

**Nonlinear Mixed Effects Models with Dropout and  
Missing Covariates When the Dropout Depends  
on the Random Effects**

by

Shijun Song

B.Sc., Peking University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
**Master of Science**

in

THE FACULTY OF GRADUATE STUDIES  
(Statistics)

**The University of British Columbia**

August 2005

© Shijun Song, 2005

# Abstract

Nonlinear mixed effects models (NLMEs) are very popular in many longitudinal studies such as HIV viral dynamic studies, pharmacokinetics analyses, and studies of growth and decay. In these studies, however, missing data problems often arise, which make some statistical analyses complicated. In this thesis, we proposed an exact method and an approximate method for NLMEs with *random-effects based* informative dropouts and missing covariates, and propose methods for simultaneous inference. Monte Carlo EM algorithms are used in both methods. The approximate method, which is based on a Taylor series expansion, avoids sampling the random effects in the E-step and thus reduces the computation burden substantially. To illustrate the proposed methods, we analyze two real datasets. The exact method is applied to a dataset with covariates and a dataset without covariates. The approximate method is applied to the dataset without covariates. The result shows that, for both datasets, dropouts may be correlated with individual random effects. Ignoring the missingness or assuming ignorable missingness may lead to unreliable inferences. A simulation study is performed to evaluate the two proposed methods under various situations.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>Dedication</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Longitudinal Data Analysis . . . . .	1
1.1.1 Longitudinal Studies . . . . .	1
1.1.2 Approaches to Longitudinal Data Analysis . . . . .	3
1.2 Missing Data Problems . . . . .	5
1.2.1 Missing Covariates and Responses . . . . .	5
1.2.2 Classification of Missing Value Mechanisms . . . . .	5
1.2.3 Literature on Missing Data Problems . . . . .	7
1.3 Motivating Examples . . . . .	8

1.4	Objectives and Outline . . . . .	11
<b>2</b>	<b>Nonlinear Mixed Effects Models</b>	<b>12</b>
2.1	Introduction . . . . .	12
2.2	Nonlinear Mixed Effects Models . . . . .	12
2.3	Literature Review on NLME Models with Informative Missing Data .	15
<b>3</b>	<b>An Exact Method for NLME Models with Informative Dropout and Missing Covariates</b>	<b>17</b>
3.1	Introduction . . . . .	17
3.2	The Models . . . . .	18
3.3	A Monte Carlo EM Method . . . . .	20
3.3.1	E-step . . . . .	22
3.3.2	M-step . . . . .	25
3.4	Sampling Methods . . . . .	27
3.4.1	Gibbs Sampler . . . . .	27
3.4.2	Multivariate Rejection Algorithm . . . . .	28
3.4.3	Importance Sampling . . . . .	29
3.5	Convergence . . . . .	30
<b>4</b>	<b>An Approximate Method for NLME Models with Informative Dropout and Missing Covariates</b>	<b>33</b>
4.1	Introduction . . . . .	33
4.2	The Approximate Method . . . . .	35
4.2.1	E-step . . . . .	36
4.2.2	M-step . . . . .	45
4.3	Monte Carlo Sampling . . . . .	45

<b>5</b>	<b>Covariates Models and Dropout Models</b>	<b>47</b>
5.1	Introduction . . . . .	47
5.2	Covariate Models . . . . .	48
5.3	Dropout Models . . . . .	49
5.4	Sensitivity Analyses . . . . .	50
<b>6</b>	<b>Data Analysis</b>	<b>52</b>
6.1	Introduction . . . . .	52
6.2	Example 1 . . . . .	53
6.2.1	Data Description . . . . .	53
6.2.2	Models . . . . .	55
6.2.3	Analysis and Results . . . . .	57
6.2.4	Sensitivity Analysis . . . . .	59
6.2.5	Conclusion . . . . .	61
6.3	Example 2 . . . . .	62
6.3.1	Data Description . . . . .	62
6.3.2	Models . . . . .	64
6.3.3	Analysis and Results . . . . .	65
6.3.4	Sensitivity Analysis . . . . .	66
6.3.5	Conclusion . . . . .	67
6.4	Computation Issues . . . . .	67
<b>7</b>	<b>Simulation Study</b>	<b>69</b>
7.1	Introduction . . . . .	69
7.2	Design of the Simulation Study . . . . .	70
7.2.1	Models . . . . .	70

7.2.2	Comparison Criteria . . . . .	71
7.3	Simulation Results . . . . .	71
7.3.1	Comparison of Methods with Varying Missing Rates . . . . .	71
7.3.2	Comparison of Methods with Different Random Effects Covari- ances . . . . .	73
7.3.3	Comparison of Methods with Varying Intra-individual Measure- ments . . . . .	74
7.3.4	Comparison of Methods with Different Variances . . . . .	75
7.4	Conclusions . . . . .	76
8	Conclusion and Discussion	78
	References	80

# List of Tables

6.1	Data summary of Example 1. . . . .	53
6.2	Estimations for response model parameters. (Example 1) . . . . .	58
6.3	Estimations for dropout model parameters. (Example 1) . . . . .	59
6.4	Sensitivity analyses for dropout models. (Example 1) . . . . .	60
6.5	Data summary of Example 2. . . . .	62
6.6	Estimates for dynamic model parameters in Model (6.11) and (6.12). (Example 2) . . . . .	65
6.7	Estimations for dropout model parameters. (Example 2) . . . . .	66
6.8	Sensitivity analyses for dropout models. (Example 2) . . . . .	67
7.1	Simulation results for varying missing rates. . . . .	73
7.2	Simulation results for different covariance matrices for random effects. . . . .	74
7.3	Simulation results for varying intra-individual measurements. . . . .	75
7.4	Simulation results for varying variances. . . . .	76

# List of Figures

1.1	Hypothetical data on the relationship between height and age. . . . .	2
1.2	Viral loads of four randomly selected patients. . . . .	10
6.1	Viral loads of four randomly selected patients. (Example 1) . . . . .	54
6.2	Q-Q plots for covariates (Example 1) . . . . .	56
6.3	Viral loads of four randomly selected patients. (Example 2) . . . . .	63



# Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Lang Wu, for his excellent guidance and immense help during my study at the University of British Columbia. Without his support, expertise and patience, this thesis would not have been completed. Also, I would like to thank my co-supervisor and second reader, Dr. Harry Joe, for his invaluable comments and suggestions on this thesis. I would also like to thank Ms. Kunling Wu, a previous student of Dr. Lang Wu. Her master thesis and personal advice benefit me a lot in completion of this thesis.

Furthermore, I would like to thank Dr. John Petkau for his invaluable advice on my consulting projects, which benefit me very much in the past and future. I thank all the faculty and staff in Department of Statistics at the University of British Columbia for providing such a nice academic environment. I should also thank all the graduate students in the Department of Statistics for making my study so enjoyable.

Most importantly I would like to thank my parents for loving me and believing in me. Their love, constant support and encouragement push me to be the best at everything I do.

SHIJUN SONG

*The University of British Columbia*

*August 2005*

To my parents.

# Chapter 1

## Introduction

### 1.1 Longitudinal Data Analysis

#### 1.1.1 Longitudinal Studies

The key characteristic of a longitudinal study is that individuals are measured repeatedly over time. Longitudinal studies differ from cross-sectional studies, in which a single outcome is measured for each individual. In many studies, especially in clinical trials, longitudinal data are very common. Even when it is possible to address the same scientific questions in a longitudinal or cross-sectional study, there may be many advantages in addressing them in a longitudinal study. An example, which can illustrate this idea, is Figure 1.1. In Figure 1.1(a), height is plotted against age for a hypothetical cross-sectional study of boys. Height appears to be shorter among older boys. In Figure 1.1(b), we connect the data points from each individual. Now, it is clear that everyone's height increases with age. This example shows that longitudinal studies can distinguish changes over time within individuals from differences among people in their baseline levels. Cross-sectional studies cannot. Longitudinal data can

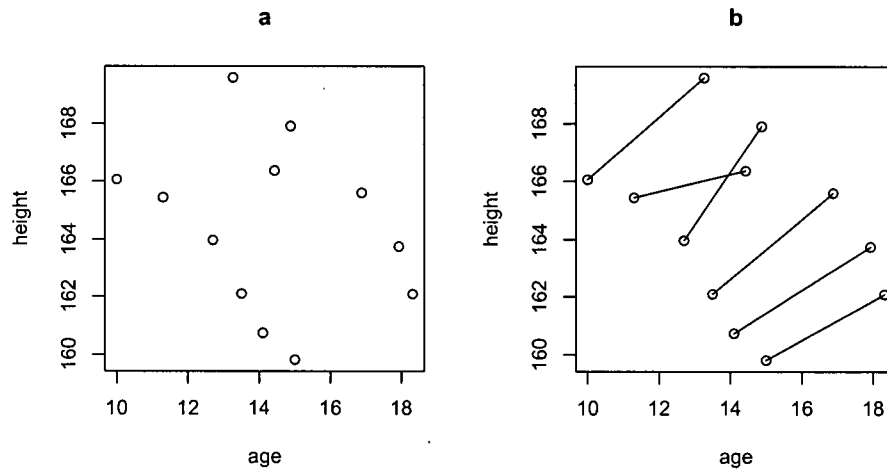


Figure 1.1: Hypothetical data on the relationship between height and age.

be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each person from historical records. Longitudinal data require special statistical methods because the set of observations on one subject tends to be inter-correlated. This correlation must be taken into account to draw valid inferences.

Correlation is also taken into account when analyzing a single long time series of measurements. In most time series studies, there is only one series available and people usually try to find clues and draw conclusions from that series itself. Analysis of longitudinal data tends to be simpler because subjects are usually assumed independent. Valid inferences can be made by borrowing information across people. That is, the consistency of a pattern across subjects is the basis for substantive conclusions. For this reason, inferences from longitudinal studies can be made more robust to model assumptions than those from time series data, particularly to assumptions about the nature of the correlation.

### 1.1.2 Approaches to Longitudinal Data Analysis

Let  $y_{ij}$  represent a response variable and  $\mathbf{x}_{ij}$  represent a  $p \times 1$  vector of  $p$  explanatory variables observed at time point  $t_{ij}$ , for measurement  $j$  on subject  $i$ ,  $j = 1, \dots, n_i$ ,  $i = 1, \dots, N$ . The mean and variance of  $y_{ij}$  are represented by  $E(y_{ij}) = \mu_{ij}$  and  $\text{Var}(y_{ij}) = v_{ij}$ . The set of repeated outcomes for subject  $i$  are collected into an  $n_i \times 1$  vector,  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$ , with mean  $E(\mathbf{y}_i) = \boldsymbol{\mu}_i$  and  $n_i \times n_i$  covariance matrix  $\text{Var}(\mathbf{y}_i) = V_i$ , where the  $(j, k)$  element of  $V_i$  is the covariance between  $y_{ij}$  and  $y_{ik}$ , denoted by  $\text{Cov}(y_{ij}, y_{ik}) = v_{ijk}$ . The covariate matrix for the  $i$ th subject is denoted as  $X_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^T$ , an  $n_i \times p$  matrix. We use  $R_i$  for the  $n_i \times n_i$  correlation matrix of  $\mathbf{y}_i$ . The responses for all subjects are denoted as  $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_N^T)^T$ , which is an  $m \times 1$ -vector with  $m = \sum_{i=1}^N n_i$ . The covariates for all units are referred to as  $X = (X_1^T, \dots, X_N^T)^T$ , which is an  $m \times p$  matrix.

There are three approaches to longitudinal data analysis. The first approach, which is often called *marginal models*, is to model univariate responses ignoring dependence. Marginal methods are mainly used for regression with dependent data with the main interest is inference for the regression parameters. For example, in a clinical trial the difference between control and treatment is most important, not the difference for any one individual.

A second approach, the *random effects model* approach, assumes that correlation arises among repeated responses because the regression coefficients vary across individuals. Here, we model  $y_{ij}$  given the individual-specific coefficients,  $\boldsymbol{\beta}_i$ , by

$$h(E(y_{ij}|\boldsymbol{\beta}_i)) = \mathbf{x}_{ij}^T \boldsymbol{\beta}_i. \quad (1.1)$$

Here,  $h(\cdot)$  is a link function. For normal responses, it is the expectation and for binary responses, it may be the log odds ratio. Usually, there are too little data

on a single person to estimate  $\beta_i$  from  $(y_i, X_i)$  alone. We further assume that the  $\beta_i$ 's are independent realizations from some distribution with mean  $\beta$ . We can write  $\beta_i = \beta + \mathbf{b}_i$ , where  $\beta$  is fixed and  $\mathbf{b}_i$  is a vector of zero-mean random variables. Then the basic assumption can be restated in terms of the latent variables  $\mathbf{b}_i$ . That is, there are unobserved factors represented by  $\mathbf{b}_i$  that are common to all responses for a given individual but which vary across individuals. Random effects models are particularly useful when inferences are to be made about individual trajectories, such as in AIDS studies. They focus on both population parameters  $\beta$  and individuals characteristics  $\mathbf{b}_i$ 's.

The third approach is called a *transition model* approach. This focuses on the conditional distribution of  $y_{ij}$  given past outcomes,  $y_{ij-1}, \dots, y_{i1}$ . Here, the data-analyst specifies a regression model for  $h(E(y_{ij}|y_{ij-1}, \dots, y_{i1}, \mathbf{x}_{ij}))$ , as an explicit function of  $\mathbf{x}_{ij}$  and of the past responses.

In each of the three approaches, we consider both the dependence of the responses on explanatory variables and the correlation among the responses. With cross-sectional data, only the dependence of the responses on the explanatory variables needs to be specified; 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:

- (1) incorrect inferences about regression coefficients  $\beta$ , particularly, confidence intervals are too short based on assumption of independence, when in fact there is positive dependence;
- (2) the estimating method of  $\beta$  may be inefficient, that is, less precise than possible;

## 1.2 Missing Data Problems

### 1.2.1 Missing Covariates and Responses

In many applications, especially in longitudinal studies, missing data are a serious problem. Ignoring missing data or using over-simplified methods to handle missing data often leads to invalid inferences. Thus, it is very important to find appropriate approaches to deal with missing data. Two kinds of missing data in longitudinal studies are common: (i) missing covariates; and (ii) missing responses due to dropout or missing visits. For example, individuals may not come to study center for measurements at scheduled time points for various reasons, or they may even dropout permanently because of drug intolerance or death. Missing data make statistical analysis in longitudinal studies much more complicated, because standard methods, which are usually designed for complete-data, are not directly applicable.

Commonly-used naive methods for missing data include the complete-data method, which only uses the complete observations and deletes all incomplete observations, the mean imputation method, which replaces the missing values by the mean values of observed data, and the last-value-carried-forward method, which imputes a missing value by the immediate previous observed data.

