 = 500$ Monte-Carlo samples, and increase the Monte-Carlo sample size as the number of iteration t increases: $k_{t+1} = k_t + k_t/c$ with $c = 4$. Convergence criterion for the iterative methods in our examples is that the relative change in the parameter estimates from successively iterations is smaller than a given tolerance level (e.g., 0.01). However, due to Monte Carlo errors induced by Gibbs sampler, it is difficult to converge for an extremely small tolerance level, otherwise it may converge very slowly. The actual tolerance level we used in our example for the two proposed joint model methods is 0.05. Convergence of the algorithms are considered to be achieved when the maximum relative change of all estimates is less than 5% in two consecutive iterations. We use the multivariate rejection sampling method for the MCEM method. Other sampling methods may also be applied.

On a SUN Sparc work-station, the MCEM method took about 90 minutes to converge while the APPR method took only 3 minutes to converge. This shows that the APPR method offers quite a substantial reduction in computing time, and is thus computationally much more efficient than the MCEM method.

Analysis Results and Conclusions

Table 3.3 presents the resulting parameter estimates and standard errors. We see that, except for the naive method, the other three methods give similar point estimates for the parameters, especially for the covariate model parameters $\boldsymbol{\alpha}$. However, for the parameters $\boldsymbol{\beta}$ of main interest, the two-step method gives *smaller standard errors* than the two joint model methods (MCEM and APPR). This is because the two-step method ignores the variability due to estimating the parameters $\boldsymbol{\alpha}$ in the covariate model. Thus, these results are consistent

Table 3.3: Parameter estimates (standard errors) for the HIV dataset.

Method	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7	δ	R
Naive	–	–	–	11.73	65.41	.41	6.82	–3.01	9.27	–1.67	.35	
	–	–	–	(.2)	(3.9)	(3.3)	(.6)	(5.6)	(8.8)	(3.6)		
Two-step	–.42	4.17	–3.74	11.73	65.78	1.53	6.84	–2.86	9.04	–1.75	.35	.52
	(.1)	(.5)	(.5)	(.2)	(4.1)	(4.7)	(.6)	(5.5)	(8.9)	(3.5)		
MCEM	–.41	4.02	–3.54	11.74	66.60	1.55	6.85	–2.78	8.91	–1.79	.35	.53
	(.1)	(.5)	(.6)	(.2)	(4.7)	(5.2)	(.7)	(6.0)	(9.0)	(3.6)		
APPR	–.42	4.17	–3.74	11.74	65.72	1.33	6.85	–2.82	8.99	–1.77	.35	.52
	(.1)	(.6)	(.6)	(.2)	(4.5)	(5.0)	(.7)	(5.9)	(9.1)	(3.6)		

Note: A and B are unstructured covariance matrices, but we only report the estimates of their diagonal elements here. $Diag(\hat{A}) = (.52, 4.06, 1.98)$ *for Two-step,* $Diag(\hat{A}) = (.53, 2.55, 1.25)$ *for MCEM, and* $Diag(\hat{A}) = (.52, 4.06, 1.99)$ *for APPR.* $Diag(\hat{B}) = (1.11, 69.94, 2.02, 24.86)$ *for Naive,* $Diag(\hat{B}) = (1.10, 69.58, 2.02, 25.04)$ *for Two-step,* $Diag(\hat{B}) = (1.11, 69.96, 2.05, 25.47)$ *for MCEM, and* $Diag(\hat{B}) = (1.10, 69.78, 2.01, 24.85)$ *for APPR.*

with the analytical results about the two-step method in Section 3.2. We also see that the naive method may severely under-estimate the effect of the covariate CD4 (which is measured by the parameter β_3). The estimates and standard errors based on the two joint model methods are similar and may be more reliable.

The commonly used two-step method and the naive method may give misleading results, and the two proposed joint model methods may be more reliable. We will confirm this conclusion via simulations in next section.

3.5.2 A Simulation Study

In this section, we conduct a simulation study to evaluate the proposed methods (MCEM and APPR), and compare them with the commonly used two-step method and the naive method by the mean-square-error (MSE). The models and the measurement schedules used in the simulation are the same as those in the real HIV dataset in the previous section (i.e., models (3.14) – (3.17)). In the simulations, the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are shown in

Table 3.4: Simulation results for parameter estimates as well as (standard errors) and (simulated standard errors)* for the estimation methods Naive, Two-step, MCEM, and APPR.

Parameter	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7
True Value	–.5	4.0	–4.0	12.0	66.0	1.5	7.0	–3.0	9.0	–2.0
Naive Method	–	–	–	11.98	65.50	**0.94**	6.92	–3.74	10.13	–1.62
	–	–	–	(.1)	(1.1)	(1.1)	(.3)	(1.7)	(2.6)	(.9)
	–	–	–	(.2)*	(1.2)*	(1.0)*	(.3)*	(1.6)*	(2.8)*	(1.0)*
Two-step	–.51	4.05	–4.02	11.98	65.66	1.53	6.92	–3.74	10.13	–1.62
	(.1)	(.3)	(.4)	(.2)	(1.2)	(1.4)	(.3)	(1.7)	(2.6)	(.9)
	(.1)*	(.3)*	(.3)*	(.1)*	(1.2)*	(1.4)*	(.2)*	(1.6)*	(2.5)*	(.9)*
MCEM	–.51	4.04	–4.02	11.98	65.85	1.48	6.98	–3.11	9.15	–1.95
	(.1)	(.4)	(.4)	(.1)	(1.5)	(1.8)	(.3)	(2.0)	(3.0)	(1.1)
	(.1)*	(.4)*	(.4)*	(.2)*	(1.5)*	(1.8)*	(.3)*	(1.9)*	(2.8)*	(1.1)*
APPR	–.51	4.05	–4.03	11.98	65.46	1.52	6.93	–3.82	10.29	–1.56
	(.1)	(.3)	(.4)	(.2)	(1.4)	(1.8)	(.3)	(2.0)	(3.1)	(1.1)
	(.1)*	(.3)*	(.3)*	(.2)*	(1.4)*	(1.7)*	(.3)*	(1.9)*	(2.8)*	(1.0)*

Table 3.4, and the other true parameter values are $\delta = .1$, $R = .3$, $A = \text{diag}(.5, 2, 1)$, and $B = \text{diag}(1, 9, 2, 4)$. We simulated 100 data sets and calculated averages of the resulting estimates and their standard errors as well as simulated standard errors based on each of the four estimation methods. Since MCEM method sometimes offers computational problems, such as slow or non-convergence, the 100 sets of parameter estimates are obtained from 116 data sets. The simulation results are shown in Table 3.4.

From Table 3.4, we see that the naive method can severely under-estimate the covariate effect β_3. The two-step method produces similar point estimates as the two joint model methods (MCEM and APPR), but it gives *smaller standard errors* than the MCEM and the APPR methods, since the two-step method fails to incorporate variability in estimating the covariate model. These results are consistent with the analytical results in Section 3.2. Note that the estimates for the covariate model parameters $\boldsymbol{\alpha}$ are similar for the three methods (Two-step, MCEM, and APPR), but the parameters $\boldsymbol{\alpha}$ are usually treated as nuisance pa-

rameters and are not of primary interest. The two proposed joint model methods (MCEM and APPR) perform better than the two-step method and the naive method, and the MCEM method is the best among all four methods in terms of bias. In the computation, the APPR method converges much faster than the MCEM method. These simulation results confirm that the naive method and the two-step method may give misleading results and that the two proposed methods are more reliable.

3.6 Discussion

We have proposed two approximate likelihood methods for semiparametric NLME models with covariate measurement errors and missing data. The first method, implemented by a Monte Carlo EM algorithm combined with Gibbs sampler, may be more accurate but may be computationally intensive and sometimes may offer computational problems such as slow or non-convergence. The second method, implemented by an iterative algorithm without Monte Carlo approximation, is computationally much more efficient, but it may be less accurate than the first method since it uses an additional approximation. Alternatively, the second method may provide excellent parameter starting values for the first method. Simulation results show that both methods perform better than the commonly used two-step method and a naive method. In particular, the commonly used two-step method may under-estimate standard errors, and the naive method may under-estimate covariate effects.

For semiparametric NLME models with covariate measurement errors and missing data, the models can be very complex. Thus, if there is not sufficient information in the data, the models can be non-identifiable. To our knowledge, there seem no existing general necessary and sufficient conditions for model identifiability, and the identifiability problem needs to be considered on a case-by-case basis. In practice, we can check model identifiability

by examining the convergence of iterative algorithms. If the model is non-identifiable, the iterative algorithms may diverge quickly. For the models considered here, we find that the iterative algorithms converged without problems, so the models seem identifiable.

The methods proposed here may be extended to semiparametric generalized linear mixed models and nonparametric mixed-effects models with covariate measurement errors and missing data. The results will be reported in the near future.

3.7 Appendix: Asymptotic Properties of $\hat{\gamma}$ Based on the APPR Method in Section 3.4

In this section, we show the asymptotic properties of $\hat{\gamma}$ obtained by the APPR method in Section 3.4. We first state some Lemmas, which are used in the proofs, in Section 3.7.1. Then we describe some regularity conditions under which the asymptotic properties hold in Section 3.7.2. In Section 3.7.3, we obtain some estimating equations, which are used for showing asymptotic properties of $\hat{\gamma}$. Consistency and asymptotic normality of $\hat{\gamma}$ are shown in Sections 3.7.4 and 3.7.5, respectively. The results are extensions of Vonesh et al. (2002).

3.7.1 Some Lemmas

The following four lemmas will be used for showing asymptotic properties of $\hat{\gamma}$.

Lemma 3.2. (Vonesh and Chinchilli, 1997). Let Y_n be a sequence of random variables satisfying $Y_n = c + O_p(a_n)$ where $a_n = o(1)$. If $f(x)$ is a function with r continuous derivatives at $x = c$, then

$$f(Y_n) = f(c) + f^{(1)}(c)(Y_n - c) + \cdots + [1/(r-1)!] f^{(r-1)}(c)(Y_n - c)^{r-1} + O_p(a_n^r),$$

where $f^{(k)}(c)$ is the kth derivative of f evaluated at c. In particular,

$$f(Y_n) = f(c) + O_p(a_n).$$

This result holds when $O_p(\cdot)$ is replaced everywhere by $o_p(\cdot)$ or when Y_n and c are replaced by a vector/matrix random variable $\mathbf{Y}_n$ and vector/matrix constant c.

Lemma 3.3. (Aitchison and Silvey, 1958). If f is a continuous function mapping R^s into itself with the property that, for every θ such that $\|\theta\| = 1$, $\theta^T f(\theta) < 0$, then there exists a point $\hat{\theta}$ such that $\|\hat{\theta}\| < 1$ and $f(\hat{\theta}) = \mathbf{0}$.

Lemma 3.4 (The Bounded Convergence Theorem). Let $\{f_n(x)\}$ be a sequence of measurable functions defined on a set of E of finite measure, and suppose there is a real number M such that $|f_n(x)| \leq M$ for all n and all x. If $f(x) = \lim_{n\to\infty} f_n(x)$ for each x in E, then

$$\lim_{n\to\infty} \int_E f_n(x)dx = \int_E \lim_{n\to\infty} f_n(x)dx = \int_E f(x)dx.$$

Lemma 3.5. Let A and B be $\nu \times \nu$ symmetric matrices with eigenvalues $\mu_1(A) \geq \mu_2(A) \geq \cdots \geq \mu_\nu(A)$ and $\mu_1(B) \geq \mu_2(B) \geq \cdots \geq \mu_\nu(B)$, respectively. If $A - B$ are nonnegative definite, denoted by $A - B \geq 0$ or $A \geq B$, then we have $\mu_i(A) \geq \mu_i(B)$, $i = 1, \ldots, \nu$.

3.7.2 Notation and Regularity Conditions

Let the τ-dimensional vector $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Gamma$ and

$$l_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) = l_i(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i; \mathbf{y}_i, \mathbf{z}_i) = \log f_Y(y_i|\mathbf{z}_i, \boldsymbol{\omega}_i; \boldsymbol{\gamma}) + \log f_Z(\mathbf{z}_i|\boldsymbol{\omega}_i; \boldsymbol{\gamma}),$$

$$N_i L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) = l_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) + \log f(\mathbf{a}_i) + \log f(\mathbf{b}_i),$$

where $N_i = n_i + m_i$. Let

$$l'_{i,\boldsymbol{\omega}_i}(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma)) = \frac{\partial}{\partial \boldsymbol{\omega}_i} l_i(\gamma, \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i(\gamma)},$$

$$l''_{i,\boldsymbol{\omega}_i\boldsymbol{\omega}_i}(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma)) = \frac{\partial^2}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_i^T} l_i(\gamma, \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i(\gamma)},$$

$$l'_{i,\gamma}(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma)) = \frac{\partial}{\partial \gamma} l_i(\gamma, \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i(\gamma)},$$

$$l''_{i,\gamma\gamma}(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma)) = \frac{\partial^2}{\partial \gamma \partial \gamma^T} l_i(\gamma, \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i(\gamma)},$$

$$l''_{i,\gamma\boldsymbol{\omega}_i}(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma)) = \frac{\partial^2}{\partial \gamma \partial \boldsymbol{\omega}_i^T} l_i(\gamma, \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i(\gamma)}, \quad \text{etc.}$$

Similarly, we can define the corresponding derivatives for $L_i(\gamma, \hat{\boldsymbol{\omega}}_i(\gamma))$ and

$$l'_{i,\boldsymbol{\omega}_i}(\gamma, \boldsymbol{\omega}_i) = \frac{\partial}{\partial \boldsymbol{\omega}_i} l_i(\gamma, \boldsymbol{\omega}_i), \qquad l''_{i,\boldsymbol{\omega}_i\boldsymbol{\omega}_i}(\gamma, \boldsymbol{\omega}_i) = \frac{\partial^2}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_i^T} l_i(\gamma, \boldsymbol{\omega}_i), \text{ etc.}$$

Also, we denote convergence in probability as $N_i \to \infty$ by $o_p(1_{N_i})$, convergence in probability as $n \to \infty$ by $o_p(1_n)$, and convergence in probability as both $N_i \to \infty$ and $n \to \infty$ by $o_p(1_{N_i,n})$. We show consistency and asymptotic normality under the following regularity conditions. An outline of the proof when some of these conditions are relaxed is provided at the end.

R1. $N_i = O(N)$ uniformly for $i = 1, \ldots, n$, where $N = \min_i N_i$.

R2. The variance-covariance parameters $\boldsymbol{\lambda} = (\delta^2, R, A, B)$ are fixed and known, and the true parameter γ_0 is in the interior of Γ. Q_i, Λ_i, Z_i, Σ_i, and D are evaluated at $\boldsymbol{\theta}$ and $\boldsymbol{\omega}_i$. When $\boldsymbol{\lambda}$ is unknown, we can simply replace it by its consistent estimate (e.g., in (3.13)).

R3. The density functions $f_Y(y_{ij}|\mathbf{z}_i, \boldsymbol{\omega}_i; \gamma)$ and $f_Z(\mathbf{z}_{ij}|\boldsymbol{\omega}_i; \gamma)$ satisfy the necessary regularity conditions (e.g., Bradley and Gart, 1962) such that, for fixed γ, the MLE of $\boldsymbol{\omega}_i$ is $\sqrt{N_i}$-consistent for $\boldsymbol{\omega}_i$ as $N_i \to \infty$. In addition, the necessary regularity conditions are assumed (e.g., Serfling, 1980, p. 27, Theorem C) such that, by the Law of Large Numbers,

the following hold:

$$-N_i^{-1}\frac{\partial^2}{\partial\gamma\partial\gamma^T}l_i(\gamma,\,\omega_i) = N_i^{-1}Q_i^T\Lambda_i^{-1}Q_i + o_p(1_{N_i}) \quad \text{as} \quad N_i \to \infty,$$

$$-N_i^{-1}\frac{\partial^2}{\partial\omega_i\partial\omega_i^T}l_i(\gamma,\,\omega_i) = N_i^{-1}Z_i^T\Lambda_i^{-1}Z_i + o_p(1_{N_i}) \quad \text{as} \quad N_i \to \infty,$$

$$-N_i^{-1}\frac{\partial^2}{\partial\gamma\partial\omega_i^T}l_i(\gamma,\,\omega_i) = N_i^{-1}Q_i^T\Lambda_i^{-1}Z_i + o_p(1_{N_i}) \quad \text{as} \quad N_i \to \infty,$$

$$-N_i^{-1}\frac{\partial^2}{\partial\omega_i\partial\gamma^T}l_i(\gamma,\,\omega_i) = N_i^{-1}Z_i^T\Lambda_i^{-1}Q_i + o_p(1_{N_i}) \quad \text{as} \quad N_i \to \infty,$$

where, under models (2.6) – (2.8),

$$E_{\boldsymbol{r}|\boldsymbol{\omega}}\left\{-\frac{\partial^2}{\partial\gamma\partial\gamma^T}l_i(\gamma,\,\omega_i)\right\} = Q_i^T\Lambda_i^{-1}Q_i,$$

$$E_{\boldsymbol{r}|\boldsymbol{\omega}}\left\{\frac{\partial^2}{\partial\omega_i\partial\omega_i^T}l_i(\gamma,\,\omega_i)\right\} = Z_i^T\Lambda_i^{-1}Z_i,$$

$$E_{\boldsymbol{r}|\boldsymbol{\omega}}\left\{\frac{\partial^2}{\partial\gamma\partial\omega_i^T}l_i(\gamma,\,\omega_i)\right\} = \left[E_{\boldsymbol{r}|\boldsymbol{\omega}}\left\{\frac{\partial^2}{\partial\gamma\partial\omega_i^T}l_i(\gamma,\,\omega_i)\right\}\right]^T = Q_i^T\Lambda_i^{-1}Z_i.$$

Finally, the matrices $N_i^{-1}Q_i^T\Lambda_i^{-1}Q_i$ and $N_i^{-1}Z_i^T\Lambda_i^{-1}Z_i$ are both assumed to be positive definite with finite determinants such that, for example, the smallest eigenvalue of $N_i^{-1}Q_i^T\Lambda_i^{-1}Q_i$ exceeds λ_0 for some $\lambda_0 > 0$.

R4. For all $\gamma \in \Gamma$ and all the b-dimensional $\omega_i \in R^b$, the function $L_i(\gamma,\,\omega_i)$ is six times differentiable and continuous in γ and ω_i for all y_{ij} and $\mathbf{z}_{ij}$, and $L_i(\gamma,\,\omega_i)$ satisfies the necessary regularity conditions needed to change the order of integration and differentiation, as indicated in the proof.

R5. For any $\gamma \in \Gamma$, there exist $d_1 > 0$ and $\lambda_1 > 0$ such that

a. For all $\gamma^* \in B_{d_1}(\gamma)$, where $B_{d_1}(\gamma)$ is the τ-dimensional sphere centered at γ with radius d_1, the following holds:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma^T}l'_{i,\gamma}(\gamma,\,\hat{\omega}_i(\gamma))\,\big|_{\gamma=\gamma^*} = \Omega(\gamma^*)^{-1} + o_p(1_n), \quad \text{as} \quad n \to \infty,$$

where $\Omega(\gamma^*)^{-1}$ is positive definite with minimum eigenvalue greater than λ_1 and

$$\begin{aligned}\frac{\partial}{\partial \gamma^T} l'_{i,\gamma}(\gamma, \hat{\omega}_i(\gamma))\,|_{\gamma=\gamma^*} &= \{l''_{i,\gamma\gamma}(\gamma^*, \hat{\omega}_i(\gamma^*)) + l''_{i,\gamma\omega_i}(\gamma^*, \hat{\omega}_i(\gamma^*)) \\ &\quad \times [l''_{i,\omega\omega}(\gamma^*, \hat{\omega}_i(\gamma^*)) + \mathbf{D}^{-1}]^{-1} l''_{i,\omega\gamma}(\gamma^*, \hat{\omega}_i(\gamma^*))\}.\end{aligned}$$

b. The first, second, and third derivatives of $\sqrt{N_i} L_i(\gamma, \omega_i)$ with respect to ω_i are uniformly bounded in $B_{d_1}(\gamma)$.

R6. At the true value γ_0, the following hold true:

$$\begin{aligned}&E_{\omega}(Q_i^T \Sigma_i^{-1} Q_i) = \varphi_i(\gamma_0) \text{ exists for all } i = 1, \ldots, n, \\ &\lim_{n\to\infty} n^{-2} \sum_{i=1}^{n} \mathrm{Cov}_{\omega}(Q_i^T \Sigma_i^{-1} Q_i) = 0,\end{aligned}$$

and

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \varphi_i(\gamma_0) = \Omega(\gamma_0)^{-1},$$

where Q_i, Z_i, and Σ_i are evaluated at γ_0 and ω_i and $\Omega(\gamma_0)^{-1}$ is positive definite.

R7. The marginal densities, $\int \exp\{N_i L_i(\gamma, \omega_i)\}\, d\omega_i$, satisfy the necessary regularity conditions such that the MLE of γ exists and satisfies $(\hat{\gamma}_{MLE} - \gamma_0) = O_p(n^{-1/2})$.

3.7.3 Estimating Equations

In Section 3.4, we obtain an approximate MLE $\hat{\theta} = (\hat{\gamma}, \hat{\lambda})$ of θ and the predictor $\hat{\omega}_i$ of ω. It is obvious from (3.10) and (3.12) that the estimate $\hat{\gamma}$ is a function of $\hat{\lambda}$ and the final estimates $\hat{\omega}_i$ of ω_i are functions of both $\hat{\gamma}$ and $\hat{\lambda}$. For fixed λ, it can be shown that $\hat{\gamma}$ and $\hat{\omega}_i = \hat{\omega}_i(\hat{\gamma})$ maximize the complete-data log-likelihood function (3.2), i.e., $\sum_{i=1}^{n} N_i L_i(\gamma, \omega_i)$.

In fact, we can write $l_c^{(i)} = l_c^{(i)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i, \boldsymbol{\lambda})$ in (3.2) as

$$\begin{aligned} l_c^{(i)} &= -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\log|2\pi A| - \frac{1}{2}\log|2\pi B| \\ &\quad -\frac{1}{2}\left[\begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i \\ \mathbf{z}_i - U_i\,\boldsymbol{\alpha} - V_i\,\mathbf{a}_i \end{pmatrix}^T \Lambda_i^{-1} \begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i \\ \mathbf{z}_i - U_i\,\boldsymbol{\alpha} - V_i\,\mathbf{a}_i \end{pmatrix}\right] \\ &\quad -\frac{1}{2}\begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{pmatrix}^T D^{-1} \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{pmatrix}. \end{aligned}$$

Taking the first derivatives of $\sum_{i=1}^{n} l_c^{(i)}$ with respect to $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{\omega}_i = (\mathbf{a}_i^T, \mathbf{b}_i^T)^T$ and setting these first derivatives to appropriate zero vectors, we can obtain the following estimating equations

$$\begin{cases} \sum_{i=1}^{n} Q_i\,\Lambda_i^{-1} \begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i \\ \mathbf{z}_i - U_i\,\boldsymbol{\alpha} - V_i\,\mathbf{a}_i \end{pmatrix} = \mathbf{0} \\ Z_i\,\Lambda_i^{-1} \begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i \\ \mathbf{z}_i - U_i\,\boldsymbol{\alpha} - V_i\,\mathbf{a}_i \end{pmatrix} - D^{-1} \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{pmatrix} = \mathbf{0}, \quad i = 1, \ldots, n, \end{cases} \tag{3.18}$$

which is equivalent to

$$\begin{cases} \sum_{i=1}^{n} Q_i\,\Lambda_i^{-1}\,Q_i\,\boldsymbol{\gamma} + \sum_{i=1}^{n} Q_i\,\Lambda_i^{-1}\,Z_i\boldsymbol{\omega}_i = \sum_{i=1}^{n} Q_i\,\Lambda_i^{-1}\,\mathbf{r}_i, \\ Z_i\,\Lambda_i^{-1}\,Q_i\,\boldsymbol{\gamma} + [Z_i\,\Lambda_i^{-1}\,Z_i + D^{-1}]\boldsymbol{\omega}_i = Z_i\,\Lambda_i^{-1}\mathbf{r}_i, \quad i = 1\ldots, n, \end{cases} \tag{3.19}$$

where

$$\mathbf{r}_i = \begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i + \frac{\partial \mathbf{g}_i}{\partial \boldsymbol{\alpha}^T}\,\boldsymbol{\alpha} + \frac{\partial \mathbf{g}_i}{\partial \boldsymbol{\beta}^T}\,\boldsymbol{\beta} + \frac{\partial \mathbf{g}_i}{\partial \mathbf{a}_i^T}\,\mathbf{a}_i + \frac{\partial \mathbf{g}_i}{\partial \mathbf{b}_i^T}\,\mathbf{b}_i \\ \mathbf{z}_i \end{pmatrix}.$$

The solution to the estimating equations (3.19) can be obtained by iteratively solving the following equations

$$\begin{cases} \sum_{i=1}^{n} \tilde{Q}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{Q}_i\,\boldsymbol{\gamma} + \sum_{i=1}^{n} \tilde{Q}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{Z}_i\,\boldsymbol{\omega}_i = \sum_{i=1}^{n} \tilde{Q}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{\boldsymbol{r}}_i, \\ \tilde{Z}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{Q}_i\,\boldsymbol{\gamma} + (\tilde{Z}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{Z}_i + \tilde{D}^{-1})\,\boldsymbol{\omega}_i = \tilde{Z}_i^T\,\tilde{\Lambda}_i^{-1}\,\tilde{\boldsymbol{r}}_i, \quad i = 1\ldots, n, \end{cases} \tag{3.20}$$

where $\tilde{\mathbf{r}}_i = (\tilde{\mathbf{y}}_i^T, \mathbf{z}_i^T)^T$ is defined in (3.7), and $\tilde{Q}_i$, $\tilde{\Lambda}_i$, $\tilde{Z}_i$, and $\tilde{D}$ are Q_i, Λ_i, Z_i, and D evaluated at $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\omega}_i = \tilde{\boldsymbol{\omega}}_i$, respectively. The solution to the equations (3.20) is given in (3.10) and (3.12). Therefore, for fixed $\boldsymbol{\lambda}$, the final estimate $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_i = \hat{\boldsymbol{\omega}}_i(\hat{\boldsymbol{\gamma}})$ satisfy the estimating equations (3.18) and maximize the complete-data log-likelihood function (3.2). These facts will be used to show the following asymptotic properties of $\hat{\boldsymbol{\gamma}}$.