### 1.2.2 Classification of Missing Value Mechanisms

At the presence of missing data, the missing data mechanism must be taken into account in order to obtain valid statistical inferences. Little and Rubin (1987) and Little (1995) give a general treatment of statistical analysis with missing values. Let  $\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(o)} \\ \mathbf{y}^{(m)} \end{pmatrix}$  with  $\mathbf{y}^{(o)}$  denoting the measurements actually obtained and  $\mathbf{y}^{(m)}$  de-

noting the measurements which would have been available had they not been missing. Let  $\mathbf{r}$  denote a set of indicator random variables, denoting which elements of  $\mathbf{y}$  fall into  $\mathbf{y}^{(o)}$  and which into  $\mathbf{y}^{(m)}$ . Now, a probability model for the missing value mechanism defines the probability distribution of  $\mathbf{r}$ . Little and Rubin (1987) classify the missing value mechanism as

- 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 quite plausible. For example, some values are missing because of reasons irrelevant to the treatment such as the medical equipment is broken down on a certain day. So missingness is MCAR if  $\mathbf{r}$  is independent of both  $\mathbf{y}^{(o)}$  and  $\mathbf{y}^{(m)}$
- 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 notations, missingness is MAR if  $\mathbf{r}$  is independent of  $\mathbf{y}^{(m)}$ .
- Missing data are *nonignorable* or *informative* (NIM) if the probability of missingness depends on unobserved data. To be specific, NIM has two cases in the context of random effects models: (i) 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  $\mathbf{r}$  is dependent on  $\mathbf{y}^{(m)}$ . (ii) The probability of missingness depends on unknown random effects (i.e. individual characteristics such as individual decay rates) which may substantially affect the responses. We call missingness *random-effect-based*



*informative* (Little, 1995) if  $\mathbf{r}$  is dependent on random effect  $\mathbf{b}_i$ .

It turns out that, for likelihood-based inference, the crucial distinction is between random and informative missing values. Both MCAR and MAR missing mechanisms are sometimes referred to without distinction as *ignorable*. Little and Rubin (1987) show that, when missing data are non-ignorable, likelihood inference must incorporate the missing data mechanism.

### 1.2.3 Literature on Missing Data Problems

Little (1992) reviewed methods of estimation in regression models with missing covariates. Six methods dealing with missing covariates are compared: complete-case methods, available-case methods, least squares on imputed data, maximum likelihood methods, Bayesian methods and multiple imputation. He suggested that the maximum likelihood method, Bayesian methods, and multiple imputation method perform well, and the maximum likelihood method is preferred in a large samples and Bayesian methods or multiple imputation method are preferred in a small samples. Ibrahim (1990) considered missing covariates (MAR) in generalized linear models (GLMs) with discrete covariates, and applied the EM algorithm to obtain MLEs under the assumption that the missing covariates are from a discrete distribution. Ibrahim, Lipsitz, and Chen (1999) proposed a Monte Carlo EM algorithm for GLMs with nonignorable missing covariates.

Wu and Carroll (1988) consider linear mixed effects models (LMEs) with informative dropout under the assumption that the informative dropout could be modeled by a probit model which includes the random effects as its covariates. Diggle and Kenward (1994) consider general approaches to informative dropouts in multivariate data and longitudinal data. They show that considering informative dropout mechanisms

in the statistical inference reduces the bias caused by considering the informative dropout as only MAR. Ten Have et al. (1998) discuss mixed effects logistics regression models for longitudinal binary responses with informative dropout. Roy and Lin (2002) consider multivariate longitudinal data with nonignorable dropouts and missing covariates. Little (1995) gives an excellent review on modeling the dropout mechanism in repeated-measures studies. Dropout models were classified into selection models and pattern-mixture models. The main difference between the two type of dropout models is that the form of missing data mechanism needs to be specified in the selection models but not in the pattern-mixture models. He classified NIM into nonignorable *outcome-based* missing data where the dropout depends on the missing values, and *random-effect-based* missing data where the dropout depends on random effects. He also suggested to examine the sensitivity of the results to the choice of missing data mechanisms when we almost know nothing about the missing data mechanism. Ibrahim, Chen and Lipsitz (2001) develop a Monte Carlo EM algorithm to obtain MLEs in GLMMs with informative dropouts. They propose that the missing data mechanism may be modelled by a logistic regression and a sequence of one-dimensional conditional distributions which may reduce the number of nuisance parameters.

### 1.3 Motivating Examples

Our research is motivated by studies of HIV viral dynamics, which have received great attention in AIDS studies in recent years (Ho et al. 1995, Perelson et al. 1996; Wu and Ding 1999). These viral dynamic models provide good understanding of the pathogenesis of HIV infection and evaluation of antiretroviral therapies, and the

dynamic parameters may reflect the efficacy of the antiviral treatments (Ding and Wu, 2001). A common problem in these studies is that some subjects may drop out of the study or miss visits due to drug intolerance and other problems (although dropout patients may return to study later), and covariates may contain missing data as well. It is important to evaluate how the dropout patients affect estimates of the viral decay rates since the decay rates may reflect the efficacy of the antiviral treatments. The dataset which motivate our research consists of 48 HIV infected patients who were treated by a potent antiviral regimen. The viral load is repeatedly measured after initiation of the treatment. After the antiviral treatment, the patient's viral loads will often decay, and the decay rate may reflect the efficacy of the treatment. We only consider the viral load data before viral rebound and the first three months data since data after three months are likely to be contaminated by long-term clinical factors. The number of measurements for each patient varies from 2 to 7. Fourteen patients have missing viral loads at scheduled time points due to dropout or other problems. The baseline covariates CD4 cell counts, total complement levels (CH50), and tumor necrosis factor (TNF) contain 3.7%, 12.3% and 16.4% missing data respectively. Four patients are randomly selected and their viral loads are plotted in Figure 1.2.

Visual inspection of the raw data seems to indicate that dropout patients appear to have slower viral decay rates, compared to the remaining patients. Thus, the dropouts are likely to be informative or nonignorable. This dataset was analyzed previously, but dropout patients were discarded and the missing viral loads were assume to be missing completely at random (Wu and Ding 1999; Wu and Wu 2001). Wu (2004) re-analyze the dataset, proposing a missing mechanism based on the unobserved responses (viral loads). In this thesis, our objectives are to model the viral load, incorporating non-ignorable missing mechanism, based on unknown

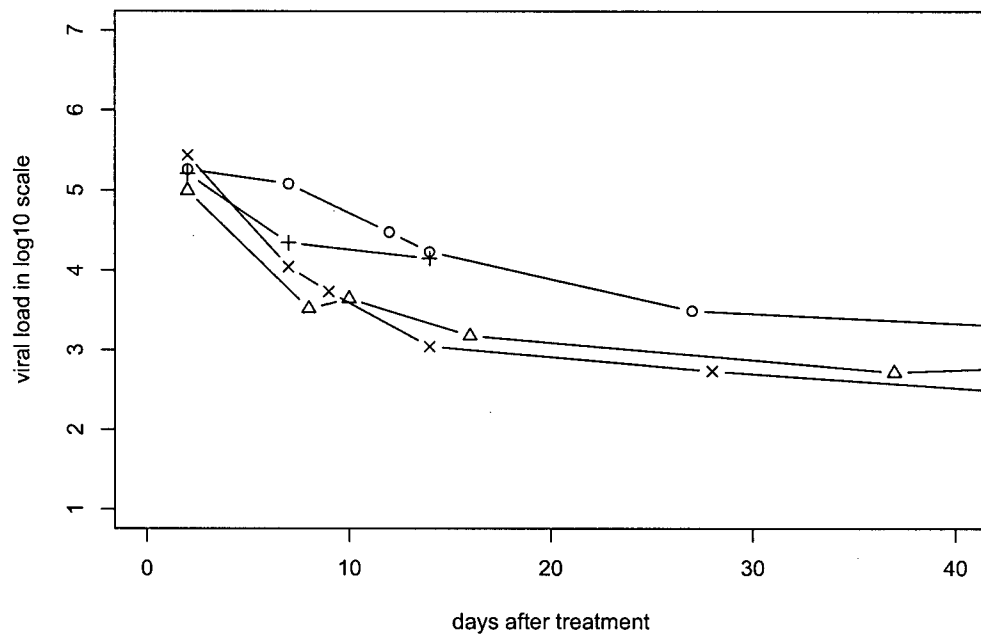


Figure 1.2: Viral loads of four randomly selected patients.

random effects, and check if the estimates of decay rates are different.

## 1.4 Objectives and Outline

In this thesis, we develop an exact inference method, implemented by a Monte Carlo EM algorithm, to make simultaneous inferences for NLMEs with informative dropout and missing covariates. To avoid computational difficulties when the dimension of random effects is not small, we also propose an approximate inference method, which integrates out the random effects in the EM algorithm for more efficient computation. Our methods differ from Wu (2004) in that the proposed dropout mechanism depends on the random effects rather than the unobserved responses.

The remainder of this thesis is organized as follows. Chapter 2 introduces NLMEs. Chapter 3 discusses the exact inference method for estimation of NLMEs with informative dropout and missing covariates. The approximate inference method based on linearization is presented in Chapter 4. We discuss dropout models and covariate models in Chapter 5. In Chapter 6, we apply our methods to real datasets. Chapter 7 presents our simulation study. We conclude the thesis with a discussion in Chapter 8.

## Chapter 2

# Nonlinear Mixed Effects Models

### 2.1 Introduction

Before we present our methods for estimating parameters in NLMEs with informative dropout and missing covariates, we give a brief introduction to NLMEs in this chapter. In Section 2.2, we introduce NLMEs for longitudinal data. Section 2.3 gives a literature review on NLMEs with informative dropout and missing covariates.

### 2.2 Nonlinear Mixed Effects Models

Linear models, such as polynomials, are often empirical models based on the observed data. Therefore, they may be only valid within the observed range of the data. There is often no theoretical consideration about the underlying mechanism, which generates the data. In many longitudinal studies such as HIV viral dynamics, pharmacokinetics analyses, and studies of growth and decay, nonlinear modeling is often required for meaningful analyses. Nonlinear mixed effects models (NLMEs), or hierarchical nonlinear models, are popular in these studies in characterizing both the

intra-individual variation and the inter-individual variation (Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1996).

As a generalization of linear models, nonlinear models have many advantages:

(1) Nonlinear models are often mechanistic, that is, they are often based on the mechanism which produces the data, so the model parameters generally have a natural physical interpretation.

(2) A nonlinear model generally uses fewer parameters than a competing linear model, such as a polynomial, offering a more parsimonious description of the data.

(3) Nonlinear models often provide more reliable prediction for the responses outside the observed data range.

However, compared with linear models, nonlinear models usually don't have a close form expression for the marginal likelihood, and thus parameter estimation is more computationally intensive.

For longitudinal data analysis, nonlinear mixed effects models are popular for inferences. Suppose that there are  $N$  individuals, with individual  $i$  having  $n_i$  measurements at times  $t_{i1}, \dots, t_{in_i}$ . Let  $y_{ij}$  be the response value for individual  $i$  at time  $t_{ij}$ , subject to informative dropout,  $i = 1, \dots, N; j = 1, \dots, n_i$ . Denote  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$ . Let  $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^T$  be a collection of incompletely observed baseline time-independent covariates for individual  $i$ . Let  $\mathbf{v}_i = (v_{i1}, \dots, v_{iq})^T$  be a collection of completely observed baseline time-independent covariates for individual  $i$ .

A general NLME model can be written as a hierarchical two-stage model as follows (Davidian and Giltinan, 1995)

$$y_{ij} = g(\mathbf{z}_{ij}, \mathbf{v}_{ij}; \boldsymbol{\beta}_i) + e_{ij}, \quad \mathbf{e}_i | \boldsymbol{\beta}_i \sim N(\mathbf{0}, \sigma^2 I) \quad (2.1)$$

$$\boldsymbol{\beta}_i = \mathbf{d}(\mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \mathbf{b}_i), \quad \mathbf{b}_i \stackrel{i.i.d}{\sim} N(\mathbf{0}, D), \quad j = 1, \dots, n_i, i = 1, \dots, N, \quad (2.2)$$

where  $g(\cdot)$  is an arbitrary nonlinear function,  $\mathbf{z}_{ij}$  and  $\mathbf{v}_{ij}$  are respectively  $(p \times 1)$  and  $(q \times 1)$  vectors of covariates,  $\mathbf{e}_i = (e_{i1}, \dots, e_{in_i})^T$  represents measurement errors,  $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{is})^T$  is a  $(s \times 1)$  vector of individual-specific regression parameters,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)^T$  is a  $(r \times 1)$  vector of population parameters (fixed effects),  $\mathbf{d}(\cdot)$  is a  $s$ -dimensional vector-valued function,  $\mathbf{b}_i = (b_{i1}, \dots, b_{is})^T$  is the vector of random effects and is independent of  $\mathbf{e}_i$ ,  $\sigma^2$  is the unknown within individual variance,  $I$  is the identity matrix, and the  $(s \times s)$  matrix  $D$  quantifies the random inter-individual covariance. We write  $D = D(\boldsymbol{\eta})$ , where  $\boldsymbol{\eta}$  denotes the collection of all distinct parameters in  $D$ .

Let  $f(\cdot)^1$  denote a generic density function and  $f(y|x)$  denote the conditional density function of  $y$  given  $x$ . After integrating out the unobserved random effects vector, the density of the responses  $\mathbf{y}_i$  is given by

$$f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2, D) = \int f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2, \mathbf{b}, D) f(\mathbf{b} | D) d\mathbf{b}, \quad (2.3)$$

and the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2, D | \mathbf{y}) = \prod_{i=1}^N \int f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2, \mathbf{b}, D) f(\mathbf{b} | D) d\mathbf{b}, \quad (2.4)$$

which generally does not have a closed-form expression. Exact likelihood calculations therefore require numerical evaluation of an integral whose dimension is equal to the number of random effects  $\mathbf{b}_i$ . This is straightforward to do by direct numerical integration when the dimension of  $\mathbf{b}_i$  is 1 or 2. However, when  $\mathbf{b}_i$  has a dimension of 3 or more, people need to consider alternative methods, such as Monte Carlo method.

---

<sup>1</sup>Here, for simplicity, we are abusing mathematical notation, by using  $f$  for many different densities, and the function can be determined from the arguments.



Lindstrom and Bates (1990) propose an approximate method based on first-order Taylor expansions about the random effects  $b_i$ . The resulting algorithm provides a computationally fast, albeit approximate, method for a wide class of non-linear models.

## 2.3 Literature Review on NLME Models with Informative Missing Data

Wu and Wu (2001) estimate parameters in nonlinear mixed effects models with missing covariates (MAR) by a three-step multiple imputation method. In first step, they fitted a hierarchical model without covariates. Then they imputed the missing covariates based on a multivariate linear model, implemented by Gibbs sampler, and created  $B$  independent complete datasets in the second step. In the last step, they used the standard complete-data method to analyze each dataset and combine  $B$  obtain the overall inference. Wu (2002) proposed a method for NLMEs with censored responses and covariates measured with errors. Wu and Wu (2002) also proposed a method for analyzing NLMEs with missing time-dependent covariates. Later, Wu (2004, a) proposed an exact and an approximate method for analyzing data with missing covariates in nonlinear mixed effects models. The exact method is implemented by a Monte Carlo EM algorithm, and the approximate method linearizes the nonlinear model based on a Taylor expansion, and it substantially reduces the computation load.

Wu (2004, b) proposed a Monte Carlo EM method for estimating parameters in NLMEs with nonignorable missing covariates and dropout, with a dropout mechanism depending on unobserved responses. However, no one has considered pa-

parameter estimation in NLMEs with informative dropout and missing covariates, with a dropout mechanism depending on unknown *random effects*. In the following chapters, we focus on NLMEs with informative dropout and missing covariates, with a *random-effects based* dropout mechanism. Since the random effects are shared by both the response model and the dropout model, this approach may also be referred to as a *shared parameter model*.

## Chapter 3

# An Exact Method for NLME Models with Informative Dropout and Missing Covariates

### 3.1 Introduction

In this chapter, we develop an exact inference method based on Monte Carlo methods to obtain MLEs for parameters in NLMEs with informative dropout and missing covariates. The proposed exact method is implemented by a Monte Carlo EM algorithm. In Section 3.2, we give a description of NLMEs with informative dropout and missing covariates. Section 3.3 describes a Monte Carlo EM algorithm. A detailed description of our sampling methods is provided in Section 3.4. Computation issues regarding our algorithm are discussed in Section 3.5.