3.7.4 Consistency

We first note that, for fixed $\boldsymbol{\lambda}$, the MLE of $\boldsymbol{\gamma}$ will satisfy the set of estimating equations

$$J(\boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \prod_{i=1}^{n} f(\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \prod_{i=1}^{n} \int \exp\{N_i L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)\} d\boldsymbol{\omega}_i = \mathbf{0}.$$

Under R4, we have

$$\begin{aligned} J(\boldsymbol{\gamma}) &= \int \cdots \int \Big\{ \sum_{i=1}^{n} N_i \frac{\partial}{\partial \boldsymbol{\gamma}} L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) \Big\} \exp\Big\{ \sum_{j=1}^{n} N_j L_j(\boldsymbol{\gamma}, \boldsymbol{\omega}_j) \Big\} d\boldsymbol{\omega}_1 \cdots d\boldsymbol{\omega}_n \\ &= \sum_{i=1}^{n} \sqrt{N_i} \int \cdots \int \Big(\sqrt{N_i} \frac{\partial}{\partial \boldsymbol{\gamma}} L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) \Big) \exp\Big\{ \sum_{j=1}^{n} N_i L_j(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) \Big\} d\boldsymbol{\omega}_1 \cdots d\boldsymbol{\omega}_n \\ &= \sum_{i=1}^{n} \sqrt{N_i} \Big[\int \Big(\sqrt{N_i} L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) \Big) \exp\{N_i L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)\} d\boldsymbol{\omega}_i \times \prod_{j \neq i}^{n} \int \exp\{N_j L_j(\boldsymbol{\gamma}, \boldsymbol{\omega}_j)\} d\boldsymbol{\omega}_j \Big]. \end{aligned}$$

Now we examine the term $L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)$ in the above expression. Since the y_{ij}'s and $\mathbf{z}_{ij}$'s are conditionally independent each other given $\boldsymbol{\omega}_i$, by the Lindeberg Central Limit Theorem, it follows that conditional on $\boldsymbol{\omega}_i$,

$$\begin{aligned} L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) &= N_i^{-1} l'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) \\ &= N_i^{-1} Q_i^T \Lambda_i^{-1} \begin{pmatrix} \mathbf{y}_i - \mathbf{g}_i \\ \mathbf{z}_i - U_i \boldsymbol{\alpha} - V_i \mathbf{a}_i \end{pmatrix} = O_p(N_i^{-1/2}). \end{aligned} \tag{3.21}$$

Furthermore, under R3 it can be shown that the estimate

$$\hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}) = \boldsymbol{\omega}_i + O_p(N_i^{-1/2}). \tag{3.22}$$

Combining the results in (3.21) and (3.22) and applying Lemma 3.2 to $L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}))$, we can show that

$$\begin{aligned} L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) &= L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i) + O_p(N_i^{-1/2}) \\ &= O_p(N_i^{-1/2}) + O_p(N_i^{-1/2}) \\ &= O_p(N_i^{-1/2}). \end{aligned} \tag{3.23}$$

Then, by direct application of the Laplace approximation to integrals of the form $\int \exp\{kp(x)\}dx$ and $\int q(x)\exp\{kp(x)\}dx$, where $q(x)$ and $p(x)$ are smooth functions in x with $p(x)$ having a unique maximum at some point $\hat{x}$ (see, e.g., Barndorff-Nielson and Cox, 1989), it can be shown that

$$\int \exp\{N_i L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)\}d\boldsymbol{\omega}_i = \exp\{N_i L_i(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}))\}\left(\frac{2\pi}{|N_i \hat{L}''_{i,\boldsymbol{\omega}\boldsymbol{\omega}}|}\right)^{b/2}(1 + O(N_i^{-1}))$$

and

$$\begin{aligned} &\int \left(\sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)\right)\exp\{N_i L_i(\boldsymbol{\gamma}, \boldsymbol{\omega}_i)\}d\boldsymbol{\omega}_i \\ = \;& \exp\{N_i L_i(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}))\} \times \left(\frac{2\pi}{|N_i \hat{L}''_{i,\boldsymbol{\omega}\boldsymbol{\omega}}|}\right)^{b/2}\left(\sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) + O(N_i^{-1})\right), \end{aligned}$$

where $\hat{L}''_{i,\boldsymbol{\omega},\boldsymbol{\omega}} = L''_{i,\boldsymbol{\omega},\boldsymbol{\omega}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}))$. Because (3.21) implies $\sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) = O_p(1)$, it follows from R1 that

$$\left(\sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) + O_p(N_i^{-1})\right) \times \prod_{j\neq i}^{n}(1 + O_p(N_j^{-1})) = \sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) + O_p(N^{-1}).$$

Hence we have

$$\begin{aligned} J(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\sqrt{N_i}\left[\exp\left\{N_i L_i(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}))\right\}\left(\frac{2\pi}{|N_i \hat{L}''_{i,\boldsymbol{\omega}\boldsymbol{\omega}}|}\right)^{b/2}\left(\sqrt{N_i}L'_{i,\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})) + O(N_i^{-1})\right)\right. \\ &\qquad \left.\times \prod_{j\neq i}^{n}\left\{\exp\left\{N_j L_j(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_j(\boldsymbol{\gamma}))\right\}\left(\frac{2\pi}{|N_l \hat{L}''_{j,\boldsymbol{\omega}\boldsymbol{\omega}}|}\right)^{b/2}\left(1 + O(N_j^{-1})\right)\right\}\right] \\ &= K(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}(\boldsymbol{\gamma}))\left[\sum_{i=1}^{n} N_i L'_{i,\boldsymbol{\gamma}}\left(\boldsymbol{\gamma}, \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma})\right) + nO_p(N^{-1/2})\right] \end{aligned}$$

where $K(\gamma, \hat{\omega}(\gamma)) = \exp\{\sum_{i=1}^{n} N_i L_i'(\gamma, \hat{\omega}_i(\gamma))\} \prod_{i=1}^{n} (2\pi/|N_i \hat{L}_{i,\omega\omega}''|)^{b/2}$. Since $K(\gamma, \hat{\omega}(\gamma)) \neq 0$ for all $\gamma \in \Gamma$, the MLE $\hat{\gamma}_{MLE}$ of γ satisfies

$$J(\gamma)\Big|_{\gamma=\hat{\gamma}_{MLE}} = \mathbf{0} \iff J_1(\gamma, \hat{\omega}(\gamma))\Big|_{\gamma=\hat{\gamma}_{MLE}} + O_p(nN^{-1/2}) = \mathbf{0}, \tag{3.24}$$

where $J_1(\gamma, \hat{\omega}(\gamma)) = \sum_{i=1}^{n} N_i L_{i,\gamma}'(\gamma, \hat{\omega}_i(\gamma)) = \sum_{i=1}^{n} l_{i,\gamma}'(\gamma, \hat{\omega}_i(\gamma))$ is the set of estimating equations for γ conditional on fixed $\boldsymbol{\lambda}$, as given in (3.21). By taking a first-order Taylor series expansion of $J_1(\gamma, \hat{\omega}(\gamma))$ about $\hat{\gamma}_{MLE}$ and noting that, from (3.24), $J_1(\hat{\gamma}_{MLE}, \hat{\omega}(\hat{\gamma}_{MLE})) = O_p(nN^{-1/2})$, we have

$$J_1(\gamma, \hat{\omega}(\gamma)) = O_p(nN^{-1/2}) + J_1'(\gamma^*, \hat{\omega}(\gamma^*))(\gamma - \hat{\gamma}_{MLE}), \tag{3.25}$$

where

$$J_1'(\gamma^*, \hat{\omega}(\gamma^*)) = \frac{\partial}{\partial \gamma^T} J_1(\gamma, \hat{\omega}(\gamma))\,|_{\gamma=\gamma^*} = \frac{\partial}{\partial \gamma^T} \sum_{i=1}^{n} N_i L_{i,\gamma}'(\gamma, \hat{\omega}_i(\gamma))\,|_{\gamma=\gamma^*}$$

and γ^* is on the line segment joining γ to $\hat{\gamma}_{MLE}$. By applying the chain rule, we have, for any $\gamma \in \Gamma$,

$$\begin{aligned} J_1'(\gamma, \hat{\omega}(\gamma)) &= \sum_{i=1}^{n} \left\{ \frac{\partial^2}{\partial \gamma \partial \gamma^T} N_i L_i(\gamma; \omega_i)\Big|_{\omega_i=\hat{\omega}_i(\gamma)} + \frac{\partial^2}{\partial \gamma \partial \omega_i^T} N_i L_i(\gamma, \omega_i)\Big|_{\omega_i=\hat{\omega}_i(\gamma)} \frac{\partial \hat{\omega}_i(\gamma)}{\partial \gamma^T} \right\} \\ &= \sum_{i=1}^{n} \left\{ l_{i,\gamma\gamma}''(\gamma, \hat{\omega}_i(\gamma)) + l_{i,\gamma\omega}''(\gamma, \hat{\omega}_i(\gamma)) \frac{\partial \hat{\omega}_i(\gamma)}{\partial \gamma^T} \right\}. \end{aligned} \tag{3.26}$$

Note that $\hat{\omega}_i(\gamma)$ maximizes $N_i L_i(\gamma, \omega)$ and satisfies the second set of equations in (3.18), i.e.

$$l_{i,\omega}'(\gamma, \hat{\omega}_i(\gamma)) - D^{-1}\hat{\omega}_i(\gamma) = \mathbf{0} \iff \hat{\omega}_i(\gamma) = D\, l_{i,\omega}'(\gamma, \hat{\omega}_i(\gamma)).$$

Applying the chain rule once again, we have

$$\begin{aligned} \frac{\partial \hat{\omega}_i(\gamma)}{\partial \gamma^T} &= \frac{\partial}{\partial \gamma^T}\left\{ D l_{i,\omega}'(\gamma, \hat{\omega}_i(\gamma)) \right\} \\ &= D\left[\frac{\partial^2}{\partial \omega_i \partial \gamma^T} l_i(\gamma, \omega_i)\Big|_{\omega_i=\hat{\omega}_i(\gamma)} \right] + D\left[\frac{\partial^2}{\partial \omega_i \partial \omega_i^T} l_i(\gamma, \omega_i)\Big|_{\omega_i=\hat{\omega}_i(\gamma)} \right] \frac{\partial \hat{\omega}_i(\gamma)}{\partial \gamma^T} \\ &= D l_{i,\omega\gamma}''(\gamma, \hat{\omega}_i(\gamma)) + D l_{i,\omega\omega}''(\gamma, \hat{\omega}_i(\gamma)) \frac{\partial \hat{\omega}_i(\gamma)}{\partial \gamma^T}. \end{aligned}$$

Solving the above equation for $[\partial\hat{\omega}_i(\gamma)/\partial\gamma^T]$, we have

$$\frac{\partial\hat{\omega}_i(\gamma)}{\partial\gamma^T} = \Big\{ - l''_{i,\omega\omega}(\gamma,\, \hat{\omega}_i(\gamma)) + D^{-1} \Big\}^{-1} l''_{i,\omega\gamma}(\gamma,\, \hat{\omega}_i(\gamma)).$$

Substituting this expression of $[\partial\hat{\omega}_i(\gamma)/\partial\gamma^T]$ in (3.26), it follows from R1, R3, and R5 that as $n \to \infty$

$$\begin{aligned} -\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\gamma^T} l'_{i,\gamma}(\gamma,\, \hat{\omega}_i(\gamma))|_{\gamma=\gamma^*} \\ &= \Omega(\gamma^*)^{-1} + o_p(1_n), \end{aligned} \tag{3.27}$$

which implies $-\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) \xrightarrow{p} \Omega(\gamma^*)^{-1}$.

Taking $\epsilon = \frac{1}{2}\Omega(\gamma^*)^{-1}$ in the definition of convergence in probability, from (3.27), we have

$$\begin{aligned} & P\left(-\frac{1}{2}\Omega(\gamma^*)^{-1} < -\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) - \Omega(\gamma^*)^{-1} < \frac{1}{2}\Omega(\gamma^*)^{-1} \right) \to 1, \qquad \text{as } n \to \infty \\ \iff\ & P\left(\frac{1}{2}\Omega(\gamma^*)^{-1} < -\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) < \frac{3}{2}\Omega(\gamma^*)^{-1} \right) \to 1, \qquad \text{as } n \to \infty, \end{aligned}$$

which implies

$$P\left(\frac{1}{2}\Omega(\gamma^*)^{-1} < -\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) \right) \to 1, \qquad \text{as } n \to \infty.$$

Therefore, for sufficiently large n,

$$-\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*)) > \frac{1}{2}\Omega(\gamma^*)^{-1} > \frac{1}{2}\lambda_1 I \tag{3.28}$$

with probability 1, where λ_1 is defined in R5. Since, in (3.25), $O_p(N^{-1/2}) = N^{-1/2} O_p(1) \xrightarrow{p} \mathbf{0}$ as $N \to \infty$, similar arguments as for $-\frac{1}{n} J'_1(\gamma^*,\, \hat{\omega}_i(\gamma^*))$ lead to with probability 1, for any $0 < \kappa < d_1$,

$$\|O_p(N^{-1/2})\| \le \frac{\lambda_1 \kappa}{4}, \tag{3.29}$$

where d_1 is defined in R5. For any γ such that $\|\gamma - \hat{\gamma}_{MLE}\| = \kappa$, we know that $\|\gamma - \hat{\gamma}_{MLE}\|/\kappa = 1$, which satisfies the condition of Lemma 3.3. We can regard $J_1(\gamma, \hat{\omega}(\gamma))$ in (3.25) as a function of $\frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})$ on the unit sphere in R^r and using the results in (3.28) and (3.29), we have

$$\begin{aligned}
& \frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})^T \left[\frac{1}{n} J_1(\gamma, \ , \hat{\omega}(\gamma))\right] \\
= \ & \frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})^T O_p(N^{-1/2}) + \kappa \left[\frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})\right]^T \left[\frac{1}{n} J_1'(\gamma^*, \hat{\omega}(\gamma^*))\right] \left[\frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})\right] \\
\leq \ & \frac{1}{\kappa}\|\gamma - \hat{\gamma}_{MLE}\| \|O_p(N^{-1/2})\| - \frac{\lambda_1 \kappa}{2} \frac{1}{\kappa^2} \|\gamma - \hat{\gamma}_{ML}\|^2 \\
\leq \ & \frac{\lambda_1 \kappa}{4} - \frac{\lambda_1 \kappa}{2} < -\frac{\lambda_1 \kappa}{4} < 0.
\end{aligned}$$

Therefore, conditional on ω, we can show

$$P_{\omega}\left\{\frac{1}{\kappa}(\gamma - \hat{\gamma}_{MLE})^T \left[\frac{1}{n} J_1(\gamma, \ , \hat{\omega}(\gamma))\right] < 0\right\} \to 1 \quad \text{as } n \to \infty, N \to \infty.$$

Thus, since $\hat{\gamma}$ satisfies $J_1(\hat{\gamma}, \hat{\omega}(\hat{\gamma})) = \mathbf{0}$, Lemma 3.3 implies that

$$\lim_{n,N\to\infty} P_{\omega}\{\|\hat{\gamma} - \hat{\gamma}_{MLE}\| < \kappa\} = 1.$$

Hence, using Lemma 3.4, we have

$$\lim_{n,N\to\infty} P\{\|\hat{\gamma} - \hat{\gamma}_{MLE}\| < \kappa\} = E_{\omega}\left\{\lim_{n,N\to\infty} P_{\omega}\{\|\hat{\gamma} - \hat{\gamma}_{MLE}\| < \kappa\}\right\} = 1. \tag{3.30}$$

Since $J_1(\hat{\gamma}, \hat{\omega}(\hat{\gamma})) = \mathbf{0}$ and $J_1(\hat{\gamma}_{MLE}, \hat{\omega}(\hat{\gamma}_{MLE})) = O_p(nN^{-1/2})$ by (3.24), it follows from (3.25) that

$$-\frac{1}{n} J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))(\hat{\gamma} - \hat{\gamma}_{MLE}) = O_p(N^{-1/2}), \tag{3.31}$$

where γ^{**} is on the line segment joining $\hat{\gamma}$ to $\hat{\gamma}_{MLE}$. From (3.30), we know that γ^{**} is in the interior. It follows from R5 and inequality (3.28) that

$$-\frac{1}{n} J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**})) > \frac{1}{2}\Omega(\gamma^{**})^{-1} > \frac{1}{2}\lambda_1 I,$$

where $\Omega(\gamma^{**})^{-1}$ is positive definite. Lemma 3.5 implies that all eigenvalues of $\left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]$ are greater than $\frac{1}{2}\lambda_1$. Therefore, all eigenvalues of $\left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]^{-1}$, which are equal to the reciprocal of all eigenvalues of $\left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]$, are less than $\frac{2}{\lambda_1}$. So we have $\left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]^{-1} < \frac{2}{\lambda_1}I$, i.e., $\left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]^{-1} = O_p(1)$. Hence, from (3.31), we have

$$(\hat{\gamma} - \hat{\gamma}_{MLE}) = \left[-\frac{1}{n}J_1'(\gamma^{**}, \hat{\omega}(\gamma^{**}))\right]^{-1} O_p(N^{-1/2}) = O_p(1)\, O_p(N^{-1/2}) = O_p(N^{-1/2}),$$

from which it follows (given R7) that

$$\begin{aligned}
\hat{\gamma} &= \hat{\gamma}_{MLE} + O_p(N^{-1/2}) \\
&= \gamma_0 + O_p(n^{-1/2}) + O_p(N^{-1/2}) \\
&= \gamma_0 + O_p\left\{\max\left[n^{-1/2}, \left(\min_i N_i\right)^{-1/2}\right]\right\}.
\end{aligned}$$

3.7.5 Asymptotic Normality of $\hat{\gamma}$

The asymptotic normality of $\hat{\gamma}$ will be shown based on the estimating equations (3.10). Let

$$\mathbf{\Phi}(\gamma) = \sum_{i=1}^{n} \tilde{Q}_i^T \tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\gamma),$$

where $\tilde{Q}_i$, $\tilde{\Sigma}_i$, and $\tilde{\mathbf{r}}_i$ are defined in (3.20). The estimator $\hat{\gamma}$ satisfies $\mathbf{\Phi}(\hat{\gamma}) = \mathbf{0}$ at convergence. Noting that $\partial\mathbf{\Phi}(\gamma)/\partial\gamma^T = -\sum_{i=1}^{n} \tilde{Q}_i^T \tilde{\Sigma}_i^{-1} \tilde{Q}_i$ is constant for γ, we take a Taylor series expansion of $\mathbf{\Phi}(\hat{\gamma})$ around the true parameter γ_0:

$$\mathbf{0} = \mathbf{\Phi}(\hat{\gamma}) \;=\; \mathbf{\Phi}(\gamma_0) + \frac{\partial\mathbf{\Phi}(\gamma)}{\partial\gamma^T}(\hat{\gamma} - \gamma_0),$$

which implies

$$\begin{aligned}
\sqrt{n}(\hat{\gamma} - \gamma_0) &= \left[-\frac{1}{n}\frac{\partial\mathbf{\Phi}(\gamma)}{\partial\gamma^T}\right]^{-1}\left[\frac{1}{\sqrt{n}}\mathbf{\Phi}(\gamma_0)\right] \\
&= \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{Q}_i^T \tilde{\Sigma}_i^{-1} \tilde{Q}_i\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{Q}_i^T \tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\gamma_0)\right).
\end{aligned}$$

Since $E[\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\boldsymbol{\gamma}_0)] = 0$ and $\mathrm{Cov}[\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\boldsymbol{\gamma}_0)] = \tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i$, by the Lindeberg Central Limit Theorem, we have

$$\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\boldsymbol{\gamma}_0)\right) \xrightarrow{d} N(\mathbf{0}, I).$$

Noting that $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0 + o_p(1_{N,n})$ and $N_i = O(N)$, and using Lemma 3.2 and (3.22), we have

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_i(\hat{\boldsymbol{\gamma}}) &= \hat{\boldsymbol{\omega}}_i(\boldsymbol{\gamma}_0) + o_p(1_{N,n}) \\
&= \boldsymbol{\omega}_i + O_p(N_i^{-1/2}) + o_p(1_{N,n}) \\
&= \boldsymbol{\omega}_i + O_p(N^{-1/2}) + o_p(1_{N,n}) \\
&= \boldsymbol{\omega}_i + o_p(1_{N,n}).
\end{aligned}$$

Hence, it follows by the Law of Large Numbers, Lemma 3.2, and R6 that

$$\begin{aligned}
\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2} &\to \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2}_{\tilde{\boldsymbol{\gamma}}=\hat{\boldsymbol{\gamma}},\tilde{\boldsymbol{\omega}}=\hat{\boldsymbol{\omega}}(\hat{\boldsymbol{\gamma}})}, \qquad u \to \infty \\
&\xrightarrow{p} \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2}_{\tilde{\boldsymbol{\gamma}}=\boldsymbol{\gamma}_0,\tilde{\boldsymbol{\omega}}=\boldsymbol{\omega}}, \qquad N \to \infty,\ n \to \infty \\
&\xrightarrow{p} [\Omega(\boldsymbol{\gamma}_0)]^{1/2}, \qquad N \to \infty,\ n \to \infty,
\end{aligned}$$

where u is the number of iteration. Using Slutsky's theorem, we can show that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) &= \left[\tfrac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2}\left[\tfrac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}\tilde{Q}_i\right]^{-1/2}\left(\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{Q}_i^T\tilde{\Sigma}_i^{-1}(\tilde{\boldsymbol{r}}_i - \tilde{Q}_i\boldsymbol{\gamma}_0)\right) \\
&\xrightarrow{d} N(\mathbf{0}, \Omega(\boldsymbol{\gamma}_0)).
\end{aligned}$$

We can extend the foregoing proofs to the case where $\boldsymbol{\lambda}$ is unknown by replacing $\boldsymbol{\lambda}$ with its consistent estimate. Note that at the estimate $\hat{\boldsymbol{\gamma}}$, the estimate $\hat{\boldsymbol{\lambda}}$ given in (3.13) can be shown to be consistent for $\boldsymbol{\lambda}$ as $n \to \infty$ and $N \to \infty$ (see, e.g., Demidenko, 2004).

Chapter 4

Simultaneous Inference for Semiparametric NLME Models with Covariate Measurement Errors and Outcome-based Informative Dropouts

4.1 Introduction

In this chapter, we develop two approximate likelihood methods to simultaneously address covariate measurement errors and *outcome-based informative* dropouts in semiparametric NLME models. In Section 4.2, we propose models for this complicated problem. We obtain approximate MLEs of all model parameters, using a Monte-Carlo EM (MCEM) algorithm along with Gibbs sampler methods, in Sections 4.3. In Section 4.4, we propose an alternative and computationally much more efficient approximate method. The proposed methods are illustrated in a HIV dataset in Section 4.5 and are evaluated via simulation in Section 4.6.

We summarize this chapter in Section 4.7 with some discussion.

4.2 Models for Nonignorable Missing Responses

In the presence of nonignorable or informative dropouts in the semiparameteric NLME model (2.6) and (2.7) with the covariate process (2.8), we can write $\mathbf{y}_i = (\mathbf{y}_{mis,i}, \mathbf{y}_{obs,i})$ for individual i, where $\mathbf{y}_{mis,i}$ collects the missing components of $\mathbf{y}_i$ and $\mathbf{y}_{obs,i}$ collects the observed components of $\mathbf{y}_i$. Here, the missing y_{ij}'s are intermittent, i.e., we allow dropout individuals to possibly return to the study at a later time. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{in_i})^T$ be a vector of missing response indicators for individual i such that $r_{ij} = 1$ if y_{ij} is missing and 0 otherwise. Note that $r_{ij} = 1$ does not necessarily imply that $r_{i,j+1} = 1$. We have the observed data $\{(\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i),\ i = 1, \ldots, n\}$.

To allow for a nonignorable missing mechanism in the response, we need models for the missing response indicators $\mathbf{r}_i$, which are called dropout models. The parameters in the dropout models are treated as nuisance parameters and are usually not of inference interest. Thus, we try to reduce the number of nuisance parameters to make the estimation of main parameters more efficient. Moreover, too many nuisance parameters may even make the response and the covariate models non-identifiable. Therefore, we should be very cautious about adding extra nuisance parameters.

In general, the probability that y_{ij} is missing may depend on many factors, such as responses, covariates, and individual random effects, etc. Since the missing response indicators $\mathbf{r}_i$ are binary, a simple model for them is a logistic regression model as follows. We will assume that r_{ij}'s are independent for simplicity (to reduce the number of nuisance parameters)

$$f(\mathbf{r}_i|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\eta}) = \prod_{i=1}^{n_i}[P(r_{ij} = 1|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\eta})]^{r_{ij}}[1 - P(r_{ij} = 1|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\eta})]^{1-r_{ij}},$$

with

$$\text{logit}[P(r_{ij}=1|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta})] = \log\frac{P(r_{ij}=1|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta})}{[1-P(r_{ij}=1|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta})]} = h(\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta}),$$

where $\boldsymbol{\eta}$ are the unknown nuisance parameters and $h(\cdot)$ is often chosen to be a linear function of $\mathbf{y}_i$, $\mathbf{z}_i$, $\mathbf{a}_i$, and $\mathbf{b}_i$. More complicated models can also be considered, but they may introduce more parameters and increase the computational burden. Note that the missingness of y_{ij} may depend on the (unobserved) true covariates $\mathbf{z}_i^*$ rather than the observed error-prone covariates $\mathbf{z}_i$. In this case, a method similar to the one described below can be developed and will be discussed in the next chapter.

The density function $f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta})$ is a general expression of the missing response mechanism. Little (1995) pointed out two ways to incorporate informative missingness:

- *outcome-based informative* if $f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta}) = f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\boldsymbol{\eta})$. That is, the probability that the current response is missing depends on the possibly unobserved response $\mathbf{y}_i$ and covariates $\mathbf{z}_i$ but not on the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$. For example, a patient does not show up because he is too sick to go to the clinic.

- *random-effect-based informative* if $f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta}) = f(\mathbf{r}_i|\mathbf{a}_i,\,\mathbf{b}_i;\,\boldsymbol{\eta})$. That is, the probability that the current response is missing depends on the underlying unobservable random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ but not on $\mathbf{y}_i$ and $\mathbf{z}_i$. For example, a patient may be more likely to drop out if his initial viral decay is slower than other patients.

In this chapter, we focus on the *outcome-based informative* missing mechanism. Diggle and Kenward (1994), Little (1995), and Ibrahim et al. (2001) discussed various specifications of the *outcome-based informative* missing mechanism. We will assume that, for example, r_{ij} are independent with

$$\text{logit}[P(r_{ij}=1|\mathbf{y}_i,\,\mathbf{z}_i;\,\boldsymbol{\eta})] = \eta_1 + \eta_2 z_{i1j} + \eta_3 z_{i2j} + \cdots + \eta_{\nu+1} z_{i\nu j} + \eta_{\nu+2} y_{ij}.$$

More complicated dropout models can be specified in a similar way. Note that the assumed dropout models are not testable based on the observed data, so it is important to carry out sensitivity analysis based on various plausible dropout models. If the main parameter estimates are not sensitive to the assumed dropout model, we may be confident about the results. Otherwise, if the estimates are sensitive to the assumed dropout model, we need further investigation of possible missing mechanisms.

4.3 Likelihood Inference

4.3.1 The Likelihood Function

We consider likelihood inference for semiparametric NLME models with *outcome-based informative* dropouts and measurement errors and missing data in time-varying covariates, based on the approximate models (2.6) – (2.8). Let $\boldsymbol{\theta} = (\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2,\ R,\ A,\ B,\ \boldsymbol{\eta})$ be the collection of all unknown model parameters. We assume that the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, δ^2, R, A, B, and $\boldsymbol{\eta}$ are all distinct. The approximate log-likelihood for the observed data $\{(\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i),\ i = 1, \ldots, n\}$ can be written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int\int\int \Big[f_Y(\mathbf{y}_i|\mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i; \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2)\, f_Z(\mathbf{z}_i|\mathbf{a}_i; \boldsymbol{\alpha}, R)\, f(\mathbf{a}_i; A) \\ f(\mathbf{b}_i; B)\, f(\mathbf{r}_i|\mathbf{y}_i,\ \mathbf{z}_i;\ \boldsymbol{\eta}) \Big]\, d\mathbf{y}_{mis,i}\, d\mathbf{a}_i\, d\mathbf{b}_i,$$

where

$$
\begin{aligned}
f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) &= \textstyle\prod_{j=1}^{n_i} f_Y(y_{ij}|\mathbf{z}_{ij},\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \\
&= \textstyle\prod_{j=1}^{n_i} (2\pi\delta^2)^{-1/2} \exp\{-[y_{ij} - g(t_{ij},\, d(\mathbf{u}_{ij}^T\boldsymbol{\alpha} + \mathbf{v}_{ij}^T\mathbf{a}_i,\, \boldsymbol{\beta},\, \mathbf{b}_i))]^2/2\delta^2\}, \\
f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) &= \textstyle\prod_{k=1}^{m_i} f_Z(\mathbf{z}_{ik}|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
&= \textstyle\prod_{k=1}^{m_i} |2\pi R|^{-1/2} \exp\{-(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)^T\, R^{-1} \\
&\qquad \times(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)/2\}, \\
f(\mathbf{a}_i;\, A) &= |2\pi A|^{-1/2} \exp\{-\mathbf{a}_i^T A^{-1}\mathbf{a}_i/2\}, \\
f(\mathbf{b}_i;\, B) &= |2\pi B|^{-1/2} \exp\{-\mathbf{b}_i^T B^{-1}\mathbf{b}_i/2\}.
\end{aligned}
$$

The observed-data log-likelihood $l(\boldsymbol{\theta})$ can be quite intractable, so we use a MCEM algorithm, which is similar to Section 3.3.2, to find the approximate MLEs of parameters $\boldsymbol{\theta}$. By treating the unobservable random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ as additional "missing" data, we have "complete data" $\{(\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i),\, i = 1, \ldots, n\}$. The complete-data log-likelihood for all individuals can then be expressed as

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_c^{(i)}(\boldsymbol{\theta}) \;\equiv\; & \sum_{i=1}^{n} \{\log f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i; \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
& + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B) + \log f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\eta})\}. \qquad (4.1)
\end{aligned}
$$

where $l_c^{(i)}$ is the complete-data log-likelihood for individual i.

4.3.2 Approximate MLEs Based on a MCEM Method

The EM algorithm iterates between an E-step, which computes the conditional expectation of the complete-data log-likelihood given the observed data, and a M-step, which maximizes this conditional expectation to update parameter estimates. For our models, the E-step is quite intractable due to nonlinearity, so we use Monte Carlo methods to approximate the intractable conditional expectation. In the M-step, we use standard complete-data optimization procedures to update parameter estimates.

Let $\boldsymbol{\theta}^{(t)}$ be the parameter estimates from the t-th EM iteration. The E-step for individual i at the $(t+1)$th EM iteration can be expressed as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E(l_c^{(i)}(\boldsymbol{\theta})|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)}) = \int\int\int \Big[\log f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \\ &\quad + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B) \\ &\quad + \log f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\eta})\Big] \times f(\mathbf{y}_{mis,i},\, \mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)})\, d\mathbf{y}_{mis,i}\, d\mathbf{a}_i\, d\mathbf{b}_i \\ &\equiv I_1^{(i)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + I_2^{(i)}(\boldsymbol{\alpha},\, R) + I_3^{(i)}(A) + I_4^{(i)}(B) + I_5^{(i)}(\boldsymbol{\eta}). \end{aligned} \tag{4.2}$$

Since the expression (4.2) is an expectation with respect to $f(\mathbf{y}_{mis,i},\, \mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)})$, it can be evaluated using the MCEM algorithm (Wei and Tanner, 1990). Specifically, we may use the Gibbs sampler to generate samples from $[\mathbf{y}_{mis,i},\, \mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)}]$ by iteratively sampling from the full conditionals $[\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}]$, $[\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}]$, and $[\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}]$ as follows.

$$\begin{aligned} f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}), \\ f(\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f_Y(\mathbf{y}_i,\, \mathbf{a}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{a}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{a}_i|\mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)})\, f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}), \\ f(\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f_Y(\mathbf{y}_i,\, \mathbf{b}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{b}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{y}_i,\, \mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f_Y(\mathbf{y}_i|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}). \end{aligned} \tag{4.3}$$

Monte Carlo samples from each of the full conditionals can be generated using rejection

sampling methods, as in Wu (2004). Alternatively, the integral (4.2) may be evaluated using the importance sampling method (see Section 3.3.3).

For individual i, let $\{(\tilde{\mathbf{y}}_{mis,i}^{(1)},\ \tilde{\mathbf{a}}_i^{(1)},\ \tilde{\mathbf{b}}_i^{(1)}),\ldots,(\tilde{\mathbf{y}}_{mis,i}^{(k_t)},\ \tilde{\mathbf{a}}_i^{(k_t)},\ \tilde{\mathbf{b}}_i^{(k_t)})\}$ denote a random sample of size k_t generated from $[\mathbf{y}_{mis,i},\ \mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \boldsymbol{\theta}^{(t)}]$. Note that each $(\tilde{\mathbf{y}}_{mis,i}^{(k)},\ \tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)})$ depends on the EM iteration number t, which is suppressed throughout. The E-step at the $(t+1)$th EM iteration can then be expressed as

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{i=1}^{n}\left\{\tfrac{1}{k_t}\sum_{k=1}^{k_t} l_c^{(i)}(\boldsymbol{\theta};\ (\tilde{\mathbf{y}}_{mis,i}^{(k)},\ \mathbf{y}_{obs,i}),\ \mathbf{z}_i,\ \mathbf{r}_i,\ \tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)})\right\} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{k_t}\tfrac{1}{k_t}\log f_Y((\tilde{\mathbf{y}}_{mis,i}^{(k)},\mathbf{y}_{obs,i})|\mathbf{z}_i,\ \tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)};\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{k_t}\tfrac{1}{k_t}\log f_Z(\mathbf{z}_i|\tilde{\mathbf{a}}_i^{(k)};\ \boldsymbol{\alpha},\ R) + \sum_{i=1}^{n}\sum_{k=1}^{k_t}\tfrac{1}{k_t}\log f(\tilde{\mathbf{a}}_i^{(k)};\ A) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{k_t}\tfrac{1}{k_t}\log f(\tilde{\mathbf{b}}_i^{(k)};\ B) + \sum_{i=1}^{n}\sum_{k=1}^{k_t}\tfrac{1}{k_t}\log f(\mathbf{r}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)},\ \mathbf{y}_{obs,i}),\ \mathbf{z}_i;\ \boldsymbol{\eta}) \\
&\equiv Q^{(1)}(\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2|\boldsymbol{\theta}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha},\ R|\boldsymbol{\theta}^{(t)}) + Q^{(3)}(A|\boldsymbol{\theta}^{(t)}) + Q^{(4)}(B|\boldsymbol{\theta}^{(t)}) + Q^{(5)}(\boldsymbol{\eta}|\boldsymbol{\theta}^{(t)}).
\end{aligned} \tag{4.4}$$

The M-step then maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to produce an updated estimate $\boldsymbol{\theta}^{(t+1)}$, so it is like a complete-data maximization. Since the parameters in $Q^{(1)}+Q^{(2)}$, $Q^{(3)}$, $Q^{(4)}$, and $Q^{(5)}$ are distinct, the M-step can be implemented by maximizing $Q^{(1)}+Q^{(2)}$, $Q^{(3)}$, $Q^{(4)}$, and $Q^{(5)}$ separately using standard optimization procedures for the corresponding complete-data models, as in Section 3.3.2.

As in Section 3.3.2, we use the approximate formula suggested by McLachlan and Krishnan (1997) to obtain the variance-covariance matrix of the approximate MLE $\hat{\boldsymbol{\theta}}$. Let $\mathbf{s}_c^{(i)} = \partial l_c^{(i)}/\partial\boldsymbol{\theta}$, where $l_c^{(i)}$ is the complete-data log-likelihood for individual i. Then an approximate formula for the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \Big[\sum_{i=1}^{n} E(\mathbf{s}_c^{(i)}|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \hat{\boldsymbol{\theta}})\ E(\mathbf{s}_c^{(i)}|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \hat{\boldsymbol{\theta}})^T\Big]^{-1}, \tag{4.5}$$

where the expectations can again be approximated by Monte Carlo empirical means, as in (4.4).

In summary, the foregoing MCEM algorithm proceeds as follows.

Step 1. Obtain an initial estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, R, A, B) = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \delta^{2(0)}, R^{(0)}, A^{(0)}, B^{(0)})$ based on a naive method, which ignores covariate measurement errors and missing data, and an initial estimate of $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$ based on the dropout model by filling in $\mathbf{y}_{mis,i}$ with the average of $\mathbf{y}_{obs,i}$. Set $\mathbf{a}_i^{(0)} = \mathbf{0}$ and $\mathbf{b}_i^{(0)} = \mathbf{0}$.

Step 2. At the t-th iteration, obtain Monte Carlo samples of the "missing data" $(\mathbf{y}_{mis,i}, \mathbf{a}_i, \mathbf{b}_i)$ using the Gibbs sampler along with rejection sampling methods by sampling from the full conditionals $[\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)}]$, $[\mathbf{a}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{r}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)}]$ and $[\mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{r}_i, \mathbf{a}_i; \boldsymbol{\theta}^{(t)}]$, or using importance sampling methods to approximate the conditional expectation in the E-step.

Step 3. Update estimates $\boldsymbol{\theta}^{(t+1)}$ using standard complete-data optimization procedures.

Step 4. Iterate between Step 2 and Step 3 until convergence.

4.3.3 Monte Carlo Sampling

Gibbs Sampler

As in Section 3.3.3, we can again use the Gibbs sampler to draw the desired samples as follows. Set initial values $(\tilde{\mathbf{y}}_{mis,i}^{(0)}, \tilde{\mathbf{a}}_i^{(0)}, \tilde{\mathbf{b}}_i^{(0)})$. Suppose that the current generated values are $(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)})$, we can obtain $(\tilde{\mathbf{y}}_{mis,i}^{(k+1)}, \tilde{\mathbf{a}}_i^{(k+1)}, \tilde{\mathbf{b}}_i^{(k+1)})$ as follows.

Step 1. Draw a sample for the "missing" response $\tilde{\mathbf{y}}_{mis,i}^{(k+1)}$ from

$$f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\theta}^{(t)}).$$

Step 2. Draw a sample for the "missing" random effects $\tilde{\mathbf{a}}_i^{(k+1)}$ from

$$f(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k+1)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\theta}^{(t)}).$$

Step 3. Draw a sample for the "missing" random effects $\tilde{\mathbf{b}}_i^{(k+1)}$ from

$$f(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k+1)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i, \tilde{\mathbf{a}}_i^{(k+1)}; \boldsymbol{\theta}^{(t)}).$$

After a sufficiently large burn-in of r iterations of Steps 1 – 3, the sampled values will achieve a steady state. Then, $\{(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}),\ k = r+1, \ldots, r+k_t\}$ can be treated as samples from the multidimensional density function

$$f(\mathbf{y}_{mis,i}, \mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i; \boldsymbol{\theta}^{(t)}).$$

And, if we choose a sufficiently large gap r' (say $r' = 10$), we can treat the sample series $\{(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}), k = r+r', r+2r', \ldots\}$ as independent samples from the multidimensional density function. There are several ways to get the initial values $(\tilde{\mathbf{y}}_{mis,i}^{(0)}, \tilde{\mathbf{a}}_i^{(0)}, \tilde{\mathbf{b}}_i^{(0)})$. A simple way is to set $\tilde{\mathbf{y}}_{mis,i}^{(0)}$ to the average of $\mathbf{y}_{obs,i}$, and $(\tilde{\mathbf{a}}_i^{(0)}, \tilde{\mathbf{b}}_i^{(0)})$ to $(\mathbf{0}, \mathbf{0})$.

Multivariate Rejection Algorithm

Sampling from the three full conditionals can be accomplished by the multivariate rejection sampling method. For example, consider sampling from $f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ in (4.3). Let $f^*(\mathbf{y}_{mis,i}) = f(\mathbf{y}_{obs,i}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})\, f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(t)})$ and $\varsigma = \sup_{\mathbf{u}}\{f^*(\mathbf{u})\}$. We assume $\varsigma < \infty$. A random sample from $f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ can then be obtained as follows by multivariate rejection sampling:

Step 1. Sample $\mathbf{y}_{mis,i}^*$ from $f_Y(\mathbf{y}_{mis,i}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$, and independently, sample w from the uniform (0, 1) distribution.

Step 2. If $w \leq f^*(\mathbf{y}_{mis,i}^*)/\varsigma$, then accept $\mathbf{y}_{mis,i}^*$, otherwise, go back to step 1.

Samples from $f(\mathbf{a}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{r}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(t)})$ and $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{r}_i, \mathbf{a}_i; \boldsymbol{\theta}^{(t)})$ can be obtained in a similar way. Therefore, the Gibbs sampler in conjunction with the multivariate rejection sampling can be used to obtain samples from $[\mathbf{y}_{mis,i}, \mathbf{a}_i, \mathbf{b}_i|\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i; \boldsymbol{\theta}^{(t)}]$.

4.4 A Computationally More Efficient Approximate Method

The estimation method described in the previous section may be computationally very intensive and may even offer potential computational problems such as slow or non-convergence, since the method involves sampling the random effects $(\mathbf{a}_i,\ \mathbf{b}_i)$, which may have high dimension (see the detailed discussion in Section 3.4.1). To overcome these difficulties, in this section we propose an alternative approximate method which further approximates model (2.5) by taking a first-order Taylor expansion around the current parameter and random effects estimates, which leads to a LME response model. For the resulting LME response model with the covariate model (2.8) and the dropout model, we can obtain approximate MLEs of model parameters by using a computationally more efficient MCEM algorithm. The random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ can be integrated out in the E-step of the EM algorithm and thus sampling the random effects in the E-step is no longer needed. Therefore, the proposed approximate method may provide substantial computational advantages over the MCEM method in the previous section. Moreover, the proposed method can be used to obtain good parameter starting values for the MCEM method in the previous section.

Denote the current estimates of $(\boldsymbol{\theta},\ \mathbf{a}_i,\ \mathbf{b}_i)$ by $(\tilde{\boldsymbol{\theta}},\ \tilde{\mathbf{a}}_i,\ \tilde{\mathbf{b}}_i)$, where $\tilde{\mathbf{a}}_i = E(\mathbf{a}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}})$ and $\tilde{\mathbf{b}}_i = E(\mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}})$, suppressing the EM iteration number. Taking a first-order Taylor expansion of g_{ij} in (3.6) around the current parameter estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ and random effects estimates $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{b}}_i$, we obtain the following LME response model

$$\mathbf{y}_i \ = \ \tilde{\mathbf{g}}_i + W_i\,\boldsymbol{\alpha} + X_i\,\boldsymbol{\beta} + H_i\,\mathbf{a}_i + T_i\,\mathbf{b}_i + \mathbf{e}_i, \tag{4.6}$$

where

$$
\begin{aligned}
W_i &= (\mathbf{w}_{i1}, \ldots, \mathbf{w}_{in_i})^T \text{ with } \mathbf{w}_{ij} = \frac{\partial g_{ij}}{\partial \boldsymbol{\alpha}} \\
X_i &= (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^T \text{ with } \mathbf{x}_{ij} = \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}} \\
H_i &= (\mathbf{h}_{i1}, \ldots, \mathbf{h}_{in_i})^T \text{ with } \mathbf{h}_{ij} = \frac{\partial g_{ij}}{\partial \mathbf{a}_i} \\
T_i &= (\mathbf{t}_{i1}, \ldots, \mathbf{t}_{in_i})^T \text{ with } \mathbf{t}_{ij} = \frac{\partial g_{ij}}{\partial \mathbf{b}_i} \\
\tilde{\mathbf{g}}_i &= \mathbf{g}_i(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{a}}_i, \tilde{\mathbf{b}}_i) - W_i \tilde{\boldsymbol{\alpha}} - X_i \tilde{\boldsymbol{\beta}} - H_i \tilde{\mathbf{a}}_i - T_i \tilde{\mathbf{b}}_i,
\end{aligned}
$$

with all the partial derivatives being evaluated at $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{a}}_i, \tilde{\mathbf{b}}_i)$.

Our proposed approximate method is to iteratively solve LME response model (4.6). For the LME response model (4.6) with the covariate model (2.8) and the dropout model, the MLEs of the model parameters can be obtained by a MCEM algorithm, in which the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ can be integrated out in the E-step, as shown below. Thus, the E-step only involves sampling $\mathbf{y}_{mis,i}$ rather than $(\mathbf{y}_{mis,i}, \mathbf{a}_i, \mathbf{b}_i)$ as in Section 4.2. Moreover, some analytic expressions for the M-step can be obtained. These result in a substantial improvement in computational efficiency.

4.4.1 A Much Simpler E-step

In this section, we show that, based on the approximate LME response model (4.6) and the covariate measurement error model (2.8) along with the dropout model, the random effects $(\mathbf{a}_i, \mathbf{b}_i)$ in the E-step in Section 4.2 can be integrated out.