## 3.2 The Models

We consider the models (2.1) and (2.2). Let  $\mathbf{r}_i = (r_{i1}, \dots, r_{in_i})^T$  be a vector of missing data indicators for individual  $i$  such that  $r_{ij} = 1$  if  $y_{ij}$  is missing and 0 otherwise. We write  $\mathbf{y}_i = (\mathbf{y}_{mis,i}^T, \mathbf{y}_{obs,i}^T)^T$ , where  $\mathbf{y}_{mis,i}$  corresponds to the missing components of  $\mathbf{y}_i$  and  $\mathbf{y}_{obs,i}$  contains the observed components of  $\mathbf{y}_i$ . We write  $\mathbf{z}_i = (\mathbf{z}_{mis,i}^T, \mathbf{z}_{obs,i}^T)^T$ , where  $\mathbf{z}_{mis,i}$  corresponds to the missing components of covariate vector  $\mathbf{z}_i$  and  $\mathbf{z}_{obs,i}$  contains the observed components of  $\mathbf{z}_i$ . We assume that the missing covariates are ignorable (or missing at random), i.e., the missing covariate mechanism may depend on the observed data but not on the covariate values being missing. The observed data are  $\{(\mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i), i = 1, \dots, N\}$ . Note that the dimensions of  $\mathbf{y}_{obs,i}$  and  $\mathbf{z}_{obs,i}$  depend on  $i$ .

To facilitate likelihood inference, we need to make a distributional assumption for the incompletely observed covariates  $\mathbf{z}_i$ , conditional on the completely observed covariates  $\mathbf{v}_i$ . We denote the covariate distribution by  $f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha})$ , where the parameters  $\boldsymbol{\alpha}$  may be viewed as nuisance parameters. To allow for informative missing responses, we assume a distribution for  $\mathbf{r}_i$  as  $f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\phi})$ , where the missingness may depend on the values being missing (i.e.,  $\mathbf{y}_{mis,i}$  and  $\mathbf{z}_{mis,i}$ ). The parameter  $\boldsymbol{\phi}$  are treated as nuisance parameters.

When the responses  $\mathbf{y}_i$  and the covariates  $\mathbf{z}_i$  contain missing data, likelihood inference becomes more complicated for (2.4). In this section, we consider a Monte Carlo EM algorithm (Wei and Tanner, 1990) for ‘exact’ likelihood inference, incorporating missing responses and missing covariates simultaneously. By treating the unobservable random effects  $\mathbf{b}_i$  as missing data, the ‘complete data’ become  $\{(\mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i), i = 1, \dots, N\}$ , and the ‘complete-data’ log-likelihood can be writ-

ten as

$$l(\psi) = \sum_{i=1}^N \left\{ \log f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi) + \log f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \beta, \sigma^2) \right. \\ \left. + \log f(\mathbf{b}_i | D(\eta)) + \log f(\mathbf{z}_i | \mathbf{v}_i; \alpha) \right\}, \quad (3.1)$$

where  $\psi = (\alpha, \beta, \sigma^2, \eta, \phi)$  denotes the collection of all parameters. We assume that the parameters  $\alpha, \beta, \sigma^2, \eta$ , and  $\phi$  do not overlap.

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

- *outcome-based informative* if  $f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi) = f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i; \phi)$ . That is, in addition to the covariates, the missing probability of the current response depends on the possibly unobserved response  $y_{ij}$ . For example, a patient does not show up because he is too sick to go to the clinic.
- *random-effect-based informative* if  $f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi) = f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi)$ . That is, in addition to the covariates, the missing probability of the current response depends on the underlying unobservable random effect  $\mathbf{b}_i$ . For example, a patient is more likely to miss the scheduled exam because the treatment is less effective on him and therefore, he does not believe the treatment. Or, a patient may be more likely to dropout early because his true (but unobservable) viral decay rate is slower than other patients.

Diggle and Kenward (1994), Little (1995), and Ibrahim et al. (2001) discussed various specifications of the outcome-based informative missing mechanism  $f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \phi)$ . In this thesis, we focus on the *random-effect-based* informative missing mechanism  $f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i, \phi)$ . We may assume, for example,  $r_{i1}, \dots, r_{in_i}$  are independent

with the same distribution:

$$\text{logit}[P(r_{ij} = 1|\mathbf{b}_i, \boldsymbol{\phi})] = \phi_0 + \phi_1 b_1 + \cdots + \phi_s b_s. \quad (3.2)$$

In this dropout model, the missing probabilities of responses only depend on the random effects of that patient. More complicated dropout models can be specified in a similar way. Note that the assumed models are not testable based on the observed data, so it is important to perform sensitivity analyses on various missing mechanisms. If the main parameter estimates  $\hat{\boldsymbol{\beta}}$  are quite independent of the assumed dropout model, we may be confident about the results. Otherwise, if the estimates are very sensitive to the assumed dropout model, we need to justify the dropout model first to get reasonable estimates of the parameters. The covariate model  $f(\mathbf{z}_i|\mathbf{v}_i)$  can be chosen in a similar way and sensitivity analyses should also be performed (Ibrahim et al. 1999).

### 3.3 A Monte Carlo EM 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, but it fails to get an estimating covariance matrix of MLEs. Each iteration of a EM algorithm consists of an E-step that evaluates the expectation of ‘complete data’ log-likelihood conditional on the observed data and previous parameter estimates, and an M-step that updates the parameter estimates by maximizing the the conditional expectation of log-likelihood. This iterative computation between the E-step and M-step till convergence leads to the MLEs.

If we treat  $(\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i) \equiv (\mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i)$  as the ‘com-

plete' data, the complete data density for individual  $i$  is given by

$$\begin{aligned} & f(\mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\eta}, \boldsymbol{\phi}) \\ &= f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\phi}) f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2) f(\mathbf{b}_i | D(\boldsymbol{\eta})) f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}). \end{aligned} \quad (3.3)$$

This leads to the complete data log-likelihood

$$\begin{aligned} l(\boldsymbol{\psi}) &= \sum_{i=1}^N l_i(\boldsymbol{\psi} | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i) \\ &= \sum_{i=1}^N \left\{ \log f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\phi}) + \log f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2) \right. \\ &\quad \left. + \log f(\mathbf{b}_i | D(\boldsymbol{\eta})) + \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}) \right\}, \end{aligned} \quad (3.4)$$

where  $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\eta}, \boldsymbol{\phi})$  and  $l_i(\boldsymbol{\psi} | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i)$  is the contribution to the complete data log-likelihood from the  $i$ th individual. Note that we mainly interested in estimating  $(\boldsymbol{\beta}, \sigma^2)$ , and treat  $(\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\eta})$  as nuisance parameters.

Ibrahim et al. (1999, 2001) proposed a Monte Carlo EM algorithm for estimating parameters in GLMMs with informative dropout without missing covariates and for GLMs with missing covariates respectively. Wu (2004, a, b) extends their methods to NLME models and provide a unified approach to address dropouts and missing covariates simultaneously, under outcome-based informative missing response mechanism. Here, we extend the EM methods to NLME models with dropouts and missing covariates, under *random effects based* informative missing mechanism.

### 3.3.1 E-step

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

$$\begin{aligned}
Q_i(\psi|\psi^{(t)}) &= E[l_i(\psi|y_i, z_i, v_i, r_i, b_i)|y_{obs,i}, z_{obs,i}, v_i, r_i; \psi^{(t)}] \\
&= \iiint \{\log f(r_i|y_i, b_i, z_i, v_i; \phi) + \log f(y_i|b_i, z_i, v_i; \beta, \sigma^2) \\
&\quad + \log f(z_i|v_i; \alpha) + \log f(b_i|D(\eta))\} \\
&\quad \times f(y_{mis,i}, z_{mis,i}, b_i|y_{obs,i}, z_{obs,i}, v_i, r_i, \psi^{(t)}) db_i dy_{mis,i} dz_{mis,i}. \quad (3.5)
\end{aligned}$$

In general, the above integration is intractable and does not have a closed form expression. However, since the integral is an expectation with respect to  $f(y_{mis,i}, z_{mis,i}, b_i|y_{obs,i}, z_{obs,i}, v_i, r_i; \psi^{(t)})$ , it may be evaluated using the Monte Carlo EM of Wei and Tanner (1990) and Ibrahim et al. (2001). Specifically, in the  $t$ th iteration, we may generate  $m_t$  samples from  $f(y_{mis,i}, z_{mis,i}, b_i|y_{obs,i}, z_{obs,i}, v_i, r_i; \psi^{(t)})$  and then approximate the expectation  $Q_i(\psi|\psi^{(t)})$  by its empirical mean, with missing data replaced by simulated values

$$\begin{aligned}
Q_i(\psi|\psi^{(t)}) &\approx \frac{1}{m_t} \sum_{j=1}^{m_t} \log f(r_i|y_{obs,i}, y_{mis,i}^{(j)}, b_i^{(j)}, z_{obs,i}, z_{mis,i}^{(j)}, v_i; \phi) \\
&\quad + \frac{1}{m_t} \sum_{j=1}^{m_t} \log f(y_{obs,i}, y_{mis,i}^{(j)}|b_i^{(j)}, z_{obs,i}, z_{mis,i}^{(j)}, v_i; \beta, \sigma^2) \\
&\quad + \frac{1}{m_t} \sum_{j=1}^{m_t} \log f(z_{obs,i}, z_{mis,i}^{(j)}|v_i; \alpha) \\
&\quad + \frac{1}{m_t} \sum_{j=1}^{m_t} \log f(b_i^{(j)}|D(\eta)). \quad (3.6)
\end{aligned}$$

where  $\{(y_{mis,i}^{(j)}, z_{mis,i}^{(j)}, b_i^{(j)}), j = 1, \dots, m_t\}$  are the  $m_t$  simulated values of missing responses, missing covariates, and unobservable random effects. We may choose  $m_0$  as a large number and  $m_t = m_{t-1} + m_{t-1}/k$ , ( $k \geq 1$ ) in the  $t$ th iteration. Increasing



$m_t$  with each iteration may speed up the EM convergence (Booth and Hobert, 1999). The E-step for all individuals at the  $(t + 1)$ st iteration can be written as

$$\begin{aligned}
Q(\psi|\psi^{(t)}) &= \sum_{i=1}^N Q_i(\psi|\psi^{(t)}) \\
&\approx \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{r}_i | \mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)}, \mathbf{b}_i^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{v}_i; \phi) \\
&\quad + \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)} | \mathbf{b}_i^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{v}_i; \beta, \sigma^2) \\
&\quad + \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)} | \mathbf{v}_i; \alpha) \\
&\quad + \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{b}_i^{(j)} | D(\eta)) \} \\
&\equiv Q^{(1)}(\phi|\psi^{(t)}) + Q^{(2)}(\beta, \sigma^2|\psi^{(t)}) + Q^{(3)}(\alpha|\psi^{(t)}) \\
&\quad + Q^{(4)}(D(\eta)|\psi^{(t)})
\end{aligned} \tag{3.7}$$

To generate independent samples from  $f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)})$ , we may use Gibbs sampler (Gelfand and Smith, 1990) by sampling from the following three full conditionals iteratively:

- $f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)})$ ,
- $f(\mathbf{z}_{mis,i} | \mathbf{z}_{obs,i}, \mathbf{y}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)})$ ,
- and  $f(\mathbf{b}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)})$ .

The details about this sampling will be discussed in Section 3.4.1.

In practice, when we generate samples with respect to  $f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i,$

$\mathbf{r}_i; \boldsymbol{\psi}^{(t)}$ ), we can note that

$$\begin{aligned}
& f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) / f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \cdot \frac{f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}{f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}. \tag{3.8}
\end{aligned}$$

Since we are focusing on random effects based missing mechanism, assuming that the probability of missingness can be explained by  $\mathbf{b}_i$  rather than  $\mathbf{y}_i$ , both the numerator and denominator of the second term in (3.8) are constant with respect to the new-generated samples of  $\mathbf{y}_{mis,i}$ . Thus,

$$f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \propto f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \tag{3.9}$$

Similarly, we can simplify the other two full conditionals as:

$$\begin{aligned}
& f(\mathbf{z}_{mis,i} | \mathbf{z}_{obs,i}, \mathbf{y}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{z}_i, \mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i, | \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) / f(\mathbf{z}_{obs,i}, \mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i, | \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= \frac{f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{b}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{y}_i | \mathbf{r}_i, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}{f(\mathbf{z}_{obs,i}, \mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i | \mathbf{v}_i; \boldsymbol{\psi}^{(t)})} \\
&\propto f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{y}_i | \mathbf{r}_i, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f_z^*(\mathbf{z}_i) \tag{3.10}
\end{aligned}$$

$$\begin{aligned}
& f(\mathbf{b}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{b}_i, \mathbf{y}_i, \mathbf{r}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) / f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= \frac{f(\mathbf{b}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{r}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}{f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})} \\
&\propto f(\mathbf{b}_i | \boldsymbol{\psi}^{(t)}) f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{r}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{b}_i | \boldsymbol{\psi}^{(t)}) f_b^*(\mathbf{b}_i), \tag{3.11}
\end{aligned}$$

where

$$f_z^*(\mathbf{z}_i) = f(\mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f(\mathbf{y}_i | \mathbf{r}_i, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}), \tag{3.12}$$

and

$$f_b^*(\mathbf{b}_i) = f(\mathbf{r}_i|\mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{r}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}). \quad (3.13)$$

### 3.3.2 M-step

We can update the estimates  $\boldsymbol{\psi}^{(t+1)}$  at the  $(t+1)$ st iteration by maximizing  $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)})$ .

Suppose that the parameters  $\boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}$ ,  $D(\boldsymbol{\eta})$ ,  $\boldsymbol{\phi}$  and  $\sigma$  are all different, we can update  $\boldsymbol{\phi}$ ,  $\boldsymbol{\beta}$ ,  $\sigma$ ,  $\boldsymbol{\alpha}$  and  $D(\boldsymbol{\eta})$  by maximizing  $Q^{(1)}$ ,  $Q^{(2)}$ ,  $Q^{(3)}$  and  $Q^{(4)}$  separately in the M-step.

The maximizer  $\boldsymbol{\phi}^{(t+1)}$  for  $Q^{(1)}$  may be computed via iteratively re-weighted least squares where the missing values are replaced by their simulated values  $\{\mathbf{y}_{mis,i}^{(j)}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{b}_i^{(j)}\}$ .

$$\begin{aligned} \boldsymbol{\phi}^{(t+1)} &= \arg \max_{\boldsymbol{\phi}} \{Q^{(1)}(\boldsymbol{\phi}|\boldsymbol{\psi}^{(t)})\} \\ &= \arg \max_{\boldsymbol{\phi}} \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{r}_i|\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)}, \mathbf{b}_i^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{v}_i; \boldsymbol{\phi}). \end{aligned} \quad (3.14)$$

And, similarly, the maximizer  $(\boldsymbol{\beta}^{(t+1)}, \sigma^{2(t+1)})$  for  $Q^{(2)}$  could be written as:

$$\begin{aligned} (\boldsymbol{\beta}, \sigma^2)^{(t+1)} &= \arg \max_{(\boldsymbol{\beta}, \sigma^2)} \{Q^{(2)}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\psi}^{(t)})\} \\ &= \arg \max_{(\boldsymbol{\beta}, \sigma^2)} \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)}|\mathbf{b}_i^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{v}_i; \boldsymbol{\beta}, \sigma^2); \end{aligned} \quad (3.15)$$

the maximizer  $\boldsymbol{\alpha}^{(t+1)}$  for  $Q^{(3)}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(t)})$  can be written as:

$$\begin{aligned} \boldsymbol{\alpha}^{(t+1)} &= \arg \max_{\boldsymbol{\alpha}} \{Q^{(3)}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(t)})\} \\ &= \arg \max_{\boldsymbol{\alpha}} \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}|\mathbf{v}_i; \boldsymbol{\alpha}); \end{aligned} \quad (3.16)$$

and the maximizer  $D(\boldsymbol{\eta})^{(t+1)}$  for  $Q^{(4)}(D(\boldsymbol{\eta})|\boldsymbol{\psi}^{(t)})$  can be written as:

$$\begin{aligned} D^{(t+1)} &= \arg \max_D \{Q^{(4)}(D|\boldsymbol{\psi}^{(t)})\} \\ &= \arg \max_D \frac{1}{m_t} \sum_{i=1}^N \sum_{j=1}^{m_t} \log f(\mathbf{b}_i^{(j)}|D(\boldsymbol{\eta})). \end{aligned} \quad (3.17)$$

Generally, (3.14) – (3.17) are nonlinear functions and have no closed-form expressions. The maximizers could be obtained via standard optimization procedures for complete data, such as the Newton-Raphson method and Quasi-Newton method. The asymptotic variance-covariance matrix of  $\hat{\boldsymbol{\psi}}$  can be obtained using well-known complete-data formulas (Louis, 1982; McLachlan and Krishnan, 1997), where the expectations in those formulas can be approximated by Monte-Carlo means. Specifically, note that the observed information matrix equals the expected complete information minus the missing information, that is,

$$I_{obs}(\hat{\boldsymbol{\psi}}) = I_{com}(\hat{\boldsymbol{\psi}}) - I_{mis}(\hat{\boldsymbol{\psi}}). \quad (3.18)$$

Let

$$S_{ij}(\boldsymbol{\psi}) = \frac{\partial l_i(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}, \quad (3.19)$$

$$\dot{Q}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) = \sum_{i=1}^N \dot{Q}_i(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) = \sum_{i=1}^N \sum_{j=1}^{m_t} \frac{1}{m_t} S_{ij}(\boldsymbol{\psi}), \quad (3.20)$$

and

$$\ddot{Q}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) = \frac{\partial^2 Q(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} = \sum_{i=1}^N \sum_{j=1}^{m_t} \frac{1}{m_t} \frac{\partial S_{ij}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}. \quad (3.21)$$

Since  $\boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\phi}$  and  $\boldsymbol{\eta}$  are all different, matrices  $\dot{Q}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}})$ ,  $\ddot{Q}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}})$  and  $S_{ij}(\hat{\boldsymbol{\psi}})$  are block diagonal. Then based on (3.18), the asymptotic observed information matrix is

$$I_{obs}(\hat{\boldsymbol{\psi}}) = -\ddot{Q}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) - \left\{ \sum_{i=1}^N \sum_{j=1}^{m_t} \frac{1}{m_t} S_{ij}(\hat{\boldsymbol{\psi}}) S_{ij}^T(\hat{\boldsymbol{\psi}}) - \sum_{i=1}^N \dot{Q}_i(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) \dot{Q}_i^T(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}) \right\}. \quad (3.22)$$

Thus, the asymptotic variance-covariance matrix of  $\hat{\psi}$  can be approximated by

$$\text{Cov}(\hat{\psi}) = I_{obs}^{-1}(\hat{\psi}). \quad (3.23)$$

## 3.4 Sampling Methods

### 3.4.1 Gibbs Sampler

From the previous sections, we can see that generating samples from the conditional distribution  $f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{V}_i, \mathbf{r}_i, \psi^{(t)})$  is an important step for implementing the E-step of the Monte Carlo EM algorithm. Gibbs sampler (Gelfand and Smith, 1990) is a popular method to generate samples from a complicated multi-dimensional distribution by sampling from each of the full conditional distributions in turn, if the distribution has a convenient representation via conditional distributions. Here, we use the Gibbs sampler to simulate the missing values as follows. Set initial values  $(\mathbf{y}_{mis,i}^{(0)}, \mathbf{z}_{mis,i}^{(0)}, \mathbf{b}_i^{(0)})$ . Supposed that the current generated values are  $(\mathbf{y}_{mis,i}^{(k)}, \mathbf{z}_{mis,i}^{(k)}, \mathbf{b}_i^{(k)})$ , we can obtain  $(\mathbf{y}_{mis,i}^{(k+1)}, \mathbf{z}_{mis,i}^{(k+1)}, \mathbf{b}_i^{(k+1)})$  as:

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

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

Step 2. draw a sample for the missing covariates  $\mathbf{z}_{mis,i}^{(k+1)}$  from

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

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

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

After a sufficiently large burn-in of  $d$  iterations, the sampled values will achieve a steady state. Then,  $\{(\mathbf{y}_{mis,i}^{(k)}, \mathbf{z}_{mis,i}^{(k)}, \mathbf{b}_i^{(k)}), k = d + 1, \dots, B\}$  can be treated as samples

from the multidimensional density function

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

And, if we choose a sufficiently large gap  $d'$  (usually smaller than  $d$ ), we can treat the sample series  $\{(\mathbf{y}_{mis,i}^{(k)}, \mathbf{z}_{mis,i}^{(k)}, \mathbf{b}_i^{(k)}), k = d + d', d + 2d', \dots\}$  as independent samples from the multidimensional density function.

There are several ways to get the initial values  $(\mathbf{y}_{mis,i}^{(0)}, \mathbf{z}_{mis,i}^{(0)}, \mathbf{b}_i^{(0)})$ . A simple way is to replace  $(\mathbf{y}_{mis,i}^{(0)}, \mathbf{z}_{mis,i}^{(0)})$  by the average values of the observed data, and set  $\mathbf{b}_i^{(0)}$  as 0.

### 3.4.2 Multivariate Rejection Algorithm

Sampling from the three full conditional distributions can be accomplished by an adaptive rejection algorithm (Gilk and Wild, 1992) if the appropriate densities are log-concave. However, for arbitrary NLME models, some densities may not be log-concave. In such cases, multivariate rejection sampling methods may be considered. Booth and Hobert (1999) discussed such a method for complete-data GLMM models, which can be extended to NLME models with dropouts and missing covariate as follows. Considering sampling from  $f(\mathbf{b}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$ . As in (3.13), let  $f_b^*(\mathbf{b}_i) = f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i, \boldsymbol{\psi}^{(t)})f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{r}_i, \mathbf{z}_i, \mathbf{v}_i, \boldsymbol{\psi}^{(t)})$ , and  $\tau = \sup_{\mathbf{u}} f_b^*(\mathbf{u})$ . We assume  $\tau < \infty$ . A random sample from  $f(\mathbf{b}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$  can be obtained as follows.

Step 1. sample  $\mathbf{b}_i^*$  from  $f(\mathbf{b}_i | \boldsymbol{\psi}^{(t)})$ , and independently, sample  $w$  from the uniform (0,1) distribution.

Step 2. if  $w \leq f^*(\mathbf{b}_i^*)/\tau$ , then accept  $\mathbf{b}_i^*$ ; otherwise, go to Step 1.

Samples from  $f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$  and  $f(\mathbf{z}_{mis,i} | \mathbf{z}_{obs,i}, \mathbf{y}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$  can be obtained in a similar way. Note that, when the dropout probability only de-

depends on random effect  $\mathbf{b}_i$  but not  $\mathbf{y}_i$ , we have

$$\begin{aligned}
& f(\mathbf{y}_{mis,i} | \mathbf{y}_{obs,i}, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) / f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \\
&= f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \cdot \frac{f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}{f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})} \\
&= f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) \cdot \frac{f(\mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})}{f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)})} \\
&= f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i; \boldsymbol{\psi}^{(t)}) f_y^*(\mathbf{y}_i). \tag{3.24}
\end{aligned}$$

The function  $f_y^*(\mathbf{y}_i) = f(\mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \boldsymbol{\psi}^{(t)}) / f(\mathbf{y}_{obs,i}, \mathbf{r}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \boldsymbol{\psi}^{(t)})$  is a constant with respect to the missing responses  $\mathbf{y}_{mis,i}$ . So, in Step 2, we always have  $f_y^*(\mathbf{y}_i^*) \equiv \tau$  for any generated  $\mathbf{y}_i^*$ . Thus  $f_y^*(\mathbf{y}_i^*) / \tau \equiv 1 \geq w$ , for any  $w$  generated from uniform(0,1) and  $\mathbf{y}_i^*$  is always accepted. That is, for any  $\mathbf{y}_i^*$  generated from  $f(\mathbf{y}_i | \mathbf{z}_i, \mathbf{b}_i, \mathbf{v}_i, \boldsymbol{\psi}^{(t)})$  in Step 1, we always accept  $\mathbf{y}_i^*$  in Step 2.

### 3.4.3 Importance Sampling

When the dimension of  $\mathbf{y}_{mis,i}$ ,  $\mathbf{z}_{mis,i}$  or  $\mathbf{b}_i$  is large, the above 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 normal density or a multivariate Student  $t$  density whose mean and variance match the mode and curvature of  $f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$ . Booth and Hobert (1999) discuss an importance sampling method for complete-data GLMM models. Here, we may extend their method to NLME models with dropout and missing covariates. Specially, we may write

$$f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)}) = c \exp(h(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)), \tag{3.25}$$

where  $c$  is a unknown normalizing constant. Let  $\dot{h}(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)$  and  $\ddot{h}(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)$  be the first and second derivatives of  $h(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)$  respectively, and let  $(\mathbf{y}_{mis,i}^*, \mathbf{z}_{mis,i}^*, \mathbf{b}_i^*)$  be the solution of  $\dot{h}(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i) = 0$ , which is the maximizer of  $h(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)$ . Then, the Laplace approximations of the mean and variance of  $f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$  are  $(\mathbf{y}_{mis,i}^*, \mathbf{z}_{mis,i}^*, \mathbf{b}_i^*)$  and  $-(\ddot{h}(\mathbf{y}_{mis,i}^*, \mathbf{z}_{mis,i}^*, \mathbf{b}_i^*))^{-1}$  respectively. Suppose that  $\{(\mathbf{y}_{mis,i}^{*(1)}, \mathbf{z}_{mis,i}^{*(1)}, \mathbf{b}_i^{*(1)}), \dots, (\mathbf{y}_{mis,i}^{*(m_t)}, \mathbf{z}_{mis,i}^{*(m_t)}, \mathbf{b}_i^{*(m_t)})\}$  is a random sample of size  $m_t$  generated from the importance function  $h^*(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i)$ , that is assumed to have the same support as  $f(\mathbf{y}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})$ . Then we have

$$Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(t)}) \approx \sum_{i=1}^N \left\{ \frac{1}{m_t} \sum_{j=1}^{m_t} w_{ij}^{(t)} l(\boldsymbol{\psi}, \mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{*(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{*(j)}, \mathbf{b}_i^{*(j)}) \right\}, \quad (3.26)$$

where

$$w_{ij}^{(t)} = \frac{f(\mathbf{y}_{mis,i}^{*(j)}, \mathbf{z}_{mis,i}^{*(j)}, \mathbf{b}_i^{*(j)} | \mathbf{y}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i, \boldsymbol{\psi}^{(t)})}{h^*(\mathbf{y}_{mis,i}^{*(j)}, \mathbf{z}_{mis,i}^{*(j)}, \mathbf{b}_i^{*(j)})} \quad (3.27)$$

are importance weights.

For the above sampling methods, the rejection sampling methods may be more efficient when the dimension of the random effects and the sample size are small, while the importance sampling method may be more efficient when the sample sizes are large since in this case the importance distribution may closely resembles the true conditional distribution.

### 3.5 Convergence

When applying the Monte Carlo EM algorithm, Monte Carlo samples for the “missing data” are drawn at each iteration to approximate the true values. 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  $m_t$ , the Monte Carlo sample size, will result in more precise but slower computation. A common strategy is to increase  $m_t$  as the number of EM iterations increase (Booth and Hobert, 1999). For sufficiently large values of  $m_t$ , the Monte Carlo EM algorithm would inherit the properties of the exact versions, such as the likelihood increasing properties of EM, but would substantially increase the computation work load. Thus, we usually use a relatively small  $m_t$  at initial iterations, and then increase  $m_t$  with the iteration.

If the Monte Carlo error associated with  $\hat{\psi}^{(t+1)}$  is large, the  $(t+1)$ th iteration of the Monte Carlo EM is wasted because the EM step has been swamped by the Monte Carlo error. Booth and Hobert (1999) proposed an automated method for choosing  $m_t$  in the context of complete-data GLMM models. Here we extend their method to NLME models with dropout and missing covariates.

Let

$$Q^{(1)}(\psi|\hat{\psi}^{(t)}) = \frac{\partial Q(\psi|\hat{\psi}^{(t)})}{\partial \psi}, \quad (3.28)$$

$$Q^{(2)}(\psi|\hat{\psi}^{(t)}) = \frac{\partial^2 Q(\psi|\hat{\psi}^{(t)})}{\partial \psi \partial \psi^T}, \quad (3.29)$$

and let  $\psi^{*(t+1)}$  be the solution of  $Q^{(1)}(\psi|\hat{\psi}^{(t)}) = 0$ . When the simulated samples are independent, it can be seen that the conditional distribution of  $f(\psi^{(t+1)}|\psi^{(t)})$  is approximately normal with mean  $\psi^{*(t+1)}$  and a variance that can be estimated by

$$\begin{aligned} & \widehat{\text{Var}}(\psi^{(t+1)}|\psi^{(t)}) \\ &= Q^{(2)}(\psi^{*(t+1)}|\hat{\psi}^{(t)})^{-1} \widehat{\text{Var}}(Q_i^{(1)}(\psi^{*(t+1)}|\hat{\psi}^{(t)})) Q^{(2)}(\psi^{*(t+1)}|\hat{\psi}^{(t)})^{-1} \end{aligned} \quad (3.30)$$

where

$$\begin{aligned}
& \widehat{\text{Var}}(Q_i^{(1)}(\psi^{*(t+1)}|\hat{\psi}^{(t)})) \\
&= \frac{1}{m_t} \sum_{j=1}^{m_t} \left\{ \left[ w_{ij} \frac{\partial}{\partial \psi} \log f(\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{b}_i^{(j)}; \psi^{(t+1)}) \right] \right. \\
&\quad \left. \times \left[ w_{ij} \frac{\partial}{\partial \psi} \log f(\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{(j)}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}^{(j)}, \mathbf{b}_i^{(j)}; \psi^{(t+1)}) \right]^T \right\}, \quad (3.31)
\end{aligned}$$

$\mathbf{y}_{mis,i}^{(j)}$ ,  $\mathbf{z}_{mis,i}^{(j)}$  and  $\mathbf{b}_i^{(j)}$  are simulated samples, and  $w_{ij}$  are the importance weights when the importance sampling is used and are all set equal to 1 when rejection sampling methods are used. After the  $(t+1)$ th iterations, we may then construct an approximate  $100(1-\alpha)\%$  confidence ellipsoid for  $\psi^{*(t+1)}$  based on the above normal approximation. The EM step is swamped by Monte Carlo error if the previous  $\psi^{(t)}$  lies in the confidence ellipsoid, and in that case we need to increase  $m_t$ . For example, we may set  $m_t$  to be  $m_{t-1} + m_{t-1}/k$  for some positive constant  $k$  and appropriate  $m_0$ .