Based on standard results for multivariate normal distributions, the distribution of

$(\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i)$ for individual i is given by

$$\left[\left(\begin{array}{c}\mathbf{y}_i\\ \mathbf{z}_i\\ \mathbf{a}_i\\ \mathbf{b}_i\end{array}\right);\tilde{\boldsymbol{\theta}}\right] \sim N\left[\left(\begin{array}{c}\tilde{\mathbf{g}}_i + W_i\,\boldsymbol{\alpha} + X_i\,\boldsymbol{\beta}\\ U_i\,\boldsymbol{\alpha}\\ \mathbf{0}\\ \mathbf{0}\end{array}\right), \left(\begin{array}{cccc} H_i\,A\,H_i^T + T_i\,B\,T_i^T + \delta^2\,I & H_i\,A\,V_i^T & H_i\,A & T_i\,B\\ V_i\,A\,H_i^T & V_i\,A\,V_i^T + I \bigotimes R & V_i\,A & 0\\ A\,H_i^T & A\,V_i^T & A & 0\\ B\,T_i^T & 0 & 0 & B\end{array}\right)\right].$$

Since $H_i\,A\,H_i^T + T_i\,B\,T_i^T + \delta^2\,I$ and $V_i\,A\,V_i^T + I \bigotimes R$ are positive definite, they are symmetric and invertible. By the inverse operation of partition matrices, we can write

$$\left(\begin{array}{cc} H_i\,A\,H_i^T + T_i\,B\,T_i^T + \delta^2\,I & H_i\,A\,V_i^T\\ V_i\,A\,H_i^T & V_i\,A\,V_i^T + I \bigotimes R\end{array}\right)^{-1} = \left(\begin{array}{cc} G_i + F_i\,E_i\,F_i^T & -F_i\,E_i\\ -E_i\,F^T & E_i\end{array}\right), \qquad (4.7)$$

where

$$\begin{aligned} G_i &= (H_i\,A\,H_i^T + T_i\,B\,T_i^T + \delta^2\,I)^{-1},\\ F_i &= G_i\,H_i\,A\,V_i^T,\\ E_i &= [(V_i\,A\,V_i^T + I \otimes R) - V_i\,A\,H_i^T\,G_i\,H_i\,A\,V_i^T]^{-1}. \end{aligned}$$

Because the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ are conditionally independent of $\mathbf{r}_i$ given $(\mathbf{y}_i, \mathbf{z}_i)$, it is straightforward to obtain the conditional distribution of $[\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_i,\ \mathbf{z}_i, \mathbf{r}_i]$ when $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$

$$\left[\left(\begin{array}{c}\mathbf{a}_i\\ \mathbf{b}_i\end{array}\right)\middle|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}\right] \sim N\left[\left(\begin{array}{c}\boldsymbol{\nu}_{\mathbf{a}_i}\\ \boldsymbol{\nu}_{\mathbf{b}_i}\end{array}\right), \left(\begin{array}{cc}\Sigma_{\mathbf{a}_i} & \Sigma_{\mathbf{a}_i\mathbf{b}_i}\\ \Sigma_{\mathbf{b}_i\mathbf{a}_i} & \Sigma_{\mathbf{b}_i}\end{array}\right)\right],$$

where

$$
\begin{aligned}
\boldsymbol{\nu}_{\mathbf{a}_i} &= [A\,H_i^T(G_i + F_i\,E_i\,F_i^T) - A\,V_i^T\,E_i\,F_i^T](\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) \\
&\quad + [A\,V_i^T\,E_i - A\,H_i^T\,F_i\,E_i](\mathbf{z}_i - U_i\,\boldsymbol{\alpha})\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\
\boldsymbol{\nu}_{\mathbf{b}_i} &= B\,T_i^T(G_i + F_i\,E_i\,F_i^T)(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) - B\,T_i^T\,F_i\,E_i(\mathbf{z}_i - U_i\,\boldsymbol{\alpha})\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\
\Sigma_{\mathbf{a}_i} &= A - [A\,H_i^T(G_i + F_i\,E_i\,F_i^T) - A\,V_i^T\,E_i\,F_i^T]\,H_i\,A - [A\,V_i^T\,E_i - A\,H_i^T\,F_i\,E_i]\,V_i\,A\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\
\Sigma_{\mathbf{b}_i} &= B - B\,T_i^T(G_i + F_i\,E_i\,F_i^T)\,T_i\,B\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\
\Sigma_{\mathbf{a}_i\mathbf{b}_i} &= (\Sigma_{\mathbf{b}_i\mathbf{a}_i})^T = [A\,V_i^T\,E_i\,F_i^T - A\,H_i^T\,(G_i + F_i\,E_i\,F_i^T)]\,T_i\,B\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}.
\end{aligned}
$$

By the expectation and covariance properties for multivariate random variables, we have

$$
\begin{aligned}
&E(\mathbf{a}_i|\mathbf{y}_i,\ \mathbf{z}_i\,\mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = \boldsymbol{\nu}_{\mathbf{a}_i}, \\
&E(\mathbf{a}_i\,\mathbf{a}_i^T|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = \Sigma_{\mathbf{a}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{a}_i}^T, \\
&E(\mathbf{b}_i|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = \boldsymbol{\nu}_{\mathbf{b}_i}, \\
&E(\mathbf{b}_i\,\mathbf{b}_i^T|\mathbf{y}_i,\ \mathbf{z}_i,\mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = \Sigma_{\mathbf{b}_i} + \boldsymbol{\nu}_{\mathbf{b}_i}\,\boldsymbol{\nu}_{\mathbf{b}_i}^T, \\
&E(\mathbf{a}_i\,\mathbf{b}_i^T|\mathbf{y}_i,\ \mathbf{z}_i,\mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = \Sigma_{\mathbf{a}_i\mathbf{b}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{b}_i}^T.
\end{aligned}
$$

Since

$$
f(\mathbf{y}_{mis,i},\ \mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}) = f(\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_i,\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}})\,f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \tilde{\boldsymbol{\theta}}),
$$

the conditional expectations in the E-step, which correspond to $I_1^{(i)} - I_5^{(i)}$ in (4.2), become

$$
\begin{aligned}
\tilde{I}_1^{(i)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \;=\; & \int E_{\mathbf{a}_i,\mathbf{b}_i}\Big[-\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta} - H_i\,\mathbf{a}_i - T_i\,\mathbf{b}_i)^T \\
& \times(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta} - H_i\,\mathbf{a}_i - T_i\,\mathbf{b}_i)|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}\Big] \\
& \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
= \; & -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\int \Big\{(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) \\
& -2(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\, E_{\mathbf{a}_i,\mathbf{b}_i}[H_i\,\mathbf{a}_i + T_i\,\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}] \\
& +E_{\mathbf{a}_i,\mathbf{b}_i}[(H_i\,\mathbf{a}_i + T_i\,\mathbf{b}_i)^T\,(H_i\,\mathbf{a}_i + T_i\,\mathbf{b}_i)|\mathbf{y}_i,\, \mathbf{z}_i, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}]\Big\} \\
& \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
= \; & -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\int \Big\{(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) \\
& -2(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\,[H_i\,E(\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) + T_i\,E(\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})] \\
& +E_{\mathbf{a}_i,\mathbf{b}_i}[\mathbf{a}_i^T\,H_i^T\,H_i\,\mathbf{a}_i + 2\mathbf{b}_i^T\,T_i^T\,H_i\,\mathbf{a}_i + \mathbf{b}_i^T\,T_i^T\,T_i\,\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}]\Big\} \\
& \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
= \; & -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\int \Big\{(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\,(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) \\
& -2(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\,(H_i\,E[\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}] + T_i\,E[\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}]) \\
& +\mathrm{tr}[H_i\,E(\mathbf{a}_i\,\mathbf{a}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\,H_i^T] + 2\,\mathrm{tr}[H_i\,E(\mathbf{a}_i\,\mathbf{b}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\,T_i^T] \\
& +\mathrm{tr}[T_i\,E(\mathbf{b}_i\,\mathbf{b}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\,T_i^T]\Big\} \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
= \; & -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\int \Big\{(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\,(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) \\
& -2(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})^T\,[H_i\,\boldsymbol{\nu}_{\mathbf{a}_i} + T_i\,\boldsymbol{\nu}_{\mathbf{b}_i}] \\
& +\mathrm{tr}[H_i\,(\Sigma_{\mathbf{a}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{a}_i}^T)\,H_i^T] + 2\mathrm{tr}[H_i\,(\Sigma_{\mathbf{a}_i\mathbf{b}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{b}_i}^T)\,T_i^T] \\
& +\mathrm{tr}[T_i\,(\Sigma_{\mathbf{b}_i} + \boldsymbol{\nu}_{\mathbf{b}_i}\,\boldsymbol{\nu}_{\mathbf{b}_i}^T)\,T_i^T]\Big\} \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i},
\end{aligned}
$$

$$
\begin{aligned}
\tilde{I}_2^{(i)}(\boldsymbol{\alpha},\, R) &= \int E_{\mathbf{a}_i,\mathbf{b}_i}\Big[-\frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha} - V_{ij}\,\mathbf{a}_i)^T\, R^{-1} \\
&\qquad \times(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha} - V_{ij}\,\mathbf{a}_i)|\mathbf{y}_i,\, \mathbf{z}_i, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}\Big] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\int \Big\{ \sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\,(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha}) \\
&\qquad -2\sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\, V_{ij}\, E_{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) \\
&\qquad +\sum_{j=1}^{m_i} E_{\mathbf{a}_i,\mathbf{b}_i}[\mathbf{a}_i^T\, V_{ij}^T\, R^{-1}\, V_{ij}\,\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}]\Big\} \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\int \Big\{ \sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\,(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha}) \\
&\qquad -2\sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\, V_{ij}\, E(\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) \\
&\qquad +\sum_{j=1}^{m_i} \mathrm{tr}[V_{ij}^T\, R^{-1}\, V_{ij}\, E(\mathbf{a}_i\,\mathbf{a}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})]\Big\} \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\Big\{ \sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\,(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha}) \\
&\qquad \int \Big[-2\sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\, R^{-1}\, V_{ij}\,\boldsymbol{\nu}_{\mathbf{a}_i} \\
&\qquad +\sum_{j=1}^{m_i} \mathrm{tr}(V_{ij}^T\, R^{-1}\, V_{ij}\,(\Sigma_{\mathbf{a}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{a}_i}^T))\Big] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}\Big\}, \\
\tilde{I}_3^{(i)}(A) &= \int E_{\mathbf{a}_i,\mathbf{b}_i}\Big[-\frac{1}{2}\log|2\pi A| - \frac{1}{2}\mathbf{a}_i^T\, A^{-1}\,\mathbf{a}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}\Big] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{1}{2}\log|2\pi A| - \frac{1}{2}\int \mathrm{tr}[A^{-1}\, E(\mathbf{a}_i\,\mathbf{a}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{1}{2}\log|2\pi A| - \frac{1}{2}\int \mathrm{tr}[A^{-1}\,(\Sigma_{\mathbf{a}_i} + \boldsymbol{\nu}_{\mathbf{a}_i}\,\boldsymbol{\nu}_{\mathbf{a}_i}^T)] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}, \\
\tilde{I}_4^{(i)}(B) &= \int E_{\mathbf{a}_i,\mathbf{b}_i}\Big[-\frac{1}{2}\log|2\pi B| - \frac{1}{2}\mathbf{b}_i^T\, B^{-1}\,\mathbf{b}_i|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}\Big] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{1}{2}\log|2\pi B| - \frac{1}{2}\int \mathrm{tr}[B^{-1}\, E(\mathbf{b}_i\,\mathbf{b}_i^T|\mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i} \\
&= -\frac{1}{2}\log|2\pi B| - \frac{1}{2}\int \mathrm{tr}[B^{-1}\,(\Sigma_{\mathbf{b}_i} + \boldsymbol{\nu}_{\mathbf{b}_i}\,\boldsymbol{\nu}_{\mathbf{b}_i}^T)] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i},
\end{aligned}
$$

$$\begin{aligned}\tilde{I}_5^{(i)}(\boldsymbol{\eta}) &= \int E_{\mathbf{a}_i,\mathbf{b}_i}\Big[\log f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\boldsymbol{\eta})\Big] \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}})\,d\mathbf{y}_{mis,i} \\ &= \int \log f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\boldsymbol{\eta}) \times f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}})\,d\mathbf{y}_{mis,i}.\end{aligned}$$

Thus, compared with the intractable high-dimensional integrals in (4.2), the above integrals $\tilde{I}_1^{(i)} - \tilde{I}_5^{(i)}$ have much lower dimension, i.e., the E-step is substantially simplified. Note that $\tilde{I}_1^{(i)} - \tilde{I}_5^{(i)}$ are expectations with respect to $f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}})$, so they may be evaluated using common numerical integration methods such as Gaussian quadrature if the dimension of $\mathbf{y}_{mis,i}$ is small. If the dimension of $\mathbf{y}_{mis,i}$ is not small, we can use the rejection sampling methods to generate samples from $[\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}]$ based on the conditional density

$$\begin{aligned}f(\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}) &\propto f(\mathbf{y}_i,\,\mathbf{z}_i,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}) \\ &= f(\mathbf{y}_i,\,\mathbf{z}_i;\,\tilde{\boldsymbol{\theta}})f(\mathbf{r}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\tilde{\boldsymbol{\theta}}),\end{aligned}$$

where the distribution of $(\mathbf{y}_i,\,\mathbf{z}_i)$ is

$$\begin{pmatrix}\mathbf{y}_i \\ \mathbf{z}_i\end{pmatrix} \sim N\left[\begin{pmatrix}\tilde{\mathbf{g}}_i + W_i\boldsymbol{\alpha} + X_i\boldsymbol{\beta} \\ U_i\boldsymbol{\alpha}\end{pmatrix}, \begin{pmatrix}H_iAH_i^T + T_iBT_i^T + \delta^2 I & H_iAV_i^T \\ V_iAH_i^T & V_iAV_i^T + I \bigotimes R\end{pmatrix}\right].$$

For individual i, let $\tilde{\mathbf{y}}_{mis,i}^{(1)}, \ldots, \tilde{\mathbf{y}}_{mis,i}^{(k_t)}$ denote a random sample of size k_t generated from $[\mathbf{y}_{mis,i}|\mathbf{y}_{obs,i},\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}]$. Note that each $\tilde{\mathbf{y}}_{mis,i}^{(k)}$ depends on the EM iteration number t, which is suppressed throughout. Then we approximate the conditional expectations $\tilde{I}_1^{(i)} - \tilde{I}_5^{(i)}$ in

the E-step by its empirical mean, with missing data replaced by simulated values, as follows.

$$
\begin{aligned}
\tilde{I}_1^{(i)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) \;\approx\; & -\frac{n_i}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\Big\{\frac{1}{k_t}\sum_{k=1}^{k_t}[(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}) - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}]^T \\
& \times[(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}) - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}] \\
& -\frac{1}{k_t}\sum_{i=1}^{k_t} 2[(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}) - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}]^T \\
& \times[H_i\,E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) + T_i\,E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})] \\
& +\frac{1}{k_t}\sum_{k=1}^{k_t}\mathrm{tr}\{H_i\,[\mathrm{Cov}(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\,E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\,H_i^T\} \\
& +\frac{1}{k_t}\sum_{k=1}^{k_t} 2\mathrm{tr}\{H_i\,[\mathrm{Cov}(\mathbf{a}_i, \mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\,E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\,T_i^T\} \\
& +\frac{1}{k_t}\sum_{k=1}^{k_t}\mathrm{tr}\{T_i\,[\mathrm{Cov}(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\,E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\,T_i^T\}\Big\},
\end{aligned}
$$

$$
\begin{aligned}
\tilde{I}_2^{(i)}(\boldsymbol{\alpha}, R) \;\approx\; & -\frac{m_i}{2}\log|2\pi R| - \frac{1}{2}\Bigg\{\sum_{j=1}^{m_i}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\,R^{-1}\,(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha}) \\
& -2\sum_{j=1}^{m_i}\frac{1}{k_t}\sum_{k=1}^{k_t}(\mathbf{z}_{ij} - U_{ij}\,\boldsymbol{\alpha})^T\,R^{-1}\,V_{ij}\,E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +\sum_{j=1}^{m_i}\frac{1}{k_t}\sum_{k=1}^{k_t}\mathrm{tr}\{V_{ij}^T\,R^{-1}\,V_{ij}\,[\mathrm{Cov}(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\,E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\}\Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\tilde{I}_3^{(i)}(A) \;\approx\; & -\frac{1}{2}\log|2\pi A| - \frac{1}{k_t}\sum_{k=1}^{k_t}\frac{1}{2}\mathrm{tr}\{A^{-1}\,[\mathrm{Cov}(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) \\
& +E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\,E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\},
\end{aligned}
$$

$$\tilde{I}_4^{(i)}(B) \approx -\frac{1}{2}\log|2\pi B| - \frac{1}{k_t}\sum_{k=1}^{k_t}\frac{1}{2}\mathrm{tr}\{B^{-1}[\mathrm{Cov}(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}}) + E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})\, E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i, \mathbf{r}_i; \tilde{\boldsymbol{\theta}})^T]\},$$

$$\tilde{I}_5^{(i)}(\boldsymbol{\eta}) \approx \frac{1}{k_t}\sum_{k=1}^{k_t}\log f(\mathbf{r}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{y}_{obs,i}), \mathbf{z}_i; \boldsymbol{\eta}).$$

4.4.2 The M-step

In this section, we derive some closed-form expressions for the M-step, so we avoid some iterative algorithms, which may be computationally inefficient.

The M-step of the EM algorithm maximizes

$$\tilde{Q}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n}[\tilde{I}_1^{(i)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) + \tilde{I}_2^{(i)}(\boldsymbol{\alpha}, R) + \tilde{I}_3^{(i)}(A) + \tilde{I}_4^{(i)}(B) + \tilde{I}_5^{(i)}(\boldsymbol{\eta})]$$

to produce an updated estimate of $\boldsymbol{\theta}$. Note that if we were to observe $(\mathbf{e}_i, \boldsymbol{\epsilon}_i, \mathbf{a}_i, \mathbf{b}_i)$, in addition to $(\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i)$, we would use the following estimates

$$\hat{\delta}^2 = \sum_{i=1}^{n}\mathbf{e}_i^T\mathbf{e}_i/\sum_{i=1}^{n}n_i, \quad \hat{R} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{\epsilon}_{ij}\boldsymbol{\epsilon}_{ij}^T/\sum_{i=1}^{n}m_i,$$

$$\hat{A} = \sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^T/n, \qquad \hat{B} = \sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T/n,$$

where $\sum_{i=1}^{n}\mathbf{e}_i^T\mathbf{e}_i$, $\sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{\epsilon}_{ij}\boldsymbol{\epsilon}_{ij}^T$, $\sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^T$, and $\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T$ are "sufficient" statistics for δ^2, R, A, and B, respectively. Since $\mathbf{e}_i$, $\boldsymbol{\epsilon}_i$, $\mathbf{a}_i$, and $\mathbf{b}_i$ are unobservable, we can "estimate" them by their conditional expectations given the observed data $(\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i)$, as in Laird and Ware (1982). Based on results in multivariate analysis, we have

$$\left[\begin{pmatrix}\mathbf{y}_i\\ \mathbf{z}_i\\ \mathbf{e}_i\end{pmatrix};\tilde{\boldsymbol{\theta}}\right] \sim N\left[\begin{pmatrix}\tilde{\mathbf{g}}_i + W_i\boldsymbol{\alpha} + X_i\boldsymbol{\beta}\\ U_i\boldsymbol{\alpha}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}H_i A H_i^T + T_i B T_i^T + \delta^2 I & H_i A V_i^T & \delta^2 I\\ V_i A H_i^T & V_i A V_i^T + I \otimes R & 0\\ \delta^2 I & 0 & \delta^2 I\end{pmatrix}\right],$$

and

$$\left[\begin{pmatrix} \mathbf{y}_i \\ \mathbf{z}_i \\ \boldsymbol{\epsilon}_{ij} \end{pmatrix}; \tilde{\boldsymbol{\theta}}\right] \sim N\left[\begin{pmatrix} \tilde{\mathbf{g}}_i + W_i\,\boldsymbol{\alpha} + X_i\,\boldsymbol{\beta} \\ U_i\,\boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} H_i\,A\,H_i^T + T_i\,B\,T_i^T + \delta^2\,I & H_i\,A\,V_i^T & 0 \\ V_i\,A\,H_i^T & V_i\,A\,V_i^T + I \bigotimes R & M_{ij} \\ 0 & M_{ij}^T & R \end{pmatrix}\right],$$

where M_{ij} is a $\nu m_i \times \nu$ matrix with the jth $\nu \times \nu$ submatrix R and zeros elsewhere. Since $\mathbf{e}_i$ and $\boldsymbol{\epsilon}_{ij}$ are conditionally independent of $\mathbf{r}_i$ given $\mathbf{y}_i$ and $\mathbf{z}_i$, we have

$$\begin{aligned} f(\mathbf{e}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}) &= f(\mathbf{e}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\tilde{\boldsymbol{\theta}}), \\ f(\boldsymbol{\epsilon}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}) &= f(\boldsymbol{\epsilon}_i|\mathbf{y}_i,\,\mathbf{z}_i;\,\tilde{\boldsymbol{\theta}}). \end{aligned}$$

Using the above results, the inverse of partition matrix in (4.7), and the definition of conditional distributions, it can be shown that

$$[\mathbf{e}_i|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}] \sim N(\boldsymbol{\nu}_{\mathbf{e}_i},\,\Lambda_{\mathbf{e}_i}),$$

where

$$\begin{aligned} \boldsymbol{\nu}_{\mathbf{e}_i} &= [\delta^2(G_i + F_i\,E_i\,F_i^T)(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta}) - \delta^2\,F_i\,E_i(\mathbf{z}_i - U_i\,\boldsymbol{\alpha})]\,\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\ \Lambda_{\mathbf{e}_i} &= \delta^2[I - \delta^2(G_i + F_i\,E_i\,F_i^T)]\,\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \end{aligned}$$

and

$$[\boldsymbol{\epsilon}_{ij}|\mathbf{y}_i,\,\mathbf{z}_i,\,\mathbf{r}_i;\,\tilde{\boldsymbol{\theta}}] \sim N(\boldsymbol{\nu}_{\boldsymbol{\epsilon}_{ij}},\,\Lambda_{\boldsymbol{\epsilon}_{ij}}),$$

where

$$\begin{aligned} \boldsymbol{\nu}_{\boldsymbol{\epsilon}_{ij}} &= [R\sum_{k=1}^{m_i}(E_i)^{jk}\,(\mathbf{z}_{ik} - U_{ik}\boldsymbol{\alpha}) - R\,(E_i\,F_i^T)^j\,(\mathbf{y}_i - \tilde{\mathbf{g}}_i - W_i\,\boldsymbol{\alpha} - X_i\,\boldsymbol{\beta})]\,\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \\ \Lambda_{\boldsymbol{\epsilon}_{ij}} &= [R - R\,(E_i)^{jj}\,R]\,\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}, \end{aligned}$$

with the jth $\nu \times n_i$ submatrix $(E_i\, F_i^T)^j$ of $E_i\, F_i^T$ and the jth row and the kth column $\nu \times \nu$ submatrix $(E_i)^{jk}$ of E_i, $j,\, k = 1, \ldots, m_i$. Note that

$$
\begin{aligned}
E(\mathbf{e}_i^T\, \mathbf{e}_i | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &= \int E(\mathbf{e}_i^T\, \mathbf{e}_i | \mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}, \\
E(\boldsymbol{\epsilon}_{ij}\, \boldsymbol{\epsilon}_{ij}^T | \mathbf{y}_{obs,i}, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &= \int E(\boldsymbol{\epsilon}_i\, \boldsymbol{\epsilon}_i^T | \mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}, \\
E(\mathbf{a}_i\, \mathbf{a}_i^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &= \int E(\mathbf{a}_i\, \mathbf{a}_i^T | \mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}, \\
E(\mathbf{b}_i\, \mathbf{b}_i^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &= \int E(\mathbf{b}_i\, \mathbf{b}_i^T | \mathbf{y}_i,\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})\, d\mathbf{y}_{mis,i}.
\end{aligned}
$$

Using the expectation and covariance properties for multivariate random variables and the properties of the matrix trace operation, we can update the estimates of $(\delta^2,\, R,\, A,\, B)$ in the M-step as follows.

$$
\begin{aligned}
\hat{\delta}^2 &= \sum_{i=1}^{n} \mathrm{tr}[E(\mathbf{e}_i \mathbf{e}_i^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})] / \sum_{i=1}^{n} n_i \\
&\approx \sum_{i=1}^{n} \sum_{k=1}^{k_t} \mathrm{tr}[E(\mathbf{e}_i\, \mathbf{e}_i^T | (\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})] / \sum_{i=1}^{n} k_t n_i, \\
\hat{R} &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} E(\boldsymbol{\epsilon}_{ij}\, \boldsymbol{\epsilon}_{ij}^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / \sum_{i=1}^{n} m_i \\
&\approx \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{k_t} E(\boldsymbol{\epsilon}_{ij}\, \boldsymbol{\epsilon}_{ij}^T | (\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / \sum_{i=1}^{n} k_t m_i, \\
\hat{A} &= \sum_{i=1}^{n} E(\mathbf{a}_i\, \mathbf{a}_i^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / n \\
&\approx \sum_{i=1}^{n} \sum_{k=1}^{k_t} E(\mathbf{a}_i\, \mathbf{a}_i^T | (\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / k_t n, \\
\hat{B} &= \sum_{i=1}^{n} E(\mathbf{b}_i\, \mathbf{b}_i^T | \mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / n \\
&\approx \sum_{i=1}^{n} \sum_{k=1}^{k_t} E(\mathbf{b}_i\, \mathbf{b}_i^T | (\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) / k_t n.
\end{aligned}
$$

Plugging $\hat{\delta}^2$ and $\hat{R}$ into $\sum_{i=1}^{n} [\tilde{I}_1^{(i)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \tilde{I}_2^{(i)}(\boldsymbol{\alpha},\, R)]$ and setting its first derivative with respect to $(\boldsymbol{\alpha},\, \boldsymbol{\beta})$ equal to $\mathbf{0}$, we can get a set of equations for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$.

$$
\begin{bmatrix}
\sum_{i=1}^{n} \left(W_i^T\, W_i + \hat{\delta}^2 \sum_{j=1}^{m_i} U_{ij}^T\, \hat{R}^{-1}\, U_{ij} \right) & \sum_{i=1}^{n} W_i^T\, X_i \\
\sum_{i=1}^{n} X_i^T\, W_i & \sum_{i=1}^{n} X_i^T\, X_i
\end{bmatrix}
\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}
=
\begin{pmatrix}
\sum_{i=1}^{n} \left\{ W_i^T\, \nu_{\mathbf{y}_i} + \hat{\delta}^2 \sum_{j=1}^{m_i} U_{ij}^T\, \hat{R}^{-1}\, \nu_{\mathbf{z}_{ij}} \right\} \\
\sum_{i=1}^{n} X_i^T\, \nu_{\mathbf{y}_i},
\end{pmatrix}
$$

where

$$\begin{aligned}
\boldsymbol{\nu}_{\mathbf{y}_i} &= E(\mathbf{y}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) - \tilde{\mathbf{g}}_i - H_i\, E(\mathbf{a}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) - T_i\, E(\mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}), \\
\boldsymbol{\nu}_{\mathbf{z}_{ij}} &= \mathbf{z}_{ij} - V_{ij}\, E(\mathbf{a}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}).
\end{aligned}$$

The foregoing conditional expectations can be approximated as follows.

$$\begin{aligned}
E(\mathbf{y}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &\approx \sum_{k=1}^{k_t} (\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i})/k_t, \\
E(\mathbf{a}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &\approx \sum_{k=1}^{k_t} E(\mathbf{a}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})/k_t, \\
E(\mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}}) &\approx \sum_{k=1}^{k_t} E(\mathbf{b}_i|(\tilde{\mathbf{y}}_{mis,i}^{(k)},\, \mathbf{y}_{obs,i}),\, \mathbf{z}_i,\, \mathbf{r}_i;\, \tilde{\boldsymbol{\theta}})/k_t.
\end{aligned}$$

Using the inverse operation of partition matrices, we obtain closed-form estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the M-step:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \Gamma_{11} \sum_{i=1}^{n} \Big[W_i^T\, \boldsymbol{\nu}_{\mathbf{y}_i} + \hat{\delta}^2 \sum_{j=1}^{m_i} U_{ij}^T\, \hat{R}^{-1}\, \boldsymbol{\nu}_{\mathbf{z}_{ij}} \Big] + \Gamma_{12} \sum_{i=1}^{n} X_i^T\, \boldsymbol{\nu}_{\mathbf{y}_i}, \\
\hat{\boldsymbol{\beta}} &= \Gamma_{21} \sum_{i=1}^{n} \Big[W_i^T\, \boldsymbol{\nu}_{\mathbf{y}_i} + \hat{\delta}^2 \sum_{j=1}^{m_i} U_{ij}^T\, \hat{R}^{-1}\, \boldsymbol{\nu}_{\mathbf{z}_{ij}} \Big] + \Gamma_{22} \sum_{i=1}^{n} X_i^T\, \boldsymbol{\nu}_{\mathbf{y}_i},
\end{aligned}$$

where $\Gamma_{11} = M + Q\, L^{-1}\, Q^T$, $\Gamma_{12} = \Gamma_{21}^T = -Q\, L^{-1}$, and $\Gamma_{22} = L^{-1}$, with

$$\begin{aligned}
M &= \Big(\sum_{i=1}^{n} W_i^T\, W_i + \hat{\delta}^2 \sum_{i=1}^{n} \sum_{j=1}^{m_i} U_{ij}^T\, \hat{R}^{-1}\, U_{ij} \Big)^{-1}, \\
L &= \sum_{i=1}^{n} X_i^T\, X_i - \Big(\sum_{i=1}^{n} X_i^T\, W_i \Big)\, M\, \Big(\sum_{i=1}^{n} W_i^T\, X_i \Big), \\
Q &= M \Big(\sum_{i=1}^{n} W_i^T\, X_i \Big).
\end{aligned}$$

We finally maximize $\sum_{i=1}^{n} \tilde{l}_5^{(i)}(\boldsymbol{\eta})$ to obtain an updated estimate of $\boldsymbol{\eta}$, which can be done by standard optimization procedures such as the Newton-Raphson method.

Iterating between the above E-step and M-step until convergence, we obtain approximate MLEs of $\boldsymbol{\theta}$. The asymptotic variance-covariance matrix of the approximate MLE $\hat{\boldsymbol{\theta}}$ of

$\boldsymbol{\theta}$ can again be obtained using the approximate formula given in (4.5). The only difference is that the density function $f_Y(\mathbf{y}_i|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$ in $l_c^{(i)}$ in (4.1) is based on the LME response model (4.6). We see that, for this approximate method, both the E-step and the M-step are computationally much less intensive than those in Section 4.3. The performance of this approximate method will be evaluated in Section 4.6.

4.5 Example

We illustrate our proposed methods in this chapter using a HIV dataset. We also analyze this dataset using the commonly-used *naive method* which ignores measurement errors and missing data for comparison.

4.5.1 Data Description

The dataset includes 53 HIV infected patients who were treated with a potent antiretroviral regimen. Viral loads (Plasma HIV-1 RNA copies) were measured on days 0, 2, 7, 10, 14, 21, 28 and weeks 8, 12, 24, and 48 after initiation of treatments. After the antiretroviral treatment, the patients' viral loads will decay, and the decay rates may reflect the efficacy of the treatment. Throughout the time course, the viral load may continue to decay, fluctuate, or even start to rise (rebound). The data at the late stage of study are likely to be contaminated by long-term clinical factors. Various covariates such as CD4 count were also recorded throughout the study on similar schedules. It is well known that CD4 counts are usually measured with substantial errors. The number of response measurements for each individual varies from 6 to 10. Five patients dropped out of the study due to drug intolerance and other problems and sixteen patients have missing viral loads at scheduled time points. There were 104 out of 403 CD4 measurements missing at viral load measurement times, due mainly to

a somewhat different CD4 measurement schedule. A detailed data description can be found in Wu and Ding (1999).

Six patients are randomly selected and their viral loads are plotted in Figure 1.1. Due to long-term clinical factors, drug resistance, and other complications, the viral load trajectories can be very complex after the initial phase viral decay. Visual inspection of the raw data seems to indicate that dropout patients appear to have slower viral decay, compared with the remaining patients. Thus, the dropouts are likely to be informative or nonignorable. The CD4 count trajectories for six randomly selected patients are plotted in Figure 1.2. There exists large variability in CD4 count between patients. The population CD4 count trajectory appears to have a quadratic polynomial shape.

4.5.2 The Response and the Covariate Models

Based on Wu (2002) and Wu and Zhang (2002), we consider the following HIV viral dynamic model (see Section 3.5 for the details)

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{4.8}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z^*_{ij} + b_{2i}, \tag{4.9}$$

$$\log(P_{2i}) = \beta_4 + b_{3i}, \qquad \lambda_{2ij} = w(t_{ij}) + h_i(t_{ij}), \tag{4.10}$$

where y_{ij} is the $\log_{10}$-transform of the viral load measurement for patient i at time t_{ij}, P_{1i} and P_{2i} are baseline values, λ_{1ij} and λ_{2ij} are viral decay rates, z^*_{ij} is the true (but unobservable) CD4 count, and $w(t_{ij})$ and $h_i(t_{ij})$ are nonparametric fixed- and random-effects functions (see Section 2.1). To avoid very small (large) estimates, which may be unstable, we standardize the CD4 counts and rescale the original time t (in days) so that the new time scale is between 0 and 1.

As discussed in Section 2.1, we employ the linear combinations of natural cubic splines

Table 4.1: AIC and BIC values for the model (4.8) – (4.10), with $q \leq p = 1,\ 2,\ 3$.

Model	p=1,q=1	p=2,q=2	p=2,q=1	p=3,q=3	p=3,q=2	p=3,q=1
AIC	622.92	590.82	591.37	584.51	593.32	583.19
BIC	685.37	676.68	662.95	677.16	671.76	657.72

with percentile-based knots to approximate $w(t)$ and $h_i(t)$. Following Wu and Zhang (2002), we take the same natural cubic splines with $q \leq p$ in order to decrease the dimension of random effects. AIC and BIC criteria are used to determine the values of p and q. Table 4.1 displays AIC and BIC values for various plausible models. Based on these AIC and BIC values, the model with $p = 3$ and $q = 1$, i.e.,

$$\lambda_{2ij} \approx \beta_5 + \beta_6\,\psi_1(t_{ij}) + \beta_7\,\psi_2(t_{ij}) + b_{4i}, \tag{4.11}$$

seems to be the best, and thus it is selected for our analysis.

For the CD4 process, in the absence of a theoretical rationale, we consider empirical polynomial LME models, and choose the best fitted model based again on AIC/BIC values for each possible model. This is done based on the observed CD4 values, and is done separately from the response model for simplicity. Specifically, since the inter-patient variation is large, we consider model (2.8) with $U_{il} = V_{il} = (1, u_{il}, \ldots, u_{il}^{a-1})$ and linear ($a = 2$), quadratic ($a = 3$), and cubic ($a = 4$) polynomials. Table 4.2 presents AIC and BIC values for these three models. The following quadratic polynomial LME model best fits the observed CD4 process:

$$\text{CD4}_{il} = (\alpha_1 + a_1) + (\alpha_2 + a_2)\,u_{il} + (\alpha_3 + a_3)\,u_{il}^2 + \epsilon_{il}, \tag{4.12}$$

where u_{il} is the time and $\boldsymbol{\alpha} = (\alpha_1,\ \alpha_2,\ \alpha_3)^T$ are the population parameters and $\mathbf{a}_i = (a_{i1},\ a_{i2},\ a_{i3})^T$ are the random effects.

Table 4.2: AIC and BIC values for the linear, quadratic, and cubic CD4 models

Model	a=2	a=3	a=4
AIC	806.58	715.80	752.15
BIC	830.00	774.34	791.18

4.5.3 Dropout Models and Sensitivity Analysis

In this study, dropout patients appear to have slower viral decay, compared with the remaining patients. Thus, dropouts are likely to be informative or nonignorable. So we need to assume a model for the dropout process in order to make valid likelihood inference. Although dropout models are not verifiable based on observed data, subject-area knowledge and sensitivity analysis based on plausible models may still lead to reasonable models. Note that we should avoid building a too complicated dropout model since a complicated model may become non-identifiable (Fitzmaurice et al. 1996). Subject-area knowledge suggests that dropout may be related to current or previous viral load and CD4 measurements. Thus, we consider the following five plausible dropout models for sensitivity analysis

$$
\begin{aligned}
&\textit{Model } I: && \text{logit}[P(r_{ij}=1|\mathbf{y}_i,\mathbf{z}_i;\boldsymbol{\eta})] = \eta_1+\eta_2 CD4_{ij}+\eta_3 y_{ij},\\
&\textit{Model } II: && \text{logit}[P(r_{ij}=1|\mathbf{y}_i,\mathbf{z}_i;\boldsymbol{\eta})] = \eta_1+\eta_2 y_{i,j-1}+\eta_3 y_{ij},\\
&\textit{Model } III: && \text{logit}[P(r_{ij}=1|\mathbf{y}_i,\mathbf{z}_i;\boldsymbol{\eta})] = \eta_1+\eta_2 CD4_{ij}+\eta_3 y_{i,j-1}+\eta_4 y_{ij},\\
&\textit{Model } IV: && \text{logit}[P(r_{ij}=1|\mathbf{y}_i,\mathbf{z}_i;\boldsymbol{\eta})] = \eta_1+\eta_2 y_{ik}, \qquad k\le j,\\
&\textit{Model } V: && \text{logit}[P(r_{ij}=1|\mathbf{y}_i,\mathbf{z}_i;\boldsymbol{\eta})] = \eta_1+\eta_2 CD4^*_{ij},
\end{aligned}
\tag{4.13}
$$

where y_{ik} $(k \le j)$ in Model IV is the last observed response and $CD4^*_{ij}$ in Model V is the estimated true CD4 value for individual i. Thus Models I – III represent possible nonignorable missing response models, Model IV represents a possible ignorable missing response model, and Model V relates dropouts to (estimated) true CD4 values. We also

considered the following ignorable missing response models:

$$\text{logit}[P(r_{ij} = 1|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\eta})] = \eta_1 + \eta_2 y_{i1}$$

$$\text{logit}[P(r_{ij} = 1|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\eta})] = \eta_1 + \eta_2 t_{ij},$$

but the resulting estimates are similar to those for Model IV, so we only present results for Model IV in Table 4.3. We assume independence of the r_{ij}'s to simplify the model.

4.5.4 Estimation Methods and Computation Issues

We estimate the model parameters using the *naive method* which ignores measurement errors and missing data and the two proposed "joint" model methods discussed in Sections 4.3 and 4.4. We denote the method in Section 4.3 by APPR1 and the method in Section 4.4 by APPR2. The two proposed joint model methods need starting values for model parameters since they are implemented by MCEM algorithms. We use the parameter estimates obtained by the naive method as parameter starting values for the two joint model methods.

For the naive method, we use the SPLUS function *nlme*() and *lme*() to obtain parameter estimates and their default standard errors. For the two proposed joint model methods, we assess the convergence of the Gibbs sampler by examining time series plots and sample autocorrelation function plots. For example, Figures 4.1 and 4.2 show the time series and the autocorrelation function plots for b_2 associated with patient 14. From these figures, we notice that the Gibbs sampler converges quickly and the autocorrelations between successive generated samples are negligible after lag 17. Time series and autocorrelation function plots for other random effects and missing responses show similar behaviors. Therefore, we discard the first 500 samples as the burn-in, and then we take one sample from every 20 simulated samples to obtain independent samples (see sampling methods in Section 4.3.3). We start with $k_0 = 500$ Monte Carlo samples, and increase the Monte Carlo sample size as the number

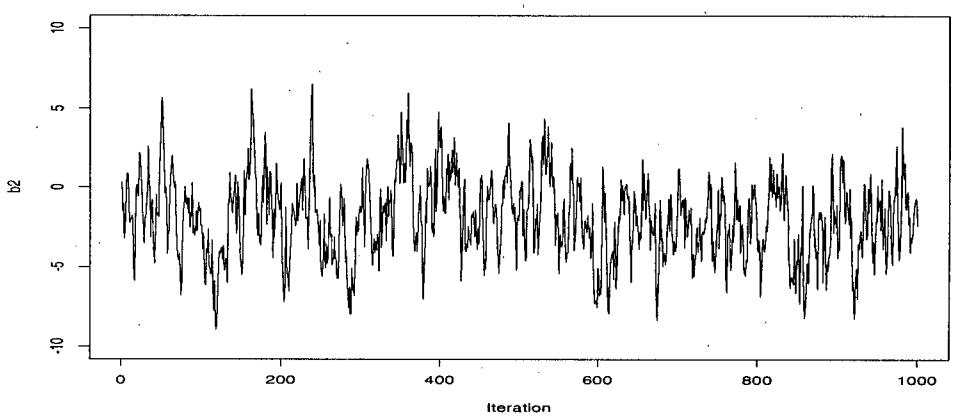

Figure 4.1: The time series plot for b_2 associated with patient 14.

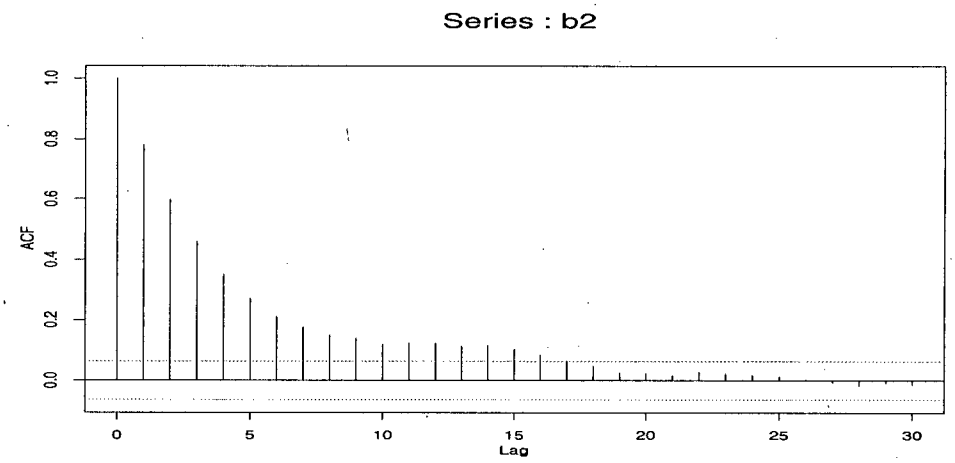

Figure 4.2: The autocorrelation function plot for b_2 associated with patient 14.

t of iteration increases: $k_{t+1} = k_t + k_t/c$ with $c = 4$ (Booth and Hobert, 1999). Convergence criterion for these two joint model methods is that the maximum relative change in the parameter estimates from successively iterations is smaller than 0.05. Convergence of the algorithms are considered to be achieved when the maximum percentage change of all estimates is less than 5% in two consecutive iterations.

We use the multivariate rejection sampling method for the two proposed joint model method. On a SUN Sparc work-station, the APPR1 method took about 140 minutes to converge while the APPR2 method took only 12 minutes to converge. This shows that APPR2 offers a big reduction in computing time, and thus is computationally much more efficient than APPR1.

4.5.5 Analysis Results

We estimate the model parameters using the naive method and the two proposed joint model methods APPR1 and APPR2. We use the parameter estimates obtained by the naive method as the parameter starting values for the APPR1 and the APPR2 methods. We also tried several other parameter starting values for the proposed methods. Different parameter starting values appear to lead to roughly the same parameter estimates for both the APPR1 and the APPR2 methods.

Table 4.3 presents the resulting parameter estimates and standard errors based on models I, IV, and V in (4.13). We find that the two joint model methods provide similar parameter estimates. We also find that the naive method may severely under-estimate the covariate CD4 effect (i.e., β_3) and may poorly estimate some other parameters as well (this will be confirmed by simulation). For the different dropout models in (4.13), we find that the resulting estimates based on the three nonignorable models (Models I, II, III) are all similar, which indicates that the estimation may be robust against the nonignorable dropout

Table 4.3: Parameter estimates (standard errors) for the models in the example.

Model	method	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7	δ	R
	NAIVE	–	–	–	11.72	65.71	0.84	6.87	–2.58	8.66	–1.90	.35	
		–	–	–	(.2)	(3.8)	(3.2)	(.6)	(5.5)	(8.9)	(3.1)		
Model I	APPR1	–.42	4.15	–3.75	11.72	67.08	1.52	6.97	–1.83	7.75	–2.54	.35	.51
		(.1)	(.5)	(.6)	(.2)	(5.2)	(6.2)	(.7)	(5.8)	(8.8)	(3.5)		
	APPR2	–.43	4.21	–3.78	11.70	66.97	1.50	6.96	–1.90	7.86	–2.63	.33	.50
		(.1)	(.6)	(.6)	(.2)	(4.4)	(5.8)	(.6)	(5.5)	(7.9)	(3.0)		
Model IV	APPR1	–.43	4.18	–3.75	11.73	66.52	1.37	6.89	–2.62	8.83	–1.92	.35	.51
		(.1)	(.5)	(.6)	(.2)	(5.0)	(6.0)	(.7)	(5.9)	(8.9)	(3.1)		
Model V	APPR1	–.43	4.21	–3.80	11.74	66.79	1.44	6.89	–2.50	8.60	–1.98	.35	.50
		(.1)	(.6)	(.6)	(.2)	(4.9)	(6.1)	(.7)	(5.9)	(8.9)	(3.1)		

Note: A and B are unstructured covariance matrices, but we only report the estimates of their diagonal elements here. $Diag(\hat{A}) = (.50, 3.65, 1.61)$ *for APPR1,* $Diag(\hat{A}) = (.52, 3.80, 1.66)$ *for APPR2.* $Diag(\hat{B}) = (1.08, 77.12, 2.03, 24.98)$ *for Naive,* $Diag(\hat{B}) = (1.10, 75.50, 2.01, 26.51)$ *for APPR1, and* $Diag(\hat{B}) = (1.09, 75.24, 1.83, 22.37)$ *for APPR2.*

models. The estimates based on the ignorable models (Models IV and V), however, appear to be somewhat different, especially for the parameters associated with the decay rates λ_{1ij} and λ_{2ij}. This suggests that the missing responses (dropouts) may be nonignorable, and reliable likelihood estimation must incorporate a reasonable nonignorable missing response model. Although some estimates in Table 4.3 are not statistically significant, the values of the estimates may still provide useful information about viral load and CD4 trajectories. The estimates of parameters η_2 and η_3 in dropout model I based on APPR1 method (or APPR2 method) are –.05 and .97 (–.04 and 1.06) respectively, with both p-values less than 0.001, which also indicates that the dropouts may be nonignorable (or informative) since the missingness may depend on the missing values. The estimates of η_2 and η_3 indicate that dropout patients seem to have lower CD4 counts and higher viral loads than the remaining patients.

4.6 A Simulation Study

In this section, we conduct a simulation study to i) evaluate the performances of the semi-parametric modelling and of the AIC/BIC knots selection method, ii) assess the two proposed methods (APPR1 and APPR2) and compare them to the naive method (NAIVE), and iii) evaluate the impact of specification of the missing response mechanisms on parameter estimation.

We generate 100 datasets from the following model, which corresponds to the model (4.8) – (4.10),

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{4.14}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z^*_{ij} + b_{2i}, \qquad \log(P_{2i}) = \beta_4 + b_{3i}, \tag{4.15}$$

$$\lambda_{2ij} = -2.2 + (5.3 + 0.1\, b_{4i}) \sin(0.04 + 3t_{ij}), \tag{4.16}$$

where the nonparametric model (4.16) is carefully chosen to closely mimic the viral load trajectory at later stages in the example of the previous section. The covariate model and the measurement time points used in the simulation are the same as those in the example of the previous section. The true values of model parameters are similar to those in the example. The true values of (β_1, β_2, β_3, β_4) and $\boldsymbol{\alpha}$ are presented in Table 4.4, and the other true parameter values are $\delta = .2$, $R = .4$, $A = \text{diag}(.5, 3, 2)$, and $B = \text{diag}(1, 9, 2, 4)$. Note that $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4}) \sim N(0, B)$, where b_{4i} is incorporated in the nonparametric model (4.16). There are 147 viral loads after the rescaled time 0.25. We regard the viral loads out of 147 greater than the 45th percentile as missing responses. Thus, we delete 20% largest response values at the last few time points to mimic an informative missing response mechanism (i.e., missingness depends on the values being missing). The structure of the data generated by the simulation study is similar to that in the example in Section 4.5.

We calculate averages of the resulting estimates and their standard errors based on

each method, and compare the methods by comparing their biases and mean-square-errors (MSEs). Here, bias and MSE are assessed in terms of percent relative bias and percent relative root mean-squared error, as defined next. For instance, the bias for β_j, the jth component of $\boldsymbol{\beta}$, is defined as

$$\text{bias}_j = \hat{\beta}_j - \beta_j,$$

where $\hat{\beta}_j$ is the estimate of β_j. The mean-squared error for β_j is defined as

$$\text{MSE}_j = \text{bias}_j^2 + s_j^2,$$

where s_j is the simulated standard error of $\hat{\beta}_j$. Then, the percent relative bias of $\hat{\beta}_j$ is defined as

$$\text{bias}_j/|\beta_j| \times 100\%,$$

and the percent relative root MSE is

$$\sqrt{\text{MSE}_j}/|\beta_j| \times 100\%.$$

First, to evaluate the nonparametric modelling, we study the performance of the AIC and BIC criteria in selecting the numbers of knots (p and q), since these numbers represent the degrees of smoothness of nonparametric functions (too large/small values may result in overfit/underfit). For the 100 datasets simulated from the semiparametric NLME model (4.14) – (4.16), we find that all BIC values and 97% of AIC values lead to the model (4.11) (i.e., $p = 3$, $q = 1$). To further evaluate the AIC and BIC methods, we also generate data from models (4.8) – (4.12) with $(\beta_5, \beta_6, \beta_7) = (-2.0, 8.0, -3.0)$ (so the true number of knots are known), and use the AIC and BIC methods to select the best model. The performance of the AIC and BIC methods is similar. These results show that the AIC and BIC criteria perform well in the current setting.