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

## Chapter 4

# An Approximate Method for NLME Models with Informative Dropout and Missing Covariates

### 4.1 Introduction

In the previous chapter, we have described a Monte Carlo EM algorithm for “exact” likelihood inference for NLME models with informative dropout and missing covariates. However, the exact method may offer potential computational problems such as slow or non-convergence, especially when the dimension of the random effects  $\mathbf{b}_i$  is large or the intra-individual data are not rich. When the dimension of the random effects  $\mathbf{b}_i$  is not small, sampling the random effects may lead to inefficient and computationally unstable Gibbs samplers, and may lead to a high degree of autocorrelation and lack of convergence. When the intra-individual data are sparse, the individual nonlinear regressions used in the M-step may fit poorly, leading to slow or non-convergence. To reduce computation work load, in this section, we propose an *ap-*

*proximate* method which iteratively solves certain LME models and avoids sampling the random effects in E-step. The proposed approximate method may be preferable when the exact method exhibits computational difficulties. Alternatively, the approximate estimates can be excellent starting values for the exact method. Note that this approximate method is *exact* for LME models and certain NLME models where the model may be nonlinear in the fixed effects but is strictly linear in the random effects.

For complete-data NLME models, approximate methods have been widely used, and these approximate methods often perform reasonably well in most cases (Lindstorm 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 models is that of Lindstorm and Bates (1990), which is equivalent to iteratively solving certain LME models (Wolfinger, 1993). For LME models with missing responses but completely observed covariates, Ibrahim et al. (2001) propose an efficient EM method which is obtained by integrating out the random effects. These methods can be extended to NLME models with informative dropout and missing covariates for approximate inference. Here, we propose an approximate method based on Lindstorm and Bates (1990), which uses first-order Taylor expansions around the updated parameter and random effects estimates and is equivalent to iteratively solving certain LME models. Then we propose to handle the missing responses and missing covariates in the LME model step, for which the random effects can be integrated out. Thus, this approximate method avoids sampling the random effects in the E-step and avoids fitting some nonlinear models in the M-step, and therefore avoids potential computational difficulties associated with the exact method. Moreover, well known efficient EM-type algorithms for complete-data LME models (Meng and van

Dyk, 1997; Liu et al. 1998) can be incorporated in the LME model step to further speed up the convergence.

## 4.2 The Approximate Method

We can rewrite the NLME model (2.1) and (2.2) as a single equation by combining the two stages:

$$y_{ij} = g_{ij}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i) + \varepsilon_{ij}, \quad j = 1, \dots, n_i; i = 1, \dots, N, \quad (4.1)$$

where  $g_{ij}(\cdot)$  is a nonlinear function. To simplify the notation, we suppress the complete observed covariates  $\mathbf{v}_i$  and denote the current estimate of  $\boldsymbol{\psi}$  in the  $t$ -th EM iteration by  $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, D(\boldsymbol{\eta})^{(t)}, \sigma^{2(t)}, \boldsymbol{\phi}^{(t)})$ . Let  $\mathbf{g}_i = (g_{i1}, \dots, g_{in_i})^T$ . Following Lindstorm and Bates (1990) and Wolfinger (1993), the proposed approximate method iteratively solves the following LME model:

$$\tilde{\mathbf{y}}_i = X_i \boldsymbol{\beta} + T_i \mathbf{b}_i + \mathbf{e}_i, \quad (4.2)$$

where

$$\tilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{g}_i(\mathbf{z}_i, \boldsymbol{\beta}^{(t)}, \mathbf{b}_i^{(t)}) + X_i \boldsymbol{\beta}^{(t)} + T_i \mathbf{b}_i^{(t)}, \quad (4.3)$$

$$X_{ij} = \left. \frac{\partial g_{ij}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i)}{\partial \boldsymbol{\beta}^T} \right|_{(\boldsymbol{\beta}^{(t)}, \mathbf{b}_i^{(t)})}, \quad (4.4)$$

$$T_{ij} = \left. \frac{\partial g_{ij}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i)}{\partial \mathbf{b}_i^T} \right|_{(\boldsymbol{\beta}^{(t)}, \mathbf{b}_i^{(t)})}, \quad (4.5)$$

$X_i \equiv X_i(\mathbf{z}_i) = (X_{i1}^T, \dots, X_{in_i}^T)^T$ ,  $T_i \equiv T_i(\mathbf{z}_i) = (T_{i1}^T, \dots, T_{in_i}^T)^T$ , and  $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \dots, \tilde{y}_{in_i})^T$ .

Note that we have

$$\begin{aligned} & f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\ &= f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}), \end{aligned} \quad (4.6)$$

where

$$\tilde{\mathbf{y}}_{mis,i} = \mathbf{y}_{mis,i} - \mathbf{g}_{mis,i}(\mathbf{z}_{mis,i}, \boldsymbol{\beta}^{(t)}, \mathbf{b}_i^{(t)}) + X_{mis,i}(\mathbf{z}_{mis,i})\boldsymbol{\beta}^{(t)} + T_{mis,i}(\mathbf{z}_{mis,i})\mathbf{b}_i^{(t)}, \quad (4.7)$$

and  $X_{mis,i}, T_{mis,i}$  are submatrices of  $X_i, T_i$  respectively.  $\mathbf{g}_{mis,i}$  is a sub-vector function of  $\mathbf{g}_i$ .  $\tilde{\mathbf{y}}_{obs,i}$  is defined similarly, and  $\tilde{\mathbf{y}}_i = (\tilde{\mathbf{y}}_{mis,i}^T, \tilde{\mathbf{y}}_{obs,i}^T)^T$ . Under the LME model (4.2), it is straightforward to show that

$$f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) \sim N(\tilde{\mathbf{b}}_i, \tilde{\Sigma}_i), \quad (4.8)$$

where

$$\tilde{\Sigma}_i = (\sigma^{-2(t)} T_i^T T_i + D^{-1(t)})^{-1}, \quad (4.9)$$

$$\tilde{\mathbf{b}}_i = \tilde{\Sigma}_i T_i^T (\tilde{\mathbf{y}}_i - X_i \boldsymbol{\beta}^{(t)}) / \sigma^{2(t)}. \quad (4.10)$$

#### 4.2.1 E-step

We can integrate out  $\mathbf{b}_i$  and obtain the following results.

$$\begin{aligned} Q_i(\boldsymbol{\psi} | \boldsymbol{\psi}^{(t)}) &= E[l_i(\boldsymbol{\psi} | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \mathbf{v}_i, \mathbf{r}_i, \mathbf{b}_i) | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}] \\ &= \iiint \{ \log f(\mathbf{r}_i | \tilde{\mathbf{y}}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\phi}^{(t)}) + \log f(\tilde{\mathbf{y}}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}^{(t)}, \sigma^{(t)2}) \\ &\quad + \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}^{(t)}) + \log f(\mathbf{b}_i | D(\boldsymbol{\eta})) \} \\ &\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned} \quad (4.11)$$

$$\begin{aligned}
I_1 &= \iiint \log f(\mathbf{r}_i | \tilde{\mathbf{y}}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) \\
&\quad \times d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iiint \log f(\mathbf{r}_i | \tilde{\mathbf{y}}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) \\
&\quad \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \psi^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iint \left\{ \int \log f(\mathbf{r}_i | \tilde{\mathbf{y}}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi^{(t)}) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \psi^{(t)}) d\mathbf{b}_i \right\} \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \quad (4.12)
\end{aligned}$$

Consider the dropout model (3.2), and suppose that the missing probabilities for each time points are conditionally mutually independent. Then, we have

$$f(\mathbf{r}_i | \tilde{\mathbf{y}}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi^{(t)}) = \prod_{j=1}^{n_i} P(R_{ij} = 1 | \phi^{(t)}, \mathbf{b}_i)^{r_{ij}} (1 - P(R_{ij} = 1 | \phi^{(t)}, \mathbf{b}_i))^{1-r_{ij}} \quad (4.13)$$

Note that, here, we use  $R_{ij}$  to represent the argument in the equation, and use  $r_{ij}$  to represent the value we observed. Define  $\tilde{\phi} = (\phi_1, \dots, \phi_s)^T$ , and we can re-write (4.12) as

$$\begin{aligned}
I_1 &= \iint \left\{ \int \log \left[ \prod_{j=1}^{n_i} P(R_{ij} = 1 | \phi^{(t)}, \mathbf{b}_i)^{r_{ij}} (1 - P(R_{ij} = 1 | \phi^{(t)}, \mathbf{b}_i))^{1-r_{ij}} \right] \right. \\
&\quad \left. f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \psi^{(t)}) d\mathbf{b}_i \right\} \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iint \left\{ \int \left( \sum_{j=1}^{n_i} \left[ r_{ij} \log(P(R_{ij} = 1 | \mathbf{b}_i, \phi^{(t)})) + (1 - r_{ij}) \log(1 - P(R_{ij} = 1 | \mathbf{b}_i, \phi^{(t)})) \right] \right. \right. \\
&\quad \left. \left. \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \psi^{(t)}) d\mathbf{b}_i \right\} f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iint \left\{ \sum_{j=1}^{n_i} \int \left[ r_{ij} \log \left( \frac{\phi_0 + \tilde{\phi}^T \mathbf{b}_i}{1 + \phi_0 + \tilde{\phi}^T \mathbf{b}_i} \right) + (1 - r_{ij}) \log \left( \frac{1}{1 + \phi_0 + \tilde{\phi}^T \mathbf{b}_i} \right) \right] \right. \\
&\quad \left. \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \psi^{(t)}) d\mathbf{b}_i \right\} f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&\stackrel{def}{=} \iint \left\{ \sum_{j=1}^{n_i} \Omega_1 \right\} f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \psi^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i}, \quad (4.14)
\end{aligned}$$

where

$$\begin{aligned} \Omega_1 = & \int \left[ r_{ij} \log \left( \frac{\phi_0 + \tilde{\phi}^T \mathbf{b}_i}{1 + \phi_0 + \tilde{\phi}^T \mathbf{b}_i} \right) + (1 - r_{ij}) \log \left( \frac{1}{1 + \phi_0 + \tilde{\phi}^T \mathbf{b}_i} \right) \right] \\ & \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i. \end{aligned} \quad (4.15)$$

Define  $a_i = \phi_0 + \tilde{\phi}^T \mathbf{b}_i$ . From (4.8), since  $\mathbf{b}_i$  is multivariate normally distributed, with mean

$$\boldsymbol{\mu}_{\mathbf{b}_i} = \tilde{\Sigma}_i T_i^T (\tilde{\mathbf{y}}_i - X_i \boldsymbol{\beta}^{(t)}) / \sigma^{2(t)} \quad (4.16)$$

and variance

$$\Sigma_{\mathbf{b}_i} = (\sigma^{-2(t)} T_i^T T_i + D^{-1(t)})^{-1}, \quad (4.17)$$

$a_i$  is a univariate normal random variable with mean

$$\mu_{a_i} = \phi_0 + \tilde{\phi}^T \tilde{\Sigma}_i T_i^T (\tilde{\mathbf{y}}_i - X_i \boldsymbol{\beta}^{(t)}) / \sigma^{2(t)} \quad (4.18)$$

and variance

$$\Sigma_{a_i} = \tilde{\phi}^T (\sigma^{-2(t)} T_i^T T_i + D^{-1(t)})^{-1} \tilde{\phi}. \quad (4.19)$$

Take an  $\mathbb{R}^s \mapsto \mathbb{R}^s$  variable transformation

$$a_i = \phi_0 + \tilde{\phi}^T \mathbf{b}_i, \quad b_{2i} = b_{2i}, \quad \dots, \quad b_{si} = b_{si}, \quad (4.20)$$

and assume that at least one  $\phi_j \neq 0, j = 1, 2, \dots, s$ , say  $\phi_1 \neq 0$ , so the Jacobi determinant

$$\left| \frac{\partial(a_i, b_{2i}, \dots, b_{si})}{\partial(b_{1i}, b_{2i}, \dots, b_{si})} \right| = \phi_1 \neq 0, \quad (4.21)$$



and we can write  $\Omega_1$  as:

$$\begin{aligned}
\Omega_1 &= \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \\
&= \int \cdots \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] \\
&\quad \times f(a_i, b_{2i}, \dots, b_{si} | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) \left| \frac{\partial(a_i, b_{2i}, \dots, b_{si})}{\partial(b_{1i}, b_{2i}, \dots, b_{si})} \right| db_{1i} db_{2i} \cdots db_{si} \\
&= \int \cdots \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] \\
&\quad \times f(a_i, b_{2i}, \dots, b_{si} | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) \phi_1 db_{1i} db_{2i} \cdots db_{si} \\
&= \int \cdots \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] \\
&\quad \times f(a_i, b_{2i}, \dots, b_{si} | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d(\phi_1 b_{1i}) db_{2i} \cdots db_{si} \\
&= \int \cdots \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] \\
&\quad \times f(a_i, b_{2i}, \dots, b_{si} | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i db_{2i} \cdots db_{si} \\
&= \int \cdots \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] \\
&\quad \times f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) f(b_{2i}, \dots, b_{si} | a_i, \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i db_{2i} \cdots db_{si} \\
&= \int \left[ \int \cdots \int f(b_{2i}, \dots, b_{si} | a_i, \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) db_{2i} \cdots db_{si} \right] \\
&\quad \times \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&= \int \left[ r_{ij} \log \left( \frac{a_i}{1+a_i} \right) + (1-r_{ij}) \log \left( \frac{1}{1+a_i} \right) \right] f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i, \quad (4.22)
\end{aligned}$$

where  $f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)})$  is the density function of  $N(\mu_{a_i}, \Sigma_{a_i})$ . By Taylor expansion of

$a_i$  at  $\mu_{a_i}$ , we can write  $\Omega_1$  approximately as:

$$\begin{aligned}
\Omega_1 &\approx \int \left\{ r_{ij} \left[ \log \left( \frac{\mu_{a_i}}{1 + \mu_{a_i}} \right) + \left( \frac{1}{\mu_{a_i}} - \frac{1}{\mu_{a_i} + 1} \right) (a_i - \mu_{a_i}) \right] \right. \\
&\quad \left. + (1 - r_{ij}) \left[ \log \left( \frac{1}{1 + \mu_{a_i}} \right) + \frac{1}{\mu_{a_i} + 1} (a_i - \mu_{a_i}) \right] \right\} f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&= \int \left[ r_{ij} \log \mu_{a_i} - \log(1 + \mu_{a_i}) \right] f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&\quad + \int \left[ \frac{r_{ij}}{\mu_{a_i}} - \frac{1}{\mu_{a_i} + 1} \right] (a_i - \mu_{a_i}) f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&= \left[ r_{ij} \log \mu_{a_i} - \log(1 + \mu_{a_i}) \right] \int f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&\quad + \left[ \frac{r_{ij}}{\mu_{a_i}} - \frac{1}{\mu_{a_i} + 1} \right] \int (a_i - \mu_{a_i}) f(a_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) da_i \\
&= \left[ r_{ij} \log \mu_{a_i} - \log(1 + \mu_{a_i}) \right] \times 1 + \left[ \frac{r_{ij}}{\mu_{a_i}} - \frac{1}{\mu_{a_i} + 1} \right] \times 0 \\
&= r_{ij} \log \mu_{a_i} - \log(1 + \mu_{a_i}). \tag{4.23}
\end{aligned}$$

Therefore,

$$\begin{aligned}
I_1 &\approx \iint \left\{ \sum_{j=1}^{n_i} r_{ij} \log \mu_{a_i} - \log(1 + \mu_{a_i}) \right\} \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iint \left\{ n_{i,mis} \log \mu_{\tilde{\mathbf{a}}_i} - n_i \log(1 + \mu_{a_i}) \right\} \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \tag{4.24}
\end{aligned}$$

where  $n_{i,mis}$  is the total number of missing responses from the  $i$ th subject, i. e.,

$$n_{i,mis} = \sum_{j=1}^{n_i} r_{ij}. \tag{4.25}$$

Next,

$$\begin{aligned}
I_2 &= \iiint \log f(\tilde{\mathbf{y}}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}^{(t)}, \sigma^{(t)2}) \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iiint \log f(\tilde{\mathbf{y}}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}^{(t)}, \sigma^{(t)2}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\
&\quad \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&= \iint \left\{ \int \log f(\tilde{\mathbf{y}}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}^{(t)}, \sigma^{(t)2}) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \right\} \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\
&\stackrel{def}{=} \iint \Omega_2 f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \tag{4.26}
\end{aligned}$$

$$\begin{aligned}
\Omega_2 &= \int \log f(\tilde{\mathbf{y}}_i | \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\beta}^{(t)}, \sigma^{(t)2}) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \\
&= \int \log \left( \frac{1}{(\sqrt{2\pi})^{n_i} |\sigma^2 I|^{1/2}} \exp \left( -\frac{1}{2\sigma^2} (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \mathbf{b}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \mathbf{b}_i) \right) \right) \\
&\quad \times \frac{1}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{b}_i - \tilde{\mathbf{b}}_i)^T \tilde{\Sigma}^{-1} (\mathbf{b}_i - \tilde{\mathbf{b}}_i) \right) d\mathbf{b}_i \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) - \frac{1}{2\sigma^2} \int \frac{(\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \mathbf{b}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \mathbf{b}_i)}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i|^{1/2}} \\
&\quad \times \exp \left( -\frac{1}{2} (\mathbf{b}_i - \tilde{\mathbf{b}}_i)^T \tilde{\Sigma}_i^{-1} (\mathbf{b}_i - \tilde{\mathbf{b}}_i) \right) d\mathbf{b}_i \tag{4.27}
\end{aligned}$$

where  $s$  is the dimension of random effects  $\mathbf{b}_i$ , as defined in Section 3.2, and  $C_2 = -\frac{n_i+s}{2} \log(2\pi)$ .

We can perform a variable transformation as  $\mathbf{b}_i = \tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i$ , and obtain:

$$\begin{aligned}
\Omega_2 &= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) \\
&\quad - \frac{1}{2\sigma^2} \int \frac{(\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i^{1/2}|} \\
&\quad \times \exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right) d(\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) \\
&\quad - \frac{1}{2\sigma^2} \int \frac{(\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i^{1/2}|} \\
&\quad \times \exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right) |\tilde{\Sigma}_i^{1/2}| d(\mathbf{k}_i) \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) \\
&\quad - \frac{1}{2\sigma^2} \int \frac{(\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i - T_i \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)}{(\sqrt{2\pi})^s} \\
&\quad \times \exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right) d\mathbf{k}_i \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) - \frac{1}{2\sigma^2} (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i) \\
&\quad - \frac{1}{2\sigma^2} \int (\mathbf{k}_i^T \tilde{\Sigma}_i^{\frac{1}{2}} T_i^T T_i \tilde{\Sigma}_i^{\frac{1}{2}} \mathbf{k}_i) \frac{\exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right)}{(\sqrt{2\pi})^s} d\mathbf{k}_i \\
&\quad - \frac{1}{2\sigma^2} \int \left(\mathbf{k}_i^T \tilde{\Sigma}_i^{\frac{1}{2}} T_i^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)\right) \frac{\exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right)}{(\sqrt{2\pi})^s} d\mathbf{k}_i \\
&\quad - \frac{1}{2\sigma^2} \int \left((\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)^T T_i \tilde{\Sigma}_i^{\frac{1}{2}} \mathbf{k}_i\right) \frac{\exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right)}{(\sqrt{2\pi})^s} d\mathbf{k}_i \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) - \frac{1}{2\sigma^2} (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i) \\
&\quad - \frac{1}{2\sigma^2} \int (\mathbf{k}_i^T \tilde{\Sigma}_i^{\frac{1}{2}} T_i^T T_i \tilde{\Sigma}_i^{\frac{1}{2}} \mathbf{k}_i) \frac{\exp\left(-\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i\right)}{(\sqrt{2\pi})^s} d\mathbf{k}_i + 0 + 0 \\
&= C_2 - \frac{n_i}{2} \log \sigma^2 - \frac{1}{2} \log(|\tilde{\Sigma}_i|) - \frac{1}{2\sigma^2} (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)^T (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i) \\
&\quad - \frac{1}{2\sigma^2} \text{Tr}(T_i^T T_i \tilde{\Sigma}_i). \tag{4.28}
\end{aligned}$$

Thus,

$$I_2 \approx C_2 - \frac{n_i}{2} \log \sigma^2 - \iint \left[ \frac{1}{2\sigma^2} \text{Tr}(T_i^T T_i \tilde{\Sigma}_i) + \frac{1}{2\sigma^2} (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i)^T \right. \\ \left. \times (\mathbf{y}_i - X_i \boldsymbol{\beta} - T_i \tilde{\mathbf{b}}_i) \right] f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i}.$$

Also,

$$I_3 = \iiint \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ = \iiint \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\ \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ = \iint \left\{ \int \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}^{(t)}) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \right\} \\ \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ = \iint \log f(\mathbf{z}_i | \mathbf{v}_i; \boldsymbol{\alpha}^{(t)}) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i}, \quad (4.29)$$

and,

$$I_4 = \iiint \log f(\mathbf{b}_i | D) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i}, \mathbf{b}_i | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ = \iiint \log f(\mathbf{b}_i | D) f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) \\ \times f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ = \iint \left\{ \int \log f(\mathbf{b}_i | D) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \right\} \\ \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i} \\ \stackrel{\text{def}}{=} \iint \Omega_4 f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i}, \quad (4.30)$$

where,

$$\Omega_4 = \int \log f(\mathbf{b}_i | D) f(\mathbf{b}_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i; \boldsymbol{\psi}^{(t)}) d\mathbf{b}_i \\ = C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \int (\mathbf{b}_i^T D^{-1} \mathbf{b}_i) \\ \times \frac{1}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{b}_i - \tilde{\mathbf{b}}_i)^T \tilde{\Sigma}^{-1} (\mathbf{b}_i - \tilde{\mathbf{b}}_i) \right) d\mathbf{b}_i, \quad (4.31)$$

and  $C_4 = -\frac{s}{2} \log(2\pi)$ .

Again, we may perform a variable transformation  $\mathbf{b}_i = \tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i$  and obtain:

$$\begin{aligned}
\Omega_4 &= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \int (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \\
&\quad \frac{1}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i|^{1/2}} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d(\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \\
&= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \int (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \\
&\quad \frac{1}{(\sqrt{2\pi})^s |\tilde{\Sigma}_i|^{1/2}} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) |\tilde{\Sigma}_i^{1/2}| d(\mathbf{k}_i) \\
&= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \int (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} (\tilde{\mathbf{b}}_i + \tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \\
&\quad \frac{1}{(\sqrt{2\pi})^s} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d\mathbf{k}_i \\
&= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \tilde{\mathbf{b}}_i^T D^{-1} \tilde{\mathbf{b}}_i \\
&\quad - \frac{1}{2} \int (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \frac{1}{(\sqrt{2\pi})^s} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d\mathbf{k}_i \\
&\quad - \frac{1}{2} \int \tilde{\mathbf{b}}_i^T D^{-1} (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \frac{1}{(\sqrt{2\pi})^s} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d\mathbf{k}_i \\
&\quad - \frac{1}{2} \int (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} \tilde{\mathbf{b}}_i \frac{1}{(\sqrt{2\pi})^s} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d\mathbf{k}_i \\
&= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \tilde{\mathbf{b}}_i^T D^{-1} \tilde{\mathbf{b}}_i \\
&\quad - \frac{1}{2} \int (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i)^T D^{-1} (\tilde{\Sigma}_i^{1/2} \mathbf{k}_i) \frac{1}{(\sqrt{2\pi})^s} \exp \left( -\frac{1}{2} \mathbf{k}_i^T \mathbf{k}_i \right) d\mathbf{k}_i - 0 - 0 \\
&= C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \tilde{\mathbf{b}}_i^T D^{-1} \tilde{\mathbf{b}}_i - \frac{1}{2} \text{Tr}(D^{-1} \tilde{\Sigma}_i). \tag{4.32}
\end{aligned}$$

Thus,

$$\begin{aligned}
I_4 &\approx C_4 - \frac{1}{2} \log |D| - \frac{1}{2} \int \left( \frac{1}{2} \tilde{\mathbf{b}}_i^T D^{-1} \tilde{\mathbf{b}}_i + \frac{1}{2} \text{Tr}(D^{-1} \tilde{\Sigma}_i) \right) \\
&\quad \times f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{v}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)}) d\tilde{\mathbf{y}}_{mis,i} d\mathbf{z}_{mis,i}. \tag{4.33}
\end{aligned}$$

From the analytical discussion above, we have integrated out  $\mathbf{b}_i$ . Consequently,  $Q_i(\boldsymbol{\psi} | \boldsymbol{\psi}^{(t)})$  can be evaluated as an integral with respect to the density function

$f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)})$ . Thus, in E-step, we can only sample  $\mathbf{y}_{mis,i}$  and  $\mathbf{z}_{mis,i}$ . This may substantially reduce computation burden since the integration with respect to  $\mathbf{b}_i$  is not needed.

### 4.2.2 M-step

The M-step in each iteration of the approximate method is similar to the M-steps of the exact method, which is discussed in Section 3.3.2. The asymptotic variance-covariance matrix of  $\hat{\boldsymbol{\psi}}$  can again be obtained using well-known complete-data formulas as described in Section 3.3.2. The only difference is that the likelihood function  $l_i$  is for the model (4.2).

## 4.3 Monte Carlo Sampling

Having integrated out the random effects  $\mathbf{b}_i$ , in E-step we only need to simulate samples from  $f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)})$ . As in Chapter 3, we can again use the Gibbs sampler to draw the desired samples. The procedure is described as follows. Set initial values  $(\tilde{\mathbf{y}}_{mis,i}^{(0)}, \mathbf{z}_{mis,i}^{(0)})$ . Supposed that the current generated values are  $(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{z}_{mis,i}^{(k)})$ .

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

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

Step 2. draw a sample for the missing covariates  $\mathbf{z}_{mis,i}^{(k+1)}$  from

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

After a burn-in period, the sampled values  $(\tilde{\mathbf{y}}_{mis,i}^{(k+1)}, \mathbf{z}_{mis,i}^{(k+1)})$  can be treated as the true sample from the density function  $f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i} \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)})$ . And, if we

choose a sufficiently large gap  $d'$  (usually smaller than  $d$ ), we can treat the sample series  $\{(\tilde{\mathbf{y}}_{mis,i}^{(k)}, \mathbf{z}_{mis,i}^{(k)}), k = d + d', d + 2d', \dots\}$  as independent samples from the multidimensional density function  $f(\tilde{\mathbf{y}}_{mis,i}, \mathbf{z}_{mis,i} | \tilde{\mathbf{y}}_{obs,i}, \mathbf{z}_{obs,i}, \mathbf{V}_i, \mathbf{r}_i; \boldsymbol{\psi}^{(t)})$ .



## Chapter 5

# Covariates Models and Dropout Models

### 5.1 Introduction

In the foregoing chapters, we have already discussed the methodology for estimation of parameters in NLMEs with informative dropout and missing covariates. To provide valid inference, we need to specify a dropout model for the missing response, and a covariate model for the incompletely observed covariates, and then incorporate them into our analyses. However, the dropout model is usually unknown and hard to be verified from the observed data. Sensitivity analyses are thus very important in that they can show us how sensitive our conclusion rely on our models. If our estimates vary a lot when we choose different dropout models or covariate models, they may be unreliable because we do not know whether our covariate model and dropout model are true. On the other hand, if our estimates are robust to model selection, we can believe that the estimates may be reliable. In Section 5.2 we introduce covariate models. In Section 5.3 we discuss possible dropout models. In Section 5.4 we discuss

sensitivity analyses for the dropout model and covariate model.

## 5.2 Covariate Models

When some covariates are missing, we need to assume a distribution for the covariates. The parameters in the covariate model are also viewed as nuisance parameters. Ibrahim (1990) proposed a saturated multinomial model for categorical covariates with missing values. A drawback of his method is that the saturated model greatly increases the number of nuisance parameters, which increases computation burden and may make the model unidentifiable. When the missing covariates are all continuous, we may assume a multivariate normal distribution for the covariates (Little and Schlucher, 1985). To allow both continuous and categorical covariates, we may write the covariate distribution as a product of one-dimensional conditional distributions, as in Ibrahim, et al. (1999)

$$\begin{aligned}
 f(\mathbf{z}_i; \boldsymbol{\alpha}) &= f(z_{i,p} | z_{i,1}, \dots, z_{i,p-1}; \boldsymbol{\alpha}_p) \\
 &\quad \times f(z_{i,p-1} | z_{i,1}, \dots, z_{i,p-2}; \boldsymbol{\alpha}_{p-1}) \\
 &\quad \times \dots \times f(z_{i,1} | \boldsymbol{\alpha}_1),
 \end{aligned} \tag{5.1}$$

where  $\mathbf{z}_i$  is the covariate vector for the  $i$ th subject,  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \dots, \boldsymbol{\alpha}_p^T)^T$  is the parameter which characterize the relationships among the covariates, and  $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_p$  are all different. The index  $p$  is the number of covariates. Note that we do not need to make distributional assumption for the completely observed covariates, which are conditioned on and are suppressed in the above expressions. Note also that this modeling scheme allows the missing covariates to be continuous, categorical and mixed. For example, suppose that  $z_1$  is continuous and  $z_2$  is binary. By the above modeling

strategy, we may specify a normal distribution for  $z_1$  and a logistic regression model for  $z_2$  conditional on  $z_1$ .

For the dataset described in Section 1.3, all of the three covariates in the models, CD4 ( $z_1$ ), TNF ( $z_2$ ) and CH50 ( $z_3$ ), contain missing values. Thus we need to assume a joint covariate model for likelihood inference. As we have discussed, we model the joint distribution of  $\mathbf{z} = (z_1, z_2, z_3)^T$  as a product of three one-dimensional conditional distributions:

$$f(z_{i1}, z_{i2}, z_{i3} | \boldsymbol{\alpha}) = f(z_{i3} | z_{i1}, z_{i2}; \boldsymbol{\alpha}_3) f(z_{i2} | z_{i1}; \boldsymbol{\alpha}_2) f(z_{i1} | \boldsymbol{\alpha}_1). \quad (5.2)$$

where  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T$ .

We focus on the following saturated model,

$$(z_{i3} | z_{i1}, z_{i2}; \boldsymbol{\alpha}_3) \sim N(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2}, \alpha_{33}), \quad (5.3)$$

$$(z_{i2} | z_{i1}; \boldsymbol{\alpha}_2) \sim N(\alpha_{20} + \alpha_{21}z_{i1}, \alpha_{22}), \quad (5.4)$$

$$(z_{i1} | \boldsymbol{\alpha}_1) \sim N(\alpha_{10}, \alpha_{11}), \quad (5.5)$$

where  $\boldsymbol{\alpha}_3 = (\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33})^T$ ,  $\boldsymbol{\alpha}_2 = (\alpha_{20}, \alpha_{21}, \alpha_{22})^T$ , and  $\boldsymbol{\alpha}_1 = (\alpha_{10}, \alpha_{11})^T$ . We will also consider other more parsimonious models for sensitivity analysis.