Table 4.4: Simulation results for the parameter estimates (standard errors) as well as their biases and MSEs for the estimation methods PARA, APPR1, and APPR2.

Parameter	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7
True Value	–0.4	4.0	–4.0	12.0	67.0	1.5	7.0			
PARA	–	–	–	11.94	60.91	–0.62	6.48	0.20		
	–	–	–	(.1)	(1.5)	(1.8)	(.2)	(.1)		
APPR1	–0.39	4.05	–3.99	12.00	66.87	1.49	7.11	–1.89	9.65	–1.66
	(.1)	(.3)	(.3)	(.1)	(1.3)	(1.6)	(.3)	(2.0)	(3.1)	(1.1)
APPR2	–0.39	4.06	–4.01	12.00	66.25	1.59	6.90	–3.11	10.18	–1.47
	(.1)	(.3)	(.3)	(.2)	(1.1)	(1.4)	(.3)	(1.7)	(2.6)	(1.0)
				Bias						
PARA	–	–	–	–.48	–9.09	–141.03	–7.43			
APPR1	1.28	1.27	.17	.01	–.19	–.60	1.29			
APPR2	1.73	1.55	–.18	.03	–1.07	6.99	–1.36			
				MSE						
PARA	–	–	–	1.71	9.38	230.93	8.18			
APPR1	21.33	10.02	9.87	1.12	2.48	98.67	4.12			
APPR2	25.40	10.49	10.56	1.34	2.73	100.94	5.06			

Note: Bias = Percent bias $= 100 \times \text{bias}_j/|\beta_j|$; *MSE = Percent* $\sqrt{\text{MSE}} = 100 \times \sqrt{\text{MSE}_j}/|\beta_j|$.

To investigate the effect of semiparametric modelling on the estimation of the fixed-effects parameters $\beta_1 - \beta_4$, we consider the two proposed methods (APPR1 and APPR2) for the semiparametric NLME model (4.14) – (4.16), along with the covariate model (4.12) and a nonignorable dropout model (Model I in (4.13)). We also use a parametric NLME model, where $\lambda_{2ij} = \beta_5 + b_{4i}$ and the other parts are the same as in the semiparametric NLME model (4.14) – (4.16), to fit the simulated datasets. To emphasize the difference between the nonparametric and the parametric modelling for λ_{2ij}, we consider an ideal case for the parametric NLME model fitting, in which there is no covariate measurement errors and dropouts. Thus, we do not need the covariate measurement error model and the dropout model in this ideal case. We use SPLUS function *nlme*() to obtain parameter estimates and their default standard errors, denoted by the PARA method. We calculate averages of the resulting estimates and their standard errors based on each method. Since APPR1 method sometimes offers computational problems, such as slow or non-convergence, the 100 sets of parameter estimates are obtained from 137 data sets. The simulation results are shown in Table 4.4. We find that estimates for the fixed-effects parameters $\beta_1 - \beta_4$ obtained by APPR1 and APPR2 are very close to their true values, and both methods perform better than the PARA method in terms of bias and MSE criteria. These results show that the semiparametric modelling based on AIC/BIC for knots selection performs well and better than the parametric modelling in the current setting.

To study the effect of missing data mechanisms, we assess the proposed methods (APPR1 and APPR2) based on a nonignorable model (Model I) and an ignorable model (Model IV), and compare them with the naive method (which ignores measurement errors and missing data). To investigate the performance of the estimate of λ_{2ij}, we generate 100 datasets from the true models (4.8) – (4.12). In the simulations, the true values of model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are shown in Table 4.5, and the other true parameter values, the missing

Table 4.5: Simulation results for the parameter estimates (standard errors) for the three estimation methods NAIVE, APPR1, and APPR2 with dropout models I and IV in (4.13).

Dropout Model	Parameter True Value	α_1 −.4	α_2 4.0	α_3 −4.0	β_1 12.0	β_2 67.0	β_3 1.5	β_4 7.0	β_5 −2.0	β_6 8.0	β_7 −3.0
	NAIVE	–	–	–	11.98	66.87	0.92	6.98	−2.41	9.64	−2.03
		–	–	–	(.1)	(1.2)	(1.1)	(0.3)	(1.9)	(2.9)	(1.0)
Model I	APPR1	−.40	4.08	−4.00	11.99	66.93	1.48	7.01	−2.04	8.11	−2.92
		(.1)	(.3)	(.3)	(.1)	(1.4)	(1.6)	(.3)	(1.9)	(3.1)	(1.2)
	APPR2	−.40	4.06	−4.01	11.99	66.86	1.53	7.03	−2.11	8.17	−2.87
		(.1)	(.3)	(.3)	(.2)	(1.3)	(1.5)	(.3)	(1.8)	(2.8)	(1.0)
Model IV	APPR1	−.40	4.06	−3.99	12.00	67.15	1.45	7.01	−1.81	8.97	−2.22
		(.1)	(.3)	(.3)	(.1)	(1.3)	(1.6)	(.3)	(2.1)	(3.2)	(1.2)
	APPR2	−.39	4.06	−4.00	11.99	66.80	1.42	6.96	−1.78	9.08	−2.28
		(.1)	(.3)	(.3)	(.2)	(1.2)	(1.5)	(.3)	(1.8)	(2.8)	(1.0)

mechanism and the missing rate are the same as above. We calculate averages of the resulting estimates and their standard errors based on each of the three methods and each of the two dropout models. We compare the methods by comparing their biases and mean-square-errors (MSEs). Since APPR1 method sometimes offers computational problems, such as slow or non-convergence, the 100 sets of parameter estimates are obtained from 130 data sets.

From the simulation results in Tables 4.5 and 4.6, we see that, when measurement errors and reasonable missing data mechanisms (Model I) are taken into account, the two proposed joint model methods (APPR1 and APPR2) perform well in terms of both bias and MSE criteria. APPR1 performs better than APPR2 as expected, but APPR2 also performs reasonably well and is computationally much more efficient. When the missing data mechanism is ignored (Model IV), however, the two methods may not perform well. The naive method, which ignores measurement errors and missing data, may lead to severely biased estimates and large MSEs (e.g., the covariate effect β_3 can be severely under-estimated).

Table 4.6: Simulation results for biases and MSEs of the parameter estimates for the three estimation methods NAIVE, APPR1, and APPR2 with dropout models I and IV in (4.13).

Dropout Model	Parameter True Value	α_1 −.4	α_2 4.0	α_3 −4.0	β_1 12.0	β_2 67.0	β_3 1.5	β_4 7.0	β_5 −2.0	β_6 8.0	β_7 −3.0
					Bias						
	NAIVE	–	–	–	−.10	−.25	−38.42	−.61	−20.53	20.45	32.33
Model I	APPR1	−.78	1.95	.01	−.08	−.11	−1.59	.18	−2.21	1.40	2.62
	APPR2	−.78	1.95	−.01	−.09	−.21	−1.78	.50	−5.70	2.08	4.30
Model IV	APPR1	−.97	2.02	.16	−.08	.22	−3.63	.19	9.26	12.10	23.10
	APPR2	1.40	2.07	−.09	−.09	−.29	−6.91	−.55	10.66	13.57	26.02
					MSE						
	NAIVE	–	–	–	1.89	2.97	148.06	6.96	168.66	68.32	70.14
Model I	APPR1	20.00	7.28	7.00	1.17	1.84	96.01	3.43	80.02	35.15	40.42
	APPR2	25.00	7.65	8.00	1.25	2.23	98.69	3.74	112.64	39.06	43.21
Model IV	APPR1	30.51	9.90	9.84	1.45	2.49	120.12	4.84	122.80	56.26	58.90
	APPR2	36.50	10.46	10.31	1.74	2.58	131.24	5.73	146.04	62.14	66.07

Note: Bias = Percent bias $= 100 \times \text{bias}_j/|\beta_j|$; *MSE = Percent* $\sqrt{\text{MSE}} = 100 \times \sqrt{\text{MSE}_j}/|\beta_j|$.

4.7 Conclusions and Discussion

We have proposed two approximate likelihood methods for semiparametric NLME models with *outcome-based informative* dropouts and covariate measurement errors and missing data, implemented by Monte Carlo EM algorithms combined with Gibbs sampler. The first method is more accurate than the second method but it may be computationally very intensive and sometimes may offer computational difficulties such as slow or non-convergence, especially when the dimensions of random effects are not small. The second method is computationally much more efficient, but it is less accurate than the first method. The second method may be used as a reasonable alternative when the first method has convergence problems or it may be used to provide excellent parameter starting values for the first method. Simulation studies indicate that the proposed methods, which incorporate measurement

errors and dropout mechanisms, produce satisfactory results, but methods ignoring measurement errors and/or ignoring dropout mechanisms may perform poorly. Moreover, the AIC and BIC criteria perform well in the current setting.

We have assumed that the dropout models depend on the observed or unobserved responses and covariates. Alternatively, we may consider dropout models which share the random-effects parameters in response and covariate processes. Such models may be appropriate if the dropout mechanism is related to the true but unobservable response/covariate values or summaries of response and covariate processes such as unobservable true viral decay rates. The methods in this chapter may be extended to such models. Finally, for Monte Carlo EM algorithms, Booth and Hobert (1999) proposed a nice automated method for choosing the number of Monte Carlo samples, which can be extended in our case as well.

Chapter 5

Semiparametric NLME Model with Random-effect-based Informative Dropouts and Covariate Measurement Errors

5.1 Introduction

In this chapter, we develop two likelihood methods to simultaneously address covariate measurement errors and *random-effect-based informative* dropouts in semiparametric NLME models. The major difference in the models in this chapter and the models in Chapter 4 is the difference in the assumed missing response (or dropout) models. The response and covariate models remain the same. In Section 5.2, we discuss the models for this problem. We obtain approximate MLEs of all model parameters, using a Monte Carlo EM (MCEM) algorithm along with Gibbs sampler methods, in Sections 5.3. To avoid potential compu-

tational problems in the method discussed in Section 5.3, we also propose an alternative approximate method by using a first-order Laplace approximation to the log-likelihood function in Section 5.4. Some asymptotic properties of the resulting estimates are also discussed. The two proposed methods are illustrated in a HIV dataset and are evaluated via simulation in Section 5.5. We conclude this chapter in Section 5.6 with some discussion. Asymptotic properties presented in Section 5.4 are proved in Section 5.7.

5.2 Missing Response Models

The dropout is *random-effect-based informative* if the missing probability of the current response depends only on the underlying unobservable random effects $\mathbf{a}_i$ and $\mathbf{b}_i$, i.e., $f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta}) = f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})$. In the presence of *random-effect-based informative* dropouts in the NLME model (2.6) and (2.7) with the covariate process (2.8), we can again write $\mathbf{y}_i = (\mathbf{y}_{mis,i}, \mathbf{y}_{obs,i})$ as before. Let $n_{obs,i}$ be the number of components in $\mathbf{y}_{obs,i}$. Here, the missing y_{ij}'s are again intermittent, i.e., we allow dropout individuals to possibly return to the study at a later time. For the vector $\mathbf{g}_i = (g_{i1}, \ldots, g_{in_i})$, where g_{ij} are defined in (3.6), we write $\mathbf{g}_i = (\mathbf{g}_{mis,i}, \mathbf{g}_{obs,i})$ with $\mathbf{g}_{obs,i}$ and $\mathbf{g}_{mis,i}$ being the conditional expectation of $\mathbf{y}_{obs,i}$ and $\mathbf{y}_{mis,i}$, respectively. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{in_i})^T$ be a vector of missing response indicators for individual i such that $r_{ij} = 1$ if y_{ij} is missing and 0 otherwise. Note that $r_{ij} = 1$ does not necessarily imply that $r_{i,j+1} = 1$. We have the observed data $\{(\mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i), i = 1, \ldots, n\}$.

To allow for an informative missing mechanism in the response, we need to assume a distribution for the missing response indicator $\mathbf{r}_i$. For reasons discussed in Sections 4.2 and 4.7, in this chapter we consider the *random-effect-based informative* dropout mechanism $f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})$ for $\mathbf{r}_i$, where $\boldsymbol{\eta}$ are the unknown nuisance parameters. For such a missing data mechanism, the missingness of y_{ij} share the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ in the response

and covariate models, suggesting that the dropout may be related to the true but unobservable response/covariate values or summaries of individual-specific response and covariate trajectories such as unobservable true viral decay rates. For example, such missing data models may be appropriate if a patient is more likely to dropout early because his true (but unobservable) viral decay rate is slower than other patients.

In this chapter, we focus on the *random-effect-based informative* missing mechanism $f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})$. Such models are related to the shared-parameter models in the literature (e.g., Wu and Carroll, 1988; Little, 1995; Ten Have et al., 1998). Although the relationship of the missingness with the random effects may be complex, a simple logistic regression model may provide a reasonable approximation. We will assume that r_{ij}'s are independent, i.e.,

$$f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta}) = \prod_{i=1}^{n_i}[P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})]^{r_{ij}}[1 - P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})]^{1-r_{ij}}, \tag{5.1}$$

with

$$\text{logit}[P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})] = \log \frac{P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})}{1 - P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})} = \eta_0 + \boldsymbol{\eta}_1^T \mathbf{a}_i + \boldsymbol{\eta}_2^T \mathbf{b}_i,$$

where $\boldsymbol{\eta} = (\eta_0, \boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T)^T$ are the unknown nuisance parameters. For example, we may assume that the missingness of response is related to the first decay rate, say $\lambda_{1ij} = \beta_2 + b_{i2}$ in Section 4.5., i.e.,

$$\begin{aligned}
\text{logit}[P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})] &= \eta_0^* + \eta_1^* \lambda_{1ij} \\
&= \eta_0^* + \eta_1^* (\beta_2 + b_{i2}) \\
&= (\eta_0^* + \eta_1^* \beta_2) + \eta_1^* b_{i2} \\
&= \eta_0 + \eta_1 b_{i2}.
\end{aligned}$$

Note that we should avoid building too complicated a dropout model since the model parameters may become non-identifiable. As the assumed dropout models are not testable

based on the observed data, it is important to carry out sensitivity analysis based on various dropout models. The *random-effect-based informative* dropout models in this chapter and the *outcome-based informative* dropout models in Chapter 4 can all be used for sensitivity analysis.

5.3 A Monte Carlo EM Method

5.3.1 The Likelihood Function

We consider likelihood inference for semiparametric NLME models with *random-effect-based informative* dropouts and measurement errors and missing data in time-varying covariates, based on approximate models (2.6) – (2.8). Let $\boldsymbol{\theta} = (\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2,\ R,\ A, B,\ \boldsymbol{\eta})$ be the collection of all unknown model parameters. We assume that the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, δ^2, R, A, B, and $\boldsymbol{\eta}$ are distinct. The approximate log-likelihood for the observed data $\{(\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i),\ i = 1, \ldots, n\}$ can be written as

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \int\int\int \Big[f_Y(\mathbf{y}_i|\mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i; \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2)\, f_Z(\mathbf{z}_i|\mathbf{a}_i; \boldsymbol{\alpha}, R)\, f(\mathbf{a}_i; A) \\
&\qquad \times f(\mathbf{b}_i; B)\, f(\mathbf{r}_i|\mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\eta}) \Big]\, d\mathbf{y}_{mis,i}\, d\mathbf{a}_i\, d\mathbf{b}_i, \\
&= \sum_{i=1}^{n} \log \int\int \Big[f_Y(\mathbf{y}_{obs,i}|\mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i; \boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2)\, f_Z(\mathbf{z}_i|\mathbf{a}_i; \boldsymbol{\alpha}, R)\, f(\mathbf{a}_i; A) \\
&\qquad \times f(\mathbf{b}_i; B)\, f(\mathbf{r}_i|\mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\eta}) \Big]\, d\mathbf{a}_i\, d\mathbf{b}_i, \qquad (5.2)
\end{aligned}
$$

where

$$
\begin{aligned}
f_Y(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) &= \textstyle\prod_{j=1}^{n_{obs,i}} f_Y(y_{obs,ij}|\mathbf{z}_{ij},\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \\
&= \textstyle\prod_{j=1}^{n_{obs,i}} (2\pi\delta^2)^{-1/2} \exp\{-[y_{obs,ij} - g_{obs,ij}]^2/2\delta^2\}, \\
f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) &= \textstyle\prod_{k=1}^{m_i} f_Z(\mathbf{z}_{ik}|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
&= \textstyle\prod_{k=1}^{m_i} |2\pi R|^{-1/2} \exp\{-(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)^T R^{-1} \\
&\qquad \times(\mathbf{z}_{ik} - \mathbf{u}_{ik}\,\boldsymbol{\alpha} - \mathbf{v}_{ik}\,\mathbf{a}_i)/2\}, \\
f(\mathbf{a}_i;\, A) &= |2\pi A|^{-1/2} \exp\{-\mathbf{a}_i^T A^{-1} \mathbf{a}_i/2\}, \\
f(\mathbf{b}_i;\, B) &= |2\pi B|^{-1/2} \exp\{-\mathbf{b}_i^T B^{-1} \mathbf{b}_i/2\},
\end{aligned}
$$

and $y_{obs,ij}$ is the observed y_{ij}. Note that unlike $l(\boldsymbol{\theta})$ in Section 4.3, the missing responses $\mathbf{y}_{mis,i}$ are integrated out in (5.2).

The observed-data log-likelihood function $l(\boldsymbol{\theta})$ generally does not have a closed-form expression since the functions in the integral can be nonlinear in the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$. So we use a Monte Carlo EM (MCEM) algorithm to find the approximate MLEs of parameters $\boldsymbol{\theta}$. By treating the unobservable random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ as "missing" data, we have "complete data" $\{(\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \mathbf{a}_i,\ \mathbf{b}_i),\ i = 1, \ldots, n\}$. The complete-data log-likelihood function for all individuals can be expressed as

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_c^{(i)}(\boldsymbol{\theta}) \;\equiv\; & \sum_{i=1}^{n} \{\log f_Y(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i; \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) \\
& + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B) + \log f(\mathbf{r}_i|\mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\eta})\}
\end{aligned} \tag{5.3}
$$

where $l_c^{(i)}$ is the complete-data log-likelihood for individual i.

5.3.2 A MCEM Algorithm

Let $\boldsymbol{\theta}^{(t)}$ be the parameter estimates from the t-th EM iteration. The E-step for individual i at the $(t+1)$th EM iteration can be expressed as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E(l_c^{(i)}(\boldsymbol{\theta})|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)}) = \int\int\int \Big[\log f_Y(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) \\ &\quad + \log f_Z(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\alpha},\, R) + \log f(\mathbf{a}_i;\, A) + \log f(\mathbf{b}_i;\, B) \\ &\quad + \log f(\mathbf{r}_i|\mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\eta})\Big] \times f(\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)})\, d\mathbf{a}_i\, d\mathbf{b}_i \\ &\equiv I_1^{(i)}(\boldsymbol{\alpha},\, \boldsymbol{\beta},\, \delta^2) + I_2^{(i)}(\boldsymbol{\alpha},\, R) + I_3^{(i)}(A) + I_4^{(i)}(B) + I_5^{(i)}(\boldsymbol{\eta}). \end{aligned} \tag{5.4}$$

Since the expression (5.4) is an expectation with respect to $f(\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)})$, it may be evaluated using the MCEM algorithm. Specifically, we may use the Gibbs sampler to generate samples from $[\mathbf{a}_i,\, \mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i;\, \boldsymbol{\theta}^{(t)}]$ by iteratively sampling from the full conditionals $[\mathbf{a}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}]$ and $[\mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}]$ as follows.

$$\begin{aligned} f(\mathbf{a}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f(\mathbf{y}_{obs,i},\, \mathbf{a}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{a}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i,\, \mathbf{a}_i|\mathbf{z}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{a}_i|\mathbf{z}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{a}_i|\mathbf{z}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{z}_i|\mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \end{aligned} \tag{5.5}$$

$$\begin{aligned} f(\mathbf{b}_i|\mathbf{y}_{obs,i},\, \mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) &\propto f(\mathbf{y}_{obs,i},\, \mathbf{b}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{b}_i|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &\propto f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{r}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i,\, \mathbf{b}_i|\mathbf{z}_i,\, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{z}_i,\, \mathbf{a}_i, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{b}_i|\mathbf{z}_i, \mathbf{a}_i;\, \boldsymbol{\theta}^{(t)}) \\ &= f(\mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\, \mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}) \cdot f(\mathbf{r}_i|\mathbf{a}_i,\, \mathbf{b}_i;\, \boldsymbol{\theta}^{(t)}). \end{aligned}$$

Monte Carlo samples from each of the full conditionals can be generated using rejection sampling methods. Alternatively, integral (5.4) may be evaluated using the importance sampling method (see Section 3.3.3). We will briefly discuss the sampling methods in the next section.

Note that, unlike the MCEM method in Section 4.3.2, here we do not need to sample $\mathbf{y}_{mis,i}$, which reduces the computational burden.

For individual i, let $\{(\tilde{\mathbf{a}}_i^{(1)}, \tilde{\mathbf{b}}_i^{(1)}), \ldots, (\tilde{\mathbf{a}}_i^{(k_t)}, \tilde{\mathbf{b}}_i^{(k_t)})\}$ denote a random sample of size k_t generated from $[\mathbf{a}_i, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i; \boldsymbol{\theta}^{(t)}]$. Note that each $(\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)})$ depends on the EM iteration number t, which is suppressed throughout. The E-step at the $(t+1)$th EM iteration can then be expressed as

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{i=1}^{n} \left\{ \frac{1}{k_t} \sum_{k=1}^{k_t} l_c^{(i)}(\boldsymbol{\theta}; \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{r}_i, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}) \right\} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f_Y(\mathbf{y}_{obs,i}|\mathbf{z}_i, \tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f_Z(\mathbf{z}_i|\tilde{\mathbf{a}}_i^{(k)}; \boldsymbol{\alpha}, R) + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f(\tilde{\mathbf{a}}_i^{(k)}; A) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f(\tilde{\mathbf{b}}_i^{(k)}; B) + \sum_{i=1}^{n} \sum_{k=1}^{k_t} \frac{1}{k_t} \log f(\mathbf{r}_i|\tilde{\mathbf{a}}_i^{(k)}, \tilde{\mathbf{b}}_i^{(k)}; \boldsymbol{\eta}) \\ &\equiv Q^{(1)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2|\boldsymbol{\theta}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha}, R|\boldsymbol{\theta}^{(t)}) + Q^{(3)}(A|\boldsymbol{\theta}^{(t)}) + Q^{(4)}(B|\boldsymbol{\theta}^{(t)}) + Q^{(5)}(\boldsymbol{\eta}|\boldsymbol{\theta}^{(t)}). \end{aligned} \tag{5.6}$$

The M-step then maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to produce an updated estimate $\boldsymbol{\theta}^{(t+1)}$, so it is like a complete-data maximization. Since the parameters in $Q^{(1)} + Q^{(2)}$, $Q^{(3)}$, $Q^{(4)}$, and $Q^{(5)}$ are distinct, the M-step can be implemented by maximizing $Q^{(1)} + Q^{(2)}$, $Q^{(3)}$, $Q^{(4)}$, and $Q^{(5)}$ separately using standard optimization procedures for the corresponding complete-data models.

As in Section 3.3.2, we use the approximate formula suggested by McLachlan and Krishnan (1997) to obtain the variance-covariance matrix of the approximate MLE $\hat{\boldsymbol{\theta}}$. Let $\mathbf{s}_c^{(i)} = \partial l_c^{(i)} / \partial \boldsymbol{\theta}$, where $l_c^{(i)}$ is the complete-data log-likelihood for individual i. Then an

approximate formula for the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \Big[\sum_{i=1}^{n} E(\mathbf{s}_c^{(i)}|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \hat{\boldsymbol{\theta}})\ E(\mathbf{s}_c^{(i)}|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \hat{\boldsymbol{\theta}})^T\Big]^{-1},$$

where the expectations can be approximated by Monte Carlo empirical means, as in (5.6).

In summary, the foregoing MCEM algorithm can be implemented as follows.

Step 1. Obtain an initial estimate of $(\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \delta^2,\ R,\ A,\ B) = (\boldsymbol{\alpha}^{(0)},\ \boldsymbol{\beta}^{(0)},\ \delta^{2(0)},\ R^{(0)},\ A^{(0)},\ B^{(0)})$ and an initial value of $(\mathbf{a}_i,\ \mathbf{b}_i) = (\mathbf{a}_i^{(0)},\ \mathbf{b}_i^{(0)})$ based on a naive method; then we obtain an initial estimate of $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$ based on the dropout model with the random effects $(\mathbf{a}_i,\ \mathbf{b}_i) = (\mathbf{a}_i^{(0)},\ \mathbf{b}_i^{(0)})$.

Step 2. At the t-th iteration, obtain Monte Carlo samples of the random effects $(\mathbf{a}_i,\ \mathbf{b}_i)$ using the Gibbs sampler along with rejection sampling methods, or using importance sampling methods to approximate the conditional expectation in the E-step.

Step 3. Obtain updated estimates $\boldsymbol{\theta}^{(t+1)}$ using standard complete-data optimization procedures.

Step 4. Iterate between Step 2 and Step 3 until convergence.

5.3.3 Sampling Methods

Gibbs Sampler

As in Section 3.3.3, we can again use the Gibbs sampler to draw the desired samples as follows. Set initial values $(\tilde{\mathbf{a}}_i^{(0)},\ \tilde{\mathbf{b}}_i^{(0)})$. Suppose that the current generated values are $(\tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)})$, we can obtain $(\tilde{\mathbf{a}}_i^{(k+1)},\ \tilde{\mathbf{b}}_i^{(k+1)})$ as follows.

Step 1. Draw a sample for the "missing" random effects $\tilde{\mathbf{a}}_i^{(k+1)}$ from $f(\mathbf{a}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \tilde{\mathbf{b}}_i^{(k)};\ \boldsymbol{\theta}^{(t)})$.

Step 2. Draw a sample for the "missing" random effects $\tilde{\mathbf{b}}_i^{(k+1)}$ from $f(\mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \tilde{\mathbf{a}}_i^{(k+1)};\ \boldsymbol{\theta}^{(t)})$.

After a sufficiently large burn-in of r iterations, the sampled values will achieve a steady state. Then, $\{(\tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)}),\ k = r+1, \ldots, r+k_t\}$ can be treated as samples from the

multidimensional density function $f(\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \boldsymbol{\theta}^{(t)})$. And, if we choose a sufficiently large gap r' (say $r' = 10$), we can treat $\{(\tilde{\mathbf{a}}_i^{(k)},\ \tilde{\mathbf{b}}_i^{(k)}), k = r + r', r + 2r', \dots\}$ as independent samples from $f(\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i, \mathbf{r}_i;\ \boldsymbol{\theta}^{(t)})$. There are several ways to get the initial values $(\tilde{\mathbf{a}}_i^{(0)},\ \tilde{\mathbf{b}}_i^{(0)})$. A simple way is to obtain $(\tilde{\mathbf{a}}_i^{(0)},\ \tilde{\mathbf{b}}_i^{(0)})$ based on a naive method.

Multivariate Rejection Algorithm

Sampling from the two full conditionals can be accomplished by the multivariate rejection sampling method. For example, we consider sampling from $f(\mathbf{a}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \mathbf{b}_i;\ \boldsymbol{\theta}^{(t)})$ in (5.5). Let $f^*(\mathbf{a}_i) = f(\mathbf{y}_{obs,i}|\mathbf{z}_i,\ \mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\theta}^{(t)})\, f(\mathbf{z}_i|\mathbf{a}_i;\ \boldsymbol{\theta}^{(t)})\, f(\mathbf{r}_i|\mathbf{a}_i,\ \mathbf{b}_i;\ \boldsymbol{\theta}^{(t)})$ and $\varsigma = \sup_{\mathbf{u}}\{f^*(\mathbf{u})\}$. We assume $\varsigma < \infty$. A random sample from $f(\mathbf{a}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \mathbf{b}_i;\ \boldsymbol{\theta}^{(t)})$ can then be obtained as follows

Step 1. Sample $\mathbf{a}_i^*$ from $f(\mathbf{a}_i;\ \boldsymbol{\theta}^{(t)})$, and independently, sample w from the uniform (0, 1) distribution.

Step 2. If $w \le f^*(\mathbf{a}_i^*)/\varsigma$, then accept $\mathbf{a}_i^*$, otherwise, go back to step 1.

Samples from $f(\mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i,\ \mathbf{a}_i;\ \boldsymbol{\theta}^{(t)})$ can be obtained in a similar way. Therefore, the Gibbs sampler in conjunction with the multivariate rejection sampling can be used to obtain samples from $[\mathbf{a}_i,\ \mathbf{b}_i|\mathbf{y}_{obs,i},\ \mathbf{z}_i,\ \mathbf{r}_i;\ \boldsymbol{\theta}^{(t)}]$.

5.4 An Alternative Approximate Method

5.4.1 The Hierarchical Likelihood Method

The approximate maximum likelihood inference using a Monte Carlo EM method in the previous section may be computationally intensive and sometimes may offer potential computational problems such as slow or non-convergence, especially when the dimensions of the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ are not small (see the detailed discussion in Section 3.4.1). To

overcome these difficulties, in this section we consider an alternative method called the *hierarchical likelihood* method (Lee and Nelder, 1996, 2001) for approximate inference. The hierarchical likelihood method avoids Monte Carlo approximation and thus is always computationally feasible, and it can also used to obtain good parameter starting values for the MCEM method in the previous section.

Let $\boldsymbol{\xi}$ denote general "nuisance parameters". Lee and Nelder (1996) considered a function $p_{\hat{\boldsymbol{\xi}}}(l)$ defined by

$$p_{\hat{\boldsymbol{\xi}}}(l) = \left[l - \frac{1}{2} \log \left| \frac{1}{2\pi} D(l, \boldsymbol{\xi}) \right| \right] \Big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}},$$

where $D(l, \boldsymbol{\xi}) = -\partial^2 l / \partial \boldsymbol{\xi}^2$, and $\hat{\boldsymbol{\xi}}$ solves $\partial l / \partial \boldsymbol{\xi} = \mathbf{0}$. Following Lee and Lelder (1996), the complete-data log-likelihood function $l_c(\boldsymbol{\theta})$ in (5.3) may also be called the hierarchical log-likelihood function since it combines the two stages of mixed-effects models. Let $\boldsymbol{\omega} = \{\boldsymbol{\omega}_i = (\mathbf{a}_i, \mathbf{b}_i),\ i = 1, \ldots, n\}$ be the collection of random effects. The function $p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))$ can be written as

$$p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta})) = \sum_{i=1}^{n} p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) \equiv \sum_{i=1}^{n} \left[l_c^{(i)}(\boldsymbol{\theta}) - \frac{1}{2} \log \left| \frac{1}{2\pi} D(l_c^{(i)}(\boldsymbol{\theta}), \boldsymbol{\omega}_i) \right| \right]_{\boldsymbol{\omega}_i=\hat{\boldsymbol{\omega}}_i}. \tag{5.7}$$

We can show that, for unobservable $\boldsymbol{\omega}$, the use of the function $p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))$ is equivalent to integrating $\boldsymbol{\omega}$ out using the first-order Laplace approximation. Thus, $p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))$ is the first-order Laplace approximation to the marginal log-likelihood $l(\boldsymbol{\theta})$ in (5.2) using the hierarchical log-likelihood function $l_c(\boldsymbol{\theta})$.

In fact, let $N_i = n_{obs,i} + m_i$ be the number of the response and covariate observations for individual i and let b be the dimension of $\boldsymbol{\omega}_i$. Assume that $N_i = O(N)$ uniformly for $i = 1, \ldots, n$, where $N = \min_i N_i$. Taking $k = N_i$, $kp(\mathbf{x}) = l_c^{(i)}(\boldsymbol{\theta})$, $\gamma = b$, and $\mathbf{x} = \boldsymbol{\omega}_i$ in the following Laplace approximation

$$\int e^{kp(\mathbf{x})} d\mathbf{x} = (2\pi/k)^{\gamma/2} \left| \frac{\partial^2 p(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=\hat{\mathbf{x}}} \right|^{-\frac{1}{2}} e^{kp(\hat{\mathbf{x}})} + O(k^{-1}),$$

where $\mathbf{x}$ is a γ-dimensional parameter vector and $\hat{\mathbf{x}}$ maximizes $kp(\mathbf{x})$, we can approximate the ith individual's contribution $l_i(\boldsymbol{\theta})$ to the overall log-likelihood $l(\boldsymbol{\theta})$ as

$$\begin{aligned}
l_i(\boldsymbol{\theta}) &= \log \int e^{l_c^{(i)}(\boldsymbol{\theta})} d\boldsymbol{\omega}_i = \log \int e^{N_i p(\boldsymbol{\omega}_i)} d\boldsymbol{\omega}_i \\
&= \log \left\{ \left(\frac{2\pi}{N_i}\right)^{b/2} \left| D(p(\boldsymbol{\omega}_i), \boldsymbol{\omega}_i) \mid_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i} \right|^{-1/2} e^{N_i p(\hat{\boldsymbol{\omega}}_i)} + O_p(N_i^{-1}) \right\} \\
&= \log \left\{ \left(\frac{2\pi}{N_i}\right)^{b/2} \left| \frac{1}{N_i} D(l_c^{(i)}(\boldsymbol{\theta}), \boldsymbol{\omega}_i) \mid_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i} \right|^{-1/2} e^{l_c^{(i)}(\boldsymbol{\theta})} + O_p(N_i^{-1}) \right\} \\
&= \log \left\{ \left| \frac{1}{2\pi} D(l_c^{(i)}(\boldsymbol{\theta}), \boldsymbol{\omega}_i) \Big|_{\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}} \right|^{-1/2} e^{l_c^{(i)}(\boldsymbol{\theta})} \mid_{\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}_i} + O_p(N_i^{-1}) \right\} \\
&= \log \left[\exp\{ p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) \} + O(N_i^{-1}) \right] \\
&= p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) + O(N_i^{-1}), \qquad (5.8)
\end{aligned}$$

in which the last step holds by Lemma 3.2 in Section 3.7. Hence, the log-likelihood $l(\boldsymbol{\theta})$ can be approximated as

$$\begin{aligned}
l(\theta) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) + O(N_i^{-1})] \\
&= p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta})) + \sum_{i=1}^{n} O(N_i^{-1}) \\
&= p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta})) + \sum_{i=1}^{n} O(N^{-1}) \\
&= p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta})) + nO(N^{-1}). \qquad (5.9)
\end{aligned}$$

As $N = \min_i N_i$ grows faster than n, the function $p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))$ approaches the marginal log-likelihood function $l(\boldsymbol{\theta})$, and hence an estimate of $\boldsymbol{\theta}$, which maximizes $p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))$, also maximizes $l(\boldsymbol{\theta})$. This lead to the following algorithm to obtain an approximate MLE of $\boldsymbol{\theta}$ called $\hat{\boldsymbol{\theta}}_{HL}$:

Step 1. Obtain an initial estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, R, A, B, \boldsymbol{\omega}) = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \delta^{2(0)}, R^{(0)}, A^{(0)}, B^{(0)}, \boldsymbol{\omega}^{(0)})$ based on a naive method, and an initial estimate of $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$ based on the dropout model with the random effects $\boldsymbol{\omega}^{(0)}$.

Step 2. Given the current parameter estimates $\boldsymbol{\theta}^{(t)}$, update random-effects estimates $\boldsymbol{\omega}_i^{(t+1)}$ by maximizing $l_c^{(i)}(\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\omega}_i$, $i = 1, \ldots, n$.

Step 3. Given the random-effects estimates $\boldsymbol{\omega}_i^{(t+1)}$, update the parameter estimates $\boldsymbol{\theta}^{(t+1)}$ by maximizing $p_{\boldsymbol{\omega}^{(t+1)}}(l_c(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$.

Step 4. Iterate between Step 2 and Step 3 until convergence.

We can use Fisher information to obtain the following approximate formula for the variance-covariance matrix of the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$

$$\text{Cov}(\hat{\boldsymbol{\theta}}_{HL}) = \left[-\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{HL}}.$$

Many optimization procedures evaluate the matrix $[-\partial^2 p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta}))/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T]$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{HL}$ (called Hessian matrix), from which it is easy to obtain $\text{Cov}(\hat{\boldsymbol{\theta}}_{HL})$.

5.4.2 Asymptotic Properties

Under suitable regularity conditions on $l(\boldsymbol{\theta})$, $g(\cdot)$, and $d(\cdot)$, we extend Vonesh (1996) to show in Section 5.7 that

$$(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = O_p \left[\max \left\{ n^{-\frac{1}{2}}, \left(\min_i N_i \right)^{-1} \right\} \right],$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. Thus, the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$ will be consistent only as both n and $(\min_i N_i) \to \infty$. Intuitively, the $n^{-\frac{1}{2}}$ term comes from standard asymptotic theory while the $(\min_i N_i)^{-1}$ term comes from the Laplace approximation.

Note that the accuracy of the first-order Laplace approximation to the log-likelihood function is $O\{n/(\min_i N_i)\}$, or, equivalently, $o(1)$ provided $(\min_i N_i)$ grows faster than n. In this case, $(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = O_p(n^{-\frac{1}{2}})$ with $\hat{\boldsymbol{\theta}}_{HL}$ being asymptotically equivalent to the "exact" MLE. This reflects the fact that, as the accuracy of the Laplace approximation to the log-likelihood increases, the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$ will behave more and more like the "exact"

MLE. However, we can decrease the growth rate of $(\min_i N_i)$ for the asymptotic normality of $\hat{\boldsymbol{\theta}}_{HL}$. In particular, as $(\min_i N_i)$ grows at a rate greater than $n^{\frac{1}{2}}$, the rate of consistency of $\hat{\boldsymbol{\theta}}_{HL}$ will still be $O_p(n^{-\frac{1}{2}})$ and the resulting estimate will be asymptotically equivalent to the "exact" MLE in the sense that it has the same asymptotic distribution as the "exact" MLE (see Section 5.7). We correct the claim by Vonesh (1996) that, as $(\min_i N_i)$ grows at a rate greater than $n^{\frac{1}{2}}$ but less than or equal to n, the rate of consistency will still be $O_p(n^{-\frac{1}{2}})$ but the resulting estimate will no longer be asymptotically equivalent to the "exact" MLE.

The proofs of the above arguments are given in Section 5.7.

5.5 Example and Simulation

5.5.1 Example

We use the same HIV dataset in Section 4.5 to illustrate our proposed methods in this chapter, but we use the *random-effect-based informative* dropout model here rather than the *outcome-based informative* dropout model in Section 4.5. These informative dropout models may be used for sensitivity analysis. We use the commonly-used *naive method*, which ignores measurement errors and missing data, for parameter starting values in the two proposed methods. See the data description in Section 4.5.1.

The Response and the Covariate Models

We consider the same HIV viral dynamic and CD4 measurement error models in Section 4.5.2. For completeness, we describe these models again here. For the viral load

process, we consider the following model

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \tag{5.10}$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + \beta_3 z^*_{ij} + b_{2i}, \tag{5.11}$$

$$\log(P_{2i}) = \beta_4 + b_{3i}, \qquad \lambda_{2ij} = w(t_{ij}) + h_i(t_{ij}), \tag{5.12}$$

where y_{ij} is the $\log_{10}$-transform of the viral load measurement for patient i at time t_{ij}, P_{1i} and P_{2i} are baseline values, λ_{1ij} and λ_{2ij} are viral decay rates, z^*_{ij} is the true (but unobservable) CD4 count, and $w(t_{ij})$ and $h_i(t_{ij})$ are nonparametric fixed- and random-effects functions (see Section 2.1). In order to reduce the number of nuisance parameters, we assume that the variance-covariance matrices A and B of the random effects are both diagonal matrices. To avoid very small (large) estimates, which may be unstable, we standardize the CD4 counts and rescale the original time t (in days) so that the new time scale is between 0 and 1.

As discussed in Section 2.1, we employ the linear combinations of natural cubic splines with percentile-based knots to approximate $w(t)$ and $h_i(t)$. Following Wu and Zhang (2002), we take the same natural cubic splines with $q \leq p$ in order to decrease the dimension of random effects. AIC and BIC criteria are used to determine the values of p and q, which leads to the following model for λ_{2ij} in (5.12) (see Table 4.1), with $p = 3$ and $q = 1$,

$$\lambda_{2ij} \approx \beta_5 + \beta_6\,\psi_1(t_{ij}) + \beta_7\,\psi_2(t_{ij}) + b_{4i}. \tag{5.13}$$

For the CD4 process, we consider empirical polynomial LME models, and choose the best fitted model based again on AIC/BIC values for each possible model. This is done based on the observed CD4 values, and is done separately from the response model for simplicity. The following quadratic polynomial LME model best fits the CD4 trajectory (see Table 4.2):

$$\text{CD4}_{il} = (\alpha_1 + a_1) + (\alpha_2 + a_2)\,u_{il} + (\alpha_3 + a_3)\,u^2_{il} + \epsilon_{il}, \tag{5.14}$$

where u_{il} is the time and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ are the population parameters and $\mathbf{a}_i = (a_{i1}, a_{i2}, a_{i3})^T$ are the random effects.

Random-effect-based Informative Dropout Models

In this study, dropout patients appear to have slower viral decay, compared with the remaining patients. Thus, dropouts are likely to be informative or nonignorable. So we need to assume a model for the dropout process in order to make valid likelihood inference. Note that we should avoid building too complicated a dropout model since a complicated model may become non-identifiable. Subject-area knowledge and preliminary checks suggest that dropout may be related to the random-effects components a_{i1}, a_{i2}, and b_{i2}, so we consider the following dropout model

$$\text{logit}[P(r_{ij} = 1|\mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\eta})] = \eta_1 + \eta_2 a_{i1} + \eta_3 a_{i2} + \eta_4 b_{i2}. \tag{5.15}$$

We assume independence of the r_{ij}'s to simplify the model. The dropout model (5.15) along with the dropout models in Section 4.5 can be used for sensitivity analysis.

Estimation Methods and Computation Issues

We estimate the model parameters using the two proposed "joint" model methods discussed in Sections 5.3 and 5.4. We denote the method in Section 5.3 by AP and the method in Section 5.4 by HL. The two proposed joint model methods need starting values for model parameters since they are implemented by a MCEM algorithm or by an iterative Laplace approximation to the log-likelihood function. We use the parameter estimates obtained by the naive method, which ignores measurement errors and missing data, as parameter starting values for the two joint model methods.

For the naive method, we use the SPLUS function *nlme*() and *lme*() to obtain parameter estimates and their default standard errors. For the proposed AP method, we assess the

convergence of the Gibbs sampler by examining time series plots and sample autocorrelation function plots. For example, Figures 5.1 and 5.2 show the time series and the autocorrelation function plots for b_2 associated with patient 14. From these figures, we notice that the Gibbs sampler converges quickly and the autocorrelations between successive generated samples are negligible after lag 15. Time series and autocorrelation function plots for other random effects show similar behaviors. Therefore, we discard the first 500 samples as the burn-in, and then we take one sample from every 20 simulated samples to obtain independent samples (see sampling methods in Section 5.3.3).

We start with $k_0 = 500$ Monte Carlo samples, and increase the Monte-Carlo sample size as the number t of EM iteration increases: $k_{t+1} = k_t + k_t/c$ with $c = 4$ (Booth and Hobert, 1999). Convergence criterion for these two joint model methods in our examples is that the relative change in the parameter estimates from successively iterations is smaller than 0.05. Convergence of the algorithms are considered to be achieved when the maximum percentage change of all estimates is less than 5% in two consecutive iterations.

We use the multivariate rejection sampling method for the AP method. Other sampling methods may also be applied and may be even more efficient. On a SUN Sparc work-station, the AP method took about 135 minutes to converge while the HL method took about 150 minutes to converge. The HL method took more time than the AP method mainly because all model parameters appear in the nonlinear function $p_{\hat{\omega}}(l_c(\boldsymbol{\theta}))$ and no separation of the parameters is possible. However, the HL method is always computationally feasible while the AP method sometimes may have convergence problems. Moreover, the HL method can be used to obtain good parameter starting values for the AP method.

Analysis Results

We estimate the model parameters using the two proposed joint model methods AP

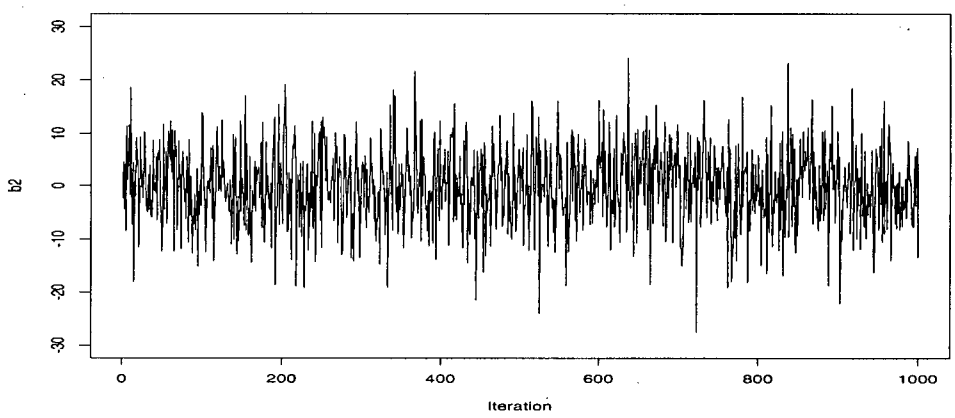

Figure 5.1: The time series plot for b_2 associated with patient 10.

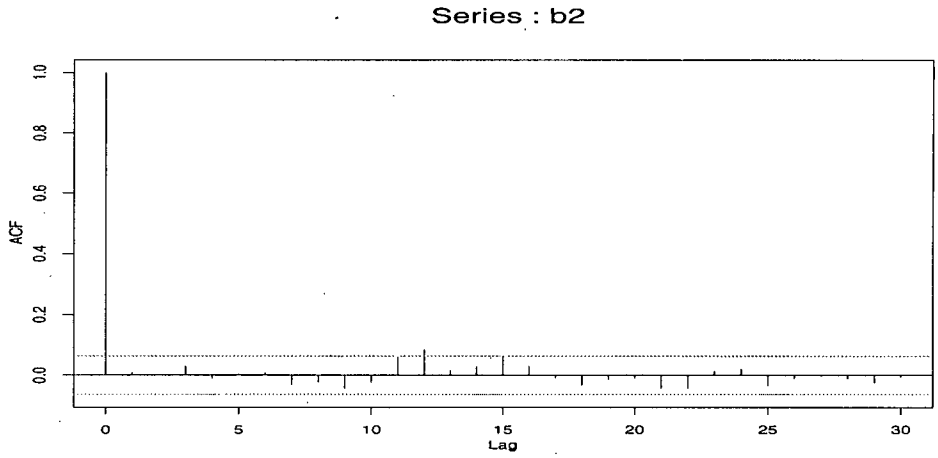

Figure 5.2: The autocorrelation function plot for b_2 associated with patient 14.

Table 5.1: Parameter estimates (standard errors) for the models in the example.

Method	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7	δ	R
AP	−.43	4.29	−3.90	11.72	66.65	1.53	6.93	−1.86	7.49	−2.36	.36	.51
	(.2)	(.5)	(.6)	(.2)	(4.3)	(4.6)	(.7)	(4.9)	(7.8)	(2.7)		
HL	−.41	4.32	−3.93	11.64	66.44	1.58	6.89	−1.92	7.46	−2.29	.35	.50
	(.1)	(.4)	(.5)	(.1)	(3.4)	(2.8)	(.6)	(4.8)	(7.5)	(2.7)		

Note: the estimated covariance matrices are $\hat{A} = \text{diag}(.62, 4.70, 4.41)$ *for AP,* $\hat{A} = \text{diag}(.51, 4.74, 4.53)$ *for HL.* $\hat{B} = \text{diag}(1.45, 91.62, 1.94, 20.16)$ *for AP, and* $\hat{B} = \text{diag}(1.42, 91.91, 1.58, 19.96)$ *for HL.*

and HL. We use the parameter estimates obtained by the naive method as the parameter starting values for the AP and the HL methods. We also tried several other parameter starting values for the proposed joint model methods. Different parameter starting values lead to roughly same parameter estimates in both the AP and the HL methods.