### 5.3 Dropout Models

Dropout models are the models for the missing responses indicators  $r_{ij}$ . The parameters in the dropout model 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  $\boldsymbol{\beta}$  more efficient. Moreover, too many nuisance parameters may even make the NLME model unidentifiable. Therefore, one should be very cautious about adding extra nuisance parameters.

Since the missing responses indicators  $r_{ij}$  are binary, a simple model for them is a logistic regression model as follows. We may assume that the missing probabilities for each time points are independent, i. e.,

$$f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \psi) = \prod_{j=1}^{n_i} \pi_{ij}^{r_{ij}} (1 - \pi_{ij})^{1-r_{ij}}, \quad (5.6)$$

and

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = h(\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \psi), \quad (5.7)$$

where  $\pi_{ij} = P(R_{ij} = 1)$  is the probability that  $y_{ij}$  is missing, and  $h(\cdot)$  is often an linear function of  $\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i$  and  $\mathbf{v}_i$ . More complicated models can also be considered, but they may introduce more parameters and increase the computational burden.

In general, the probability that  $y_{ij}$  is missing may depend on many factors, such as past or current responses, individual random effects, covariates, etc. However, since in this thesis we focus on random effects based informative dropouts, we may assume that the missing probabilities of responses are only explainable through  $\mathbf{b}_i$ , the random effects, i. e.,

$$f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \psi) = f(\mathbf{r}_i|\mathbf{b}_i, \phi). \quad (5.8)$$

Thus, we have

$$f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi) = \prod_{j=1}^{n_i} P(R_{ij} = 1|\phi, \mathbf{b}_i)^{r_{ij}} (1 - P(R_{ij} = 1|\phi, \mathbf{b}_i))^{1-r_{ij}} \quad (5.9)$$

Again, note that, as in Section 3.3.2, we use  $R_{ij}$  to represent the argument in the function and  $r_{ij}$  to represent the value of  $R_{ij}$  as we have observed.

## 5.4 Sensitivity Analyses

The dropout model and covariate model are not verifiable based on the observed data, so it is important to conduct sensitivity analyses. That is, we need to try other

plausible dropout models and covariate models, and then assess the sensitivity of results to those different models. If there is not much difference between the results based on different models, we can draw relatively reliable conclusions. Otherwise, the results may depend on assumed the models and the conclusions may not be reliable.

# Chapter 6

## Data Analysis

### 6.1 Introduction

In the previous chapters, we have discussed an exact method and an approximate method for NLMEs with *random effects based informative* dropouts and missing covariates. In this chapter, we will analyze two real datasets. In Section 6.2, we analyze the data described in Section 1.3. This dataset has both missing covariates and missing responses. We will only apply the exact method. In Section 6.3, we analyze another AIDS dataset, which does not have covariates. We will apply both the exact method and approximate method on the second dataset. In Section 6.4, we discuss some computational issues.

Table 6.1: Data summary of Example 1.

Variable	Sample mean	Sample standard deviation	Percentage of missing values
Viral load	3.57	0.99	6.7%
CD4	177.3	86.9	3.7%
CH50	241.1	48.6	16.4%
TNF	61.0	28.8	12.3%
# of patients: $N=48$			
# of observations per patient: $n_i=2$ to 7.			

## 6.2 Example 1

### 6.2.1 Data Description

The dataset which motivates our research consists of 48 HIV infected patients who were treated by a potent antiviral regimen. The Plasma HIV-1 RNA (viral load) is repeatedly measured on days 2, 7, 10, 14, 28, and weeks 8, 12 and 24, after initiation of the treatment. After the antiviral treatment, the patient's viral loads will decay, and the decay rate may reflect the efficacy of the treatment. Throughout the time course, due to individual characteristics, the viral load may continue to decay, fluctuate, or even start to rise (rebound). We only consider the data for the first three months data since data after three months are likely to be contaminated by long-term clinical factors. The number of measurements for each patient varies from 2 to 7. Fourteen patients have missing viral loads at scheduled time points due to dropout or other problems. The baseline covariates CD4 cell counts, total complement levels (CH50), and tumor necrosis factor (TNF) contain 3.7% 16.4%, and 12.3% missing data respectively. We summarize our data in Table 6.1 (viral load is in  $\log_{10}$  scale).

Four patients are randomly selected and their viral loads are plotted in Figure

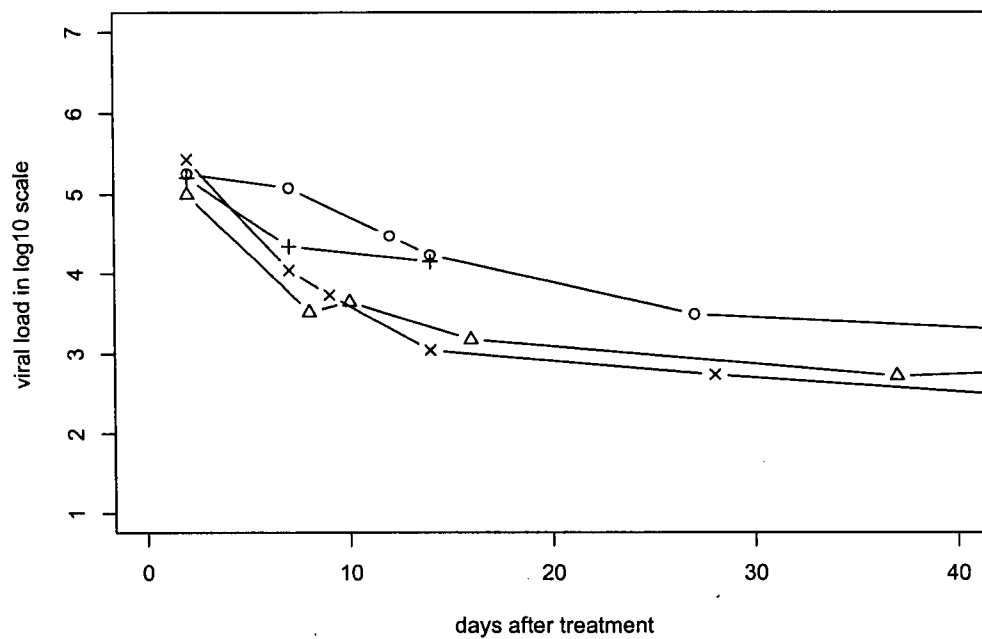


Figure 6.1: Viral loads of four randomly selected patients. (Example 1)

#### 6.1.

Visual inspection of the raw data seems to indicate that dropout patients appear to have slower viral decay, compared to the remaining patients. Thus, the dropouts are likely to be informative or nonignorable. This dataset was analyzed previously, but dropout patients were discarded and the missing viral loads were assume to be missing completely at random (Wu and Ding 1999; Wu and Wu 2001). Wu (2004) re-analyzed the dataset, assuming that the missing mechanism depends on the unobserved responses (viral loads). In this section, our objectives are to consider random effects based nonignorable missing mechanism, and check if the estimates of decay rates are different.



### 6.2.2 Models

The following two-phase HIV viral dynamic model has been proposed for this study (Wu and Ding, 1999; Wu and Wu, 2001)

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{1i}} + P_{2i}e^{-\lambda_{2i}t_{2i}}) + \varepsilon_{ij}, \quad (6.1)$$

$$\log_{10}(P_{1i}) = \beta_1 + \beta_2 TNF + b_{1i}, \quad \lambda_{1i} = \beta_3 + \beta_4 TNF + \beta_5 CD4 + b_{2i},$$

$$\log_{10}(P_{2i}) = \beta_6 + \beta_7 TNF + b_{3i}, \quad \lambda_{2i} = \beta_8 + \beta_9 CH50 + b_{4i}, \quad (6.2)$$

where  $y_{ij}$  is the  $\log_{10}$  transformation of viral load for patient  $i$  at the  $j$ th visit,  $i = 1, \dots, N; j = 1, \dots, n_i$ ,  $N = 48$  and  $n_i$  varies from 2 to 7,  $\lambda_{1i}$  and  $\lambda_{2i}$  represents two viral decay rates,  $P_{1i}$  and  $P_{2i}$  are baseline values,  $b_{ki}, k = 1, \dots, 4$ , are random effects, and  $\varepsilon_{ij}$  represents within individual errors.

In this study, the baseline CD4 ( $z_1$ ), CH50 ( $z_2$ ) and TNF( $z_3$ ) all contains missing values. To make a valid likelihood inference, we need to specify a model for these three covariates. We model the joint distribution of  $\mathbf{z} = (z_1, z_2, z_3)^T$  as a product of three one-dimensional conditional distributions:

$$f(z_{i1}, z_{i2}, z_{i3}|\boldsymbol{\alpha}) = f(z_{i3}|z_{i1}, z_{i2}; \boldsymbol{\alpha}_3) \cdot f(z_{i2}|z_{i1}; \boldsymbol{\alpha}_2) \cdot f(z_{i1}|\boldsymbol{\alpha}_1). \quad (6.3)$$

where  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T$ . We first focus on the following saturated model:

$$(z_{i3}|z_{i1}, z_{i2}; \boldsymbol{\alpha}_3) \sim N(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2}, \alpha_{33}), \quad (6.4)$$

$$(z_{i2}|z_{i1}; \boldsymbol{\alpha}_2) \sim N(\alpha_{20} + \alpha_{21}z_{i1}, \alpha_{22}), \quad (6.5)$$

$$(z_{i1}|\boldsymbol{\alpha}_1) \sim N(\alpha_{10}, \alpha_{11}). \quad (6.6)$$

Figure 6.2 shows the Q-Q plots for each of the three covariates. It appears that the normality assumption may be plausible.

The responses  $y_{ij}$  contain 6.7% missing values. Thus, we also need to assume a model for the dropout mechanism in order to make valid likelihood inference. Note

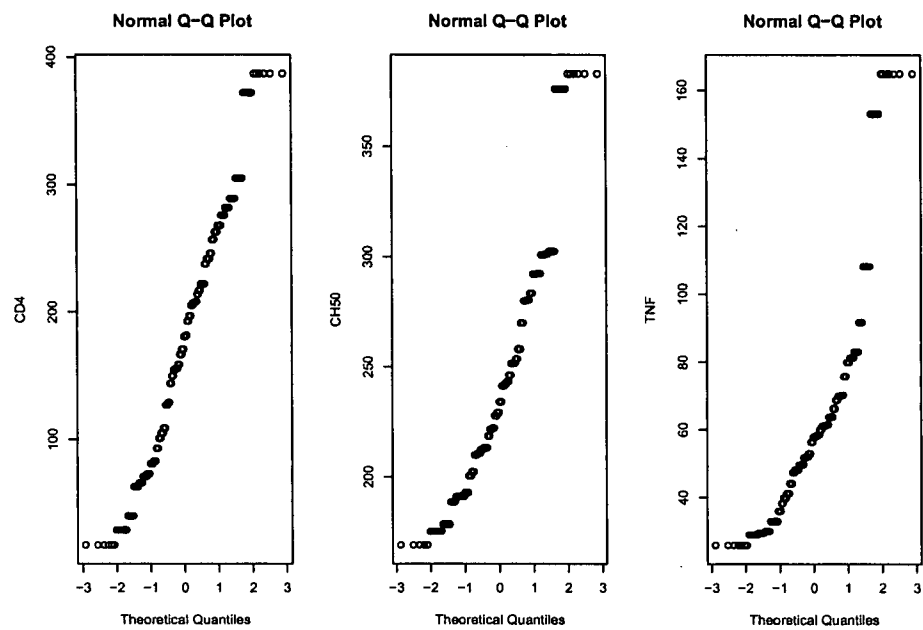


Figure 6.2: Q-Q plots for covariates (Example 1)

that although dropout models are not verifiable based on observed data, subject-area knowledge and sensitivity analyses based on plausible models may still lead to reasonable models. Since, in this thesis, we focus on random-effects based nonignorable missing data, we propose the following missing model.

$$f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \boldsymbol{\phi}) = \prod_{j=1}^{n_i} P(R_{ij} = 1 | \boldsymbol{\phi}, \mathbf{b}_i)^{r_{ij}} (1 - P(R_{ij} = 1 | \boldsymbol{\phi}, \mathbf{b}_i))^{1-r_{ij}}$$

$$\log \left( \frac{P(R_{ij} = 1 | \boldsymbol{\phi}, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \boldsymbol{\phi}, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i} + \phi_4 b_{4i}, \quad (6.7)$$

where  $\mathbf{r}_i = (r_{i1}, \dots, r_{in_i})^T$  is a vector of missing data indicators for individual  $i$  such that  $r_{ij} = 1$  if  $y_{ij}$  is missing and 0 otherwise.

### 6.2.3 Analysis and Results

We consider estimating the population parameters  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_9)^T$  using three methods: the complete case method, exact method assuming ignorable missingness, and exact method with nonignorable missingness. Details of the computation are given in Section 6.4. The results are shown in Table 6.2; Jackknife standard errors are included for comparison to check on the accuracy of method for standard error estimation described in Section 3.3.2.

We see that the results are somewhat different under different methods. For the most important parameter  $\beta_3$ , the initial decay rate, the exact method assuming non-ignorable missing gives the smallest estimate, the exact method assuming ignorable missing gives an moderate estimate, and the complete case method gives the largest. This suggests that studies assuming ignorable missing data mechanism or discarding dropout patients may over-estimate the initial viral decay rate.

Another informative parameter is  $\beta_1$ , which is the intercept term of the baseline viral load. Although it is not of much interest for testing the efficiency of the new

Table 6.2: Estimations for response model parameters. (Example 1)

	Exact (Ignor.)			Exact (Non-ignor.)			Complete Case		
	Estimate	SE	JSE	Estimate	SE	JSE	Estimate	SE	JSE
$\beta_1$	12.29	0.02	0.04	12.27	0.02	0.04	13.45	0.32	0.08
$\beta_2$	0.61	0.01	0.06	0.54	0.10	0.04	0.59	0.28	0.09
$\beta_3$	37.54	0.34	0.50	35.84	0.39	0.71	40.65	4.40	0.90
$\beta_4$	0.49	0.14	1.26	2.85	1.75	0.63	1.07	3.43	2.01
$\beta_5$	7.69	0.20	0.45	5.59	0.76	0.57	5.46	2.06	0.55
$\beta_6$	7.83	0.02	0.04	7.77	0.02	0.05	7.60	0.32	0.04
$\beta_7$	0.30	0.01	0.08	0.30	0.03	0.03	0.34	0.14	0.07
$\beta_8$	2.08	0.03	0.07	1.99	0.03	0.08	1.92	0.49	0.07
$\beta_9$	0.13	0.01	0.05	0.092	0.01	0.05	0.20	0.22	0.09
$\sigma$	0.58	0.12		0.61	0.02		0.32	0.56	

Complete Case is based on 30 subjects without any missing measurements.

treatment, it shows the difference between different estimation methods. The exact methods almost give the same estimate for  $\beta_1$ , no matter whether we assume ignorable missingness or not. However, the complete case method gives a larger estimate. A possible interpretation is that, since the complete case method overestimate the initial decay rate, the baseline response “intercept” is correspondingly larger. The JSE is often but not always larger. Based on comparing the estimated SEs from the sample sizes of 30 complete cases and 48 subjects in total, it looks like JSE is more reliable than SE using method of Section 3.3.2 and *nlme*( ). The reliability of different approaches for SE estimation is a topic of future research.