Table 5.1 presents the resulting parameter estimates and standard errors based on the *random-effect-based informative* model (5.15). We find that the two joint model methods provide similar parameter estimates. Comparing the *random-effect-based informative* dropout model with the *outcome-based informative* dropout model I in (4.13), we find that the resulting estimates are similar. This indicates again that the estimation may be robust against the nonignorable dropout models. Although some estimates in Table 5.1 are not statistically significant, the values of the estimates may still provide useful information about viral load and CD4 trajectories. The estimates of parameters $\boldsymbol{\eta}$ based on the AP method (or the HL method) are -2.32, .31, $-.05$,, and $-.07$ (or -2.41, .27, $-.04$, and $-.08$) respectively, with all p-values less than .00001, which also indicates that the dropouts may be nonignorable (or informative) since the missingness may depend on the unobservable random effects. The estimates of η_2, η_3, and η_4 indicate that dropout patients seems to have higher baseline CD4 count, decrease in CD4 count faster over time, and have slower first decay rate than the remaining patients.

5.5.2 The Simulation Study

We evaluate the proposed methods (AP and HL) for the *random-effect-based informative* model (5.15) via simulation. The response and covariate models, the *random-effect-based informative* dropout model, and the measurement time points used in the simulation are all the same as those in the example in the previous section (i.e., (5.10) – (5.14)). We choose appropriate values of $\boldsymbol{\eta}$ to mimic certain missing rate, and we use the SPLUS function *sample*() to generate binary data r_{ij} based on the values of parameters $\boldsymbol{\eta}$ and the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$. If $r_{ij} = 1$, then y_{ij} is deleted, and if $r_{ij} = 0$, y_{ij} is considered to be observed.

In the simulations, the true values of model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are shown in Table 5.2, and the other true parameter values are $\delta = .2$, $R = .4$, $A = \text{diag}(.5, 3, 2)$, and $B = \text{diag}(1, 9, 2, 4)$. We set $\boldsymbol{\eta} = (-1.4, 0.1, -0.1, -0.1)^T$ to get an average missing rate of 20%. We always regard the first two responses on each individual as observed, i.e., every individual has at least two observed responses.

We simulated 100 data sets and calculated averages of the resulting estimates and their standard errors based on each of the two methods. We compare the methods by comparing their biases and mean-square-errors (MSEs). Here, bias and MSE are assessed in terms of percent relative bias and percent relative root mean-squared error, as defined in Section 4.6. Since AP method sometimes offers computational problems, such as slow or non-convergence, the 100 sets of parameter estimates are obtained from 128 data sets.

From the simulation results in Tables 5.2 and 5.3, we see that the two proposed joint model methods (AP and HL) perform well. The AP method performs better than the HL method in the sense that the AP yields smaller relative MSE and bias than the HL method. The HL method also performs reasonably well and it is always computationally feasible. Therefore, the HL method may be a good alternative method when the AP method exhibits computational difficulties, and the HL method can also be used to obtain good parameter

Table 5.2: Simulation results for the parameter estimates (standard errors) for the estimation methods AP and HL.

Parameter	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7
True Value	–.4	4.0	–4.0	12.0	67.0	1.5	7.0	–2.0	8.0	–3.0
AP	–.40	4.01	–3.98	11.99	66.95	1.49	6.99	–2.01	8.04	–2.97
	(.2)	(.5)	(.6)	(.1)	(1.4)	(1.6)	(.3)	(2.1)	(3.2)	(1.1)
HL	–.38	3.96	–3.93	11.99	67.10	1.56	6.95	–2.08	8.09	–2.89
	(.1)	(.4)	(.5)	(.1)	(1.0)	(1.3)	(.3)	(1.8)	(2.9)	(1.0)

Table 5.3: Simulation results for bias and MSE of the parameter estimates for the estimation methods AP and HL.

Parameter	α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	β_6	β_7
True Value	–.4	4.0	–4.0	12.0	67.0	1.5	7.0	–2.0	8.0	–3.0
				Bias						
AP	1.59	.15	.95	–.07	–.1	–.44	–.13	–.53	.46	.25
HL	2.17	–1.02	1.31	–.08	.22	1.08	–.26	–3.78	.62	.93
				MSE						
AP	7.93	6.46	7.39	1.66	1.74	66.69	1.98	71.84	28.95	22.26
HL	9.07	8.33	9.31	1.69	2.43	93.67	4.24	98.75	37.30	31.71

Note: Bias = Percent bias $= 100 \times \text{bias}_j/|\beta_j|$; *MSE = Percent* $\sqrt{\text{MSE}} = 100 \times \sqrt{\text{MSE}_j}/|\beta_j|$.

starting values for the AP method.

5.6 Discussion

We have proposed two approximate likelihood methods for semiparametric NLME models with *random-effect-based informative* dropouts and covariate measurement errors and missing data, implemented by a Monte Carlo EM algorithm combined with Gibbs sampler or by an iterative Laplace approximation to the log-likelihood function respectively. The first method may be more accurate than the second method but it sometimes may offer computational difficulties such as slow or non-convergence, especially when the dimensions of random effects are not small. The second method is always computationally feasible but may be less accurate than the first method. The second method may be used as a reasonable alternative when the first method has convergence problems or it may be used to provide excellent parameter starting values for the first method. Simulation studies indicate that both methods produce satisfactory results.

Although it does not need to generate Monte Carlo samples for random effects, the second method may not be computationally more efficient than the first method. A possible reason is that there are too many model parameters appear in the nonlinear functions in optimization procedures and no separation of the parameters is possible. A possible solution is to use Bayesian method to address this problem. Specifically, we may assume the known hyperprior distributions for the model parameters.

In many longitudinal data sets, dropouts, censoring, measurement errors, and missing covariates are all present simultaneously. To our knowledge, there are almost no unified methods in the literature which address these problems simultaneously. Wu (2002) proposed a unified method to address censoring and measurement errors simultaneously and

showed that the proposed method offered significant improvement over existing methods currently in use. The ideas in Wu (2002) and in this chapter can be extended to address dropouts, censoring, measurement errors, and missing covariate simultaneously in semiparametric/nonparametric NLME models.

5.7 Appendix: Asymptotic Properties of the Approximate MLE $\hat{\boldsymbol{\theta}}_{\mathrm{HL}}$ in Section 5.4

5.7.1 Consistency

We will show that the following result

$$(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = O_p\left[\max\left\{n^{-\frac{1}{2}}, \left(\min_i N_i\right)^{-1}\right\}\right]$$

holds under the usual regularity conditions on $l(\boldsymbol{\theta})$, $g(\cdot)$ and $d(\cdot)$, where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$.

Proof. Let $\hat{\boldsymbol{\omega}}_i$ maximize $l_c^{(i)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\omega}_i$ for fixed $\boldsymbol{\theta}$. Denote $N_i = n_{obs,i} + m_i$. Suppose that $N_i = O(N)$ uniformly for $i = 1, \ldots, n$, where $N = \min_i N_i$. Based on (5.8) in Section 5.4.3, the ith individual's contribution $l_i(\boldsymbol{\theta})$ to the overall log-likelihood may be approximated as

$$l_i(\boldsymbol{\theta}) = p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) + O(N_i^{-1}) = p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta})) + O(N^{-1}).$$

Hence, the log-likelihood $l(\boldsymbol{\theta})$ can be written as (see (5.9))

$$l(\boldsymbol{\theta}) = l^*(\boldsymbol{\theta}) + O\{n\, N^{-1}\}, \tag{5.16}$$

where $l^*(\boldsymbol{\theta}) = p_{\hat{\boldsymbol{\omega}}}(l_c(\boldsymbol{\theta})) = \sum_{i=1}^n p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}))$. Let $\mathbf{u}^*(\boldsymbol{\theta}) = \partial l^*(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\theta}}_{HL}$ be the approximate maximum likelihood estimate satisfying $\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) = \mathbf{0}$. Under suitable regularity

conditions on $l(\boldsymbol{\theta})$ and assuming $\hat{\boldsymbol{\theta}}_{HL}$ is an interior point in a neighborhood containing $\boldsymbol{\theta}_0$, Taylor's theorem tells us that there exists a vector $\tilde{\boldsymbol{\theta}}$ on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{HL}$ such that

$$n^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) = n^{-1}\mathbf{u}(\boldsymbol{\theta}_0) + n^{-1}M(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0), \tag{5.17}$$

where $\mathbf{u}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $M(\boldsymbol{\theta}) = \partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ are the first and second order derivatives of the true but intractable marginal log-likelihood $l(\boldsymbol{\theta})$. The first term $n^{-1}\mathbf{u}(\boldsymbol{\theta}_0)$ on the right of (5.17) is

$$\frac{1}{n}\mathbf{u}(\boldsymbol{\theta}_0) = \frac{1}{n}\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Given sufficient regularity conditions on $l(\boldsymbol{\theta})$, we know from the Lindeberg Central Limit Theorem that

$$\frac{1}{\sqrt{n}}\mathbf{u}(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)), \tag{5.18}$$

where the matrix $\bar{I}(\boldsymbol{\theta}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I_i(\boldsymbol{\theta})$ and $I_i(\boldsymbol{\theta})$ is the information matrix for individual i. That implies

$$\frac{1}{\sqrt{n}}\mathbf{u}(\boldsymbol{\theta}_0) = O_p(1) \Longleftrightarrow \frac{1}{n}\mathbf{u}(\boldsymbol{\theta}_0) = O_p(n^{-1/2}).$$

The matrix $n^{-1}M(\tilde{\boldsymbol{\theta}})$ on the right of (5.17) is

$$\frac{1}{n}M(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \xrightarrow{p} -\bar{I}(\tilde{\boldsymbol{\theta}}), \tag{5.19}$$

by the Law of Large Numbers. Since $\bar{I}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta}$, the probability that the matrix $n^{-1}M(\tilde{\boldsymbol{\theta}})$ is invertible tends to 1. By writing $n^{-1}M(\tilde{\boldsymbol{\theta}}) = -\bar{I}(\tilde{\boldsymbol{\theta}}) + o_p(1)$ and applying Lemma 3.2 in Section 3.7 to the inverse function, we have

$$[n^{-1}M(\tilde{\boldsymbol{\theta}})]^{-1} = -\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1). \tag{5.20}$$

Given suitable regularity conditions on $g(\cdot)$ and $d(\cdot)$, for example that third order derivatives exist and are continuous in an open neighborhood about $\boldsymbol{\theta}_0$, application of Lemma 3.2 in Section 3.7 to the partial derivative function in the expression (5.16) leads to

$$n^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) = n^{-1}\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) + O(N^{-1}). \tag{5.21}$$

Note that $\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) = \mathbf{0}$ and $n^{-1}\mathbf{u}(\boldsymbol{\theta}_0) = O_p(n^{-\frac{1}{2}})$. From (5.17), we have

$$\begin{aligned}
& n^{-1}M(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = n^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) - n^{-1}\mathbf{u}(\boldsymbol{\theta}_0) \\
\Longrightarrow\ & (\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = [n^{-1}M(\tilde{\boldsymbol{\theta}})]^{-1}[n^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) - n^{-1}\mathbf{u}(\boldsymbol{\theta}_0)] \\
\Longrightarrow\ & (\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = (-\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1))[n^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) - n^{-1}\mathbf{u}(\boldsymbol{\theta}_0)] \qquad \text{(by (5.20))} \\
\Longrightarrow\ & (\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = (-\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1))[n^{-1}\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) + O(N^{-1}) + O_p(n^{-\frac{1}{2}})] \qquad \text{(by (5.21))} \\
\Longrightarrow\ & (\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = -\bar{I}(\tilde{\boldsymbol{\theta}})^{-1}\, O_p\Big[\max\Big\{n^{-\frac{1}{2}},\, N^{-1}\Big\}\Big] + o_p\Big[\max\Big\{n^{-\frac{1}{2}},\, N^{-1}\Big\}\Big] \\
\Longrightarrow\ & (\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) = O_p\left[\max\Big\{n^{-\frac{1}{2}}, \Big(\min_i N_i\Big)^{-1}\Big\}\right].
\end{aligned}$$

Finally, let $\hat{\boldsymbol{\theta}}_{ML}$ denote the "exact" maximum likelihood estimate with $\mathbf{u}(\hat{\boldsymbol{\theta}}_{ML}) = \mathbf{0}$. Let $\min_i N_i = O(n^\tau)$ for $\tau > 1$ so that the accuracy of the Laplace approximation to the marginal log-likelihood is approximately $O(n^{1-\tau}) = o(1)$ from the formula (5.16). Then, under the same regularity conditions as before, by multiplying n on the both sides of the equation (5.21) and noting that $\mathbf{u}(\hat{\boldsymbol{\theta}}_{ML}) = \mathbf{0}$, we have

$$\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) = \mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) + o_p(1) = \mathbf{0} + o_p(1) \equiv \mathbf{u}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(1).$$

Thus $\mathbf{u}(\hat{\boldsymbol{\theta}}_{HL}) - \mathbf{u}(\hat{\boldsymbol{\theta}}_{ML}) = o_p(1)$ and hence $\hat{\boldsymbol{\theta}}_{HL}$ is asymptotically equivalent to the "exact" maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{ML}$. □

5.7.2 Asymptotic Normality of $\hat{\boldsymbol{\theta}}_{\mathbf{HL}}$

In this section, we will show that as N grows at a rate greater than $n^{\frac{1}{2}}$, i.e., $N = O(n^\tau)$ for $\tau > \frac{1}{2}$, the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$ and the "exact" MLE $\hat{\boldsymbol{\theta}}_{ML}$ have the same asymptotic

distribution.

Proof. Noting that the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$ satisfies a set of equations $\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) = \mathbf{0}$, we take a first-order Taylor series expansion of $\mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL})$ around the true parameter $\boldsymbol{\theta}_0$

$$\mathbf{0} = \mathbf{u}^*(\hat{\boldsymbol{\theta}}_{HL}) \;=\; \mathbf{u}^*(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ is on the line segment joining $\boldsymbol{\theta}_0$ to $\hat{\boldsymbol{\theta}}_{HL}$, which implies

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) &= \left[-\frac{1}{n}\frac{\partial \mathbf{u}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}\right]^{-1}\left[\frac{1}{\sqrt{n}}\mathbf{u}^*(\boldsymbol{\theta}_0)\right] \\ &= \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}}\right]. \qquad (5.22)\end{aligned}$$

Now we consider the two product terms on the right of (5.22). Applying Lemma 3.2 in Section 3.7 to the first and second partial derivative functions in the expression in (5.16), we know that, for any fixed $\boldsymbol{\theta}$,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\mathbf{u}^*(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{u}(\boldsymbol{\theta}) + O(n^{\frac{1}{2}}N^{-1}) \\ \iff\quad &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + O(n^{\frac{1}{2}}N^{-1}), \qquad (5.23)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\frac{\partial \mathbf{u}^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \frac{1}{n}\frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} + O(N^{-1}) \\ \iff\quad &\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + O(N^{-1}) \qquad (5.24)\end{aligned}$$

Assume that $N = O(n^\tau)$, where $\tau > \frac{1}{2}$. Then $O(n^{\frac{1}{2}}N^{-1}) = O(n^{\frac{1}{2}-\tau}) = o(1)$. From (5.23) and (5.24), we have

$$\begin{aligned}\lim_{n\to\infty}\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}} &= \lim_{n\to\infty}\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}, \\ \lim_{n\to\infty}\tfrac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_c^{(i)}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} &= \lim_{n\to\infty}\tfrac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}.\end{aligned} \qquad (5.25)$$

Note that $\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0 = O_p[\max\{n^{-\frac{1}{2}}, N^{-1}\}] = O_p(n^{-\frac{1}{2}})$, i.e., $\hat{\boldsymbol{\theta}}_{HL}$ is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\theta}_0$. Since $\boldsymbol{\theta}^*$ is on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{HL}$, $\boldsymbol{\theta}^* \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$. Under the same regularity conditions as before, it follows from (5.18) and (5.19) that

$$\begin{aligned} &\tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)), \\ &-\tfrac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \xrightarrow{p} \bar{I}(\boldsymbol{\theta}_0). \end{aligned} \tag{5.26}$$

Combining the results in (5.25) and (5.26) and using Slutsky's theorem, we can show that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{HL} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)^{-1}),$$

which implies that when $N = O(n^\tau)$ for $\tau > \frac{1}{2}$, the approximate MLE $\hat{\boldsymbol{\theta}}_{HL}$ and the "exact" MLE $\hat{\boldsymbol{\theta}}_{ML}$ have the same asymptotic distribution. □

Chapter 6

Conclusions and Future Research

6.1 Conclusions

In this thesis, we have developed approximate maximum likelihood inference in the following three problems: (1). semiparametric NLME models with measurement errors and missing data in time-varying covariates; (2). semiparametric NLME models with covariate measurement errors and *outcome-based informative* missing responses; (3). semiparametric NLME models with covariate measurement errors and *random-effect-based informative* missing responses. Measurement errors, dropouts, and missing data are addressed simultaneously in a unified way. For each problem, we have proposed two joint model methods to simultaneously obtain approximate maximum likelihood estimates (MLEs) of all model parameters. The first method, implemented by a Monte Carlo EM algorithm, may be more accurate than the second method but it may be computationally very intensive and sometimes may offer computational difficulties such as slow or non-convergence, especially when the dimensions of random effects are not small. The second method, which approximates joint log-likelihood functions by using a first-order Taylor expansion or by using a first-order Laplace approxima-

tion, is computationally more appealing, but it may be less accurate than the first method. The performance of the second method may need further investigation. We have showed some asymptotic results for the estimates based on the second method. The second method may be used as a reasonable alternative when the first method has convergence problems or be used to provide excellent parameter starting values for the first method.

Simulation results have shown that all proposed methods perform better than the commonly used two-step method and the naive method which ignores measurement errors, in the sense that the proposed methods yield smaller bias and MSE. In particular, the commonly used two-step method may under-estimate standard errors, which is consistent with analytic results, and the naive method may under-estimate covariate effects and poorly estimate other parameters.

6.2 Future Research Topics

Finally, we discuss possible future work relevant to this thesis as follows.

1. In many longitudinal studies such as HIV viral dynamics, another common problem is that the response measurements may be subject to left censoring due to a detection limit. Censored responses in practice were often substituted by the detection limit or half the detection limit (Wu and Ding, 1999; Wu and Wu, 2001), which may lead to substantial biases in the results (Wu, 2002). In the presence of both dropouts and censoring, unified approaches which address these problems simultaneously in semiparametric/nonparametric NLME models are needed in order to make reliable statistical inference.

2. In many longitudinal datasets, dropouts, censoring, measurement errors, and missing covariates are all present simultaneously. To our knowledge, there are almost no

unified methods in the literature which address these problems simultaneously. Wu (2002) proposed a unified method to address censoring and measurement errors simultaneously and showed that the proposed method offered significant improvement over existing methods currently in use. The ideas in Wu (2002) and in this thesis can be extended to address dropouts, censoring, measurement errors, and missing covariates simultaneously in semiparametric/nonparametric NLME models.

3. For the response process, we only consider semiparametric nonlinear mixed-effects models with independent and normal distributed error terms $\mathbf{e}_i$. In the future, we may consider more complicated covariance structure for $\mathbf{e}_i$ such as an AR(1) structure.

4. In our study, we only consider semiparametric nonlinear mixed-effects models for normal data. Generally, our proposed methods may be extended to other models, such as semiparametric/nonparametric generalized linear mixed-effects models and semiparametric/nonparametric generalized nonlinear mixed-effects models.

5. Computational efficiency is an important issue in our study. Multivariate rejection sampling methods have been used in our data analyses and simulation. In general, other sampling methods, such as adaptive rejection sampling methods and importance sampling methods, may also be used and may be even more efficient. We plan to compare computational efficiency among several sampling methods in our current setting.

6. In our alternative methods, we have approximated log-likelihood functions by using a first-order Taylor expansion or by using a first-order Laplace approximation. Sometimes, these are not necessarily accurate approximations. In the future, we may investigate better approximations, such as higher order Taylor expansions and Laplace approximations.

7. One problem in our models is that there are too many parameters. If the data are not rich enough, the proposed methods may have convergence problems and identifiability problems. We plan to develop Bayesian methods for our problems.

References

Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, **29**, 813-828.

Amemiya, T. (1983). Nonlinear regression models. In *Handbook of Econometrics, Volume I*, Z. Griliches and M. D. Intriligator, eds., pp.333-389. Noth Holland, Amsterdam.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* New York: Chapman and Hall.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed models likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Ser. B,* **61**, 265-285.

Bradley, R. A., and Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, **49**, 205-214.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, **90**, 242-252.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurements Data.* Chapman & Hall.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Demidenko, E. (2004). *Mixed Models Theory and Applications.* John Wiley & Sons.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **39**, 1-38.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with Discussion). *Applied Statistics*, **43**, 49-93.

Ding, A. and Wu, H. (2001). Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics,* **2(1)**, 13-29.

Euband, R. L. (1988). *Spline smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, **91**, 99-108.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Geweke, J. (1996). *Handbook of Computational Economics.* Amsterdam: North-Holland.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337-348.

Green, P. J. and Solverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Grossman, Z., Polis, M., Feinberg, M. B., Grossman, Z., Levi, I., Jankelevich, S., Yarchoan, R., Boon, J., De Wolf, F., Lange, J. M. A., Goudsmit, J., Dimitrov, D. S., and Paul, W. E. (1999). Ongoing HIV dissemination during HAART. *Nature Medicine*, **5**, 1099-1103.

Higgins, D. M., Davidian, M., and Giltinan, D. M. (1997). A two-step approach to measurement error in time-dependent covariates in nonlinear mixed-effects models, with application to IGF-I pharmacokinetics. *Journal of the American Statistical association*, **92**, 436-448.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, **373**, 123-126.

Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551-564.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Ser. B*, **61**, 173-190.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussions). *Journal of the American Statistical Association*, **96**, 1272-1298.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Ser. B*, **58**, 619-678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987-1006.

Liang, H., Wu, H., and Carroll, R. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement errors. *Biostatistics*, **4**, 297-312.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673-687.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, **90**, 1112-1121.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, **44**, 226-233.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162-170.

McLachlan, G. J. and Krishnan, T. (1997). *The EM-Algorithm and Extension.* New York, Wiley.

Ogden, R. T. and Tarpey, T. (2006). Estimation in Regression models with externally estimated parameters. *Biostatistics*, **7**, 115-129.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, **271**, 1582-1586.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12-35.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed-effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Seber, G. A. F. (1984). *Multivariate Observations.* New York: Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Shah, A., Laird, N., and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, **92**, 775-779.

Ten Have, T. R., Pulkstenis, E., Kunselman, A., and Landis, J. R. (1998). Mixed effects logistics regression models for longitudinal binary response data with informative dropout. *Biometrics*, **54**, 367-383.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, **83**, 447-452.

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

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, **97**, 271-283.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, **85**, 699-704.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791-795.

Wu, H., Kuritzkes, D. R., McClernon, D. R., Kessler, H., Connick, E., Landay, A., Spear, G., Heath-Chiozzi, M., Rousseau, F., Fox, L., Spritzler, J., Leonard, J. M., and Lederman, M. M. (1999). Characterization of viral dynamics in human immuno-deficiency virus type 1-infected patients treated with combination antiretroviral therapy: relationships to host factors, cellular restoration and virological endpoints. *Journal of Infectious Diseases*, **179**, 799-807.

Wu, H. and Ding, A. (1999). Population HIV-1 dynamics in vivo: application models and inferential tools for virological data from AIDS clinical trials. *Biometrics*, **55**, 410-418.

Wu, H. (2005). Statistical Methods for HIV Dynamic Studies in AIDS Clinical Trials. *Statistical Methods in Medical Research*, **14**, 171-192.

Wu, H. and Zhang, J. (2002). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine*, **21**, 3655-3675.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of the American Statistical Association*, **97**, 955-964.

Wu, L. (2004). Exact and approximation inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association*, **99**, 700-709.

Wu, L. and Wu, H. (2001). A multiple imputation method for missing covariates in non-linear mixed-effect models, with application to HIV dynamics. *Statistics in Medicine*, **20**, 1755-1769.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-188.