Although the parameters in the dropout models are nuisance parameters and are usually not of interest, they sometimes may contain useful information. We summarize the estimates of  $\phi = (\phi_0, \dots, \phi_4)^T$  in Table 6.3. From Table 6.3, we could see that the estimate of  $\phi_1$  is positive and the estimate of  $\phi_2$  is negative. Consider our model (6.1) and (6.2). We may conclude that patients who have higher baseline

Table 6.3: Estimations for dropout model parameters. (Example 1)

	Parameters				
	$\phi_0$	$\phi_1$	$\phi_2$	$\phi_3$	$\phi_4$
Estimate	2.64	0.14	-0.11	-0.12	0.07
Standard Error	0.0017	0.11	0.067	0.17	0.42
<i>p</i> -value	< 0.001	0.21	0.10	0.48	0.86

viral loads and slower (true, but unobservable) initial decay rate are more likely to dropout. The new treatment may be less efficient on such patients, and ignoring these patients may result in over-optimistic conclusion on the treatment effects.

Note that the *p*-value associated with  $\phi_1$  is 0.21, which may be acceptable in practice. However,  $\phi_3$  and  $\phi_4$  are not statistically significant, and we may consequently remove them to simplify the model. Note also that the Wald tests are only approximate here, so these *p*-values should only serve as rough guidance. In next section, we will perform sensitivity analyses on alternative dropout models.

#### 6.2.4 Sensitivity Analysis

It is important to check sensitivity of parameter estimates to various plausible dropout models. Subject-area knowledge may help us to determine alternative dropouts models. It is conceivable that dropout may be related to individual's random effects, current and previous viral load measurements, or covariates such as CD4 cell counts. Such relationship may be very complicated, but simple logistic regressions may provide reasonable approximation. Note that we should make use of the conclusions from preliminary studies and try to propose simple but plausible dropout models. We should avoid building a too complicated dropout model since the parameters may become non-identifiable (Fitzmaurice et al. 1996). Here we consider the following

Table 6.4: Sensitivity analyses for dropout models. (Example 1)

	Model (6.8)			Model (6.9)			Model (6.10)		
	Estimate	SE	JSE	Estimate	SE	JSE	Estimate	SE	JSE
$\beta_1$	12.27	0.02	0.04	12.27	0.02	0.04	11.50	0.02	0.03
$\beta_2$	0.54	0.10	0.04	0.63	0.08	0.04	0.55	0.06	0.05
$\beta_3$	35.84	0.39	0.71	36.57	0.50	0.70	35.90	0.29	0.57
$\beta_4$	2.85	1.75	0.63	1.87	1.12	0.60	1.77	0.07	0.68
$\beta_5$	5.59	0.76	0.57	7.25	0.80	0.57	6.63	0.17	0.55
$\beta_6$	7.77	0.02	0.05	7.66	0.02	0.04	7.09	0.01	0.04
$\beta_7$	0.30	0.03	0.03	0.32	0.03	0.03	0.28	0.01	0.04
$\beta_8$	1.99	0.03	0.08	1.40	0.02	0.05	1.88	0.02	0.07
$\beta_9$	0.09	0.01	0.05	0.07	0.02	0.05	0.09	0.01	0.06
$\sigma$	0.61	0.02		0.62	0.02		0.80	0.01	

plausible dropout models

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i} + \phi_4 b_{4i}, \quad (6.8)$$

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} \quad (6.9)$$

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 y_{ij} \quad (6.10)$$

Since we only focus on random-effects based informative missingness in this thesis, model (6.8) and (6.9) are of main interest. Model (6.10) is responses based informative missingness, and Wu (2004) discusses and analyzes a similar dataset based on such approach. Here, we consider this kind of dropout model only for comparison. The parameter estimates under different dropout models are summarized in Table 6.4. The standard error estimates are almost consistent among different dropout models. For model (6.10), the SE estimates described in Section 3.3.2 tends to yield smaller estimates.

We find that the resulting parameter estimates for viral dynamic parameters  $\beta$

are all comparable. This suggests that the estimation of the viral dynamic parameters may be robust against plausible dropout models, and thus the estimates in Table 6.2 may be reliable. Compared with models (6.8) and (6.9), model (6.10) gives smaller standard errors on the estimates. However the jackknife SEs for the three models are much closer for each beta. Hence this is an indication of potential problems with the method of Section 3.3.2. Or it may be because model (6.10) is more parsimonious than model (6.8) and (6.9) (i.e., less nuisance parameters). Another possible explanation is that model (6.10) models the missing probability at each scheduled time point, while the other two models model the missing probability for each subject. Thus, model (6.10) captured more information.

### 6.2.5 Conclusion

Based on our analyses, we conclude that, for the HIV viral load data, complete case analyses and analyses assuming ignorable missingness may over-estimate the initial decay rate. Assuming non-ignorable missingness incorporates possible mechanism which leads to patients' dropouts, and therefore may give parameter estimates which may be more reliable. Either responses-based or random-effects based non-ignorable dropout models may be used to get valid inferences, and either may be used as a tool for sensitivity analysis for the other one.

Table 6.5: Data summary of Example 2.

Variable	Sample mean	Sample standard deviation	Percentage of missing values
Viral load	4.27	1.12	6.5%
# of patients: $N=51$ , Total # of observations: 415, # of visits per patient: $n_i=4$ to $10$ ( $\text{mean}(n_i)=8.1, \text{SD}(n_i)=1.6$ )), # of missing values: 27 (from 15 patients).			

## 6.3 Example 2

### 6.3.1 Data Description

Our second example is from another HIV study. The data contain HIV viral measurements from 51 patients. The Plasma HIV-1 RNA (viral load) is repeatedly measured on days 1, 2, 3, 7, 10, 14, 28, and weeks 8, 12 and 24, after initiation of an anti-HIV treatment. After the antiviral treatment, the patient's viral loads will decay, and the decay rate may reflect the efficacy of the treatment. As in the first dataset in Example 1, throughout the time course, due to individual characteristics, the viral load may continue to decay, fluctuate, or even start to rise (rebound). We only consider the first three months data since data after three months are likely to be contaminated by long-term clinical factors. The number of measurements for each patient varies from 2 to 7. 15 patients have missing viral loads at scheduled time points due to dropout or other problems. However, different from Example 1, this study does not contain baseline covariates. Four patients are randomly selected and their viral load measurements are plotted in Figure 6.3. We summarize our data in Table 6.5.

As in Example 1, visual inspection of the raw data seems to indicate that dropout patients appear to have slower viral decay, compared to the remaining pa-



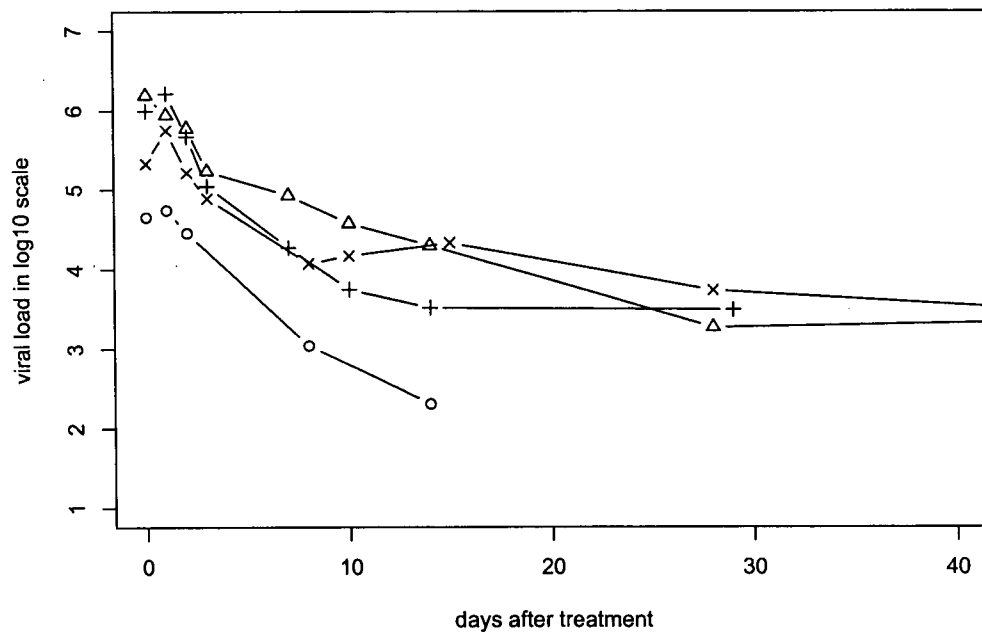


Figure 6.3: Viral loads of four randomly selected patients. (Example 2)

tients. Thus, the dropouts are likely to be informative or nonignorable.

### 6.3.2 Models

We consider the following HIV viral dynamic model, which does not contain covariates (Wu and Ding, 1999)

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{1i}} + P_{2i}e^{-\lambda_{2i}t_{2i}}) + \varepsilon_{ij}, \quad (6.11)$$

$$\begin{aligned} \log_{10}(P_{1i}) &= \beta_1 + b_{1i}, & \lambda_{1i} &= \beta_2 + b_{2i}, \\ \log_{10}(P_{2i}) &= \beta_3 + b_{3i}, & \lambda_{2i} &= \beta_4 + b_{4i}, \end{aligned} \quad (6.12)$$

where  $y_{ij}$  is the  $\log_{10}$  transformation of viral load for patient  $i$  at the  $j$ th visit,  $i = 1, \dots, N$ ;  $j = 1, \dots, n_i$ , where  $N = 51$  and  $n_i$  varies from 4 to 10,  $\lambda_{1i}$  and  $\lambda_{2i}$  represents two viral decay rates,  $P_{1i}$  and  $P_{2i}$  are baseline values,  $b_{ki}, k = 1, \dots, 4$  are random effects, and  $\varepsilon_{ij}$  represents within individual errors.

The responses  $y_{ij}$  contain 6.5% missing values. Thus, we also need to assume a model for the dropout mechanism in order to make valid likelihood inference. Note that although dropout models are not verifiable based on observed data, subject-area knowledge and sensitivity analyses based on plausible models may still lead to reasonable models. Since, in this thesis, we focus on random-effects based nonignorable missing data, we propose the following missing response model.

$$\begin{aligned} f(\mathbf{r}_i | \mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i, \mathbf{v}_i; \phi) &= \prod_{j=1}^{n_i} P(R_{ij} = 1 | \phi, \mathbf{b}_i)^{r_{ij}} (1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i))^{1-r_{ij}} \\ \log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) &= \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i} + \phi_4 b_{4i}, \end{aligned} \quad (6.13)$$

where  $\mathbf{r}_i = (r_{i1}, \dots, r_{in_i})^T$  is a vector of missing data indicators for individual  $i$  such that  $r_{ij} = 1$  if  $y_{ij}$  is missing and 0 otherwise.

Table 6.6: Estimates for dynamic model parameters in Model (6.11) and (6.12). (Example 2)

	Exact (Nonig.)			Exact (Ignor.)			Approx. (Nonig.)			Comp. Case		
	Est	SE	JSE	Est	SE	JSE	Est	SE	JSE	Est	SE	JSE
$\beta_1$	12.23	0.002	0.004	13.33	0.003	0.004	11.45	0.002	0.005	12.27	0.23	0.01
$\beta_2$	60.13	0.15	0.27	66.07	1.12	0.35	55.78	0.08	0.25	58.53	2.27	0.59
$\beta_3$	8.60	0.05	0.07	9.55	0.11	0.08	9.01	0.004	0.10	8.56	0.23	0.19
$\beta_4$	3.96	0.13	0.20	4.16	0.14	0.22	2.88	0.076	0.25	3.80	0.30	0.39
$\sigma$	0.05	0.08		0.52	0.08		0.32	0.02		0.20	0.17	

### 6.3.3 Analysis and Results

We consider estimating the population parameters  $\beta = (\beta_1, \dots, \beta_4)^T$  using four methods: the complete-case method, the exact method assuming ignorable missingness, the exact method with nonignorable missingness, and the approximate method with nonignorable missingness. The results are shown in Table 6.6.

We see that the results are somewhat different under different methods. For the most important parameter  $\beta_2$ , the initial decay rate, the exact method with nonignorable missingness gives a smaller estimate than the exact method with ignorable missingness, the approximate method gives the smallest estimate, and the complete case method gives a moderate estimate but with large estimation variance. This suggests that studies assuming ignorable missing data mechanism or discarding dropout patients may over-estimate the initial viral decay rate. Note that the approximate method gives smaller estimates and standard errors. For some parameters, the estimate of standard errors in approximate method seems too small. It may indicate that the standard error estimate in Section 4.2.2 is not reliable. For the approximate method, the Jackknife standard errors may be more reliable.

Table 6.7: Estimations for dropout model parameters. (Example 2)

	Parameters				
	$\phi_0$	$\phi_1$	$\phi_2$	$\phi_3$	$\phi_4$
Estimate	2.95	0.611	-2.046	-0.562	0.1034
Standard Error	0.0346	0.0296	0.448	0.0465	0.125
<i>p</i> -value	< 0.001	< 0.001	< 0.001	< 0.001	0.41

The estimates for parameters in the dropout model are summarized in Table 6.7. We find that, except for  $\phi_4$ , all other coefficients are statistically significant. This suggest that the missing mechanism may be non-ignorable, and the missingness may be related to the underlying unobservable individual characteristics (random effects).

#### 6.3.4 Sensitivity Analysis

From the previous section, we conclude that the dropout mechanism may be non-ignorable because most *p*-values associated with the coefficients  $\phi$  in the dropout model are statistically significant. We still need to check whether different dropout models may affect the estimates of the parameters, because the validation of dropout models are not verifiable based on the observed data. Thus, it is important to check sensitivity of parameter estimates to various plausible dropout models. Subject-area knowledge may help us to determine alternative dropouts models. We consider the following two possible models:

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i} + \phi_4 b_{4i}, \quad (6.14)$$

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i}, \quad (6.15)$$

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} \quad (6.16)$$

Table 6.8: Sensitivity analyses for dropout models. (Example 2)

	Model (6.14)			Model (6.15)			Model (6.16)		
	Estimate	S.E.	J.S.E.	Estimate	S.E.	J.S.E.	Estimate	S.E.	J.S.E.
$\beta_1$	12.23	0.002	0.004	12.23	0.002	0.004	12.25	0.002	0.04
$\beta_2$	60.13	0.15	0.27	60.13	0.14	0.27	60.09	0.14	0.25
$\beta_3$	8.60	0.05	0.07	8.59	0.06	0.07	8.66	0.06	0.09
$\beta_4$	3.96	0.13	0.20	3.99	0.13	0.19	4.02	0.11	0.25
$\sigma$	0.50	0.08		0.49	0.08		0.44	0.08	

The parameter estimates by exact method under different dropout models are summarized in Table 6.8. We find that the three model almost result in the same estimates, and therefore, we may conclude that the parameter estimates may be robust against different dropout models.

### 6.3.5 Conclusion

From the above analyses, we conclude that, for this dataset, the dropout maybe nonignorable. Simply ignoring the dropouts may over-estimate the initial viral decay rate  $\beta_2$ . The approximate method is much faster than the exact method.

## 6.4 Computation Issues

**Starting values.** For the EM algorithms in our examples, the starting values for  $\beta$  were obtained based on the complete-case methods. For example 1, the starting values for  $\alpha$  in the covariate model were obtained from linear regression models (6.4)–(6.6) using completely observed cases. The starting values for  $\phi$  were set to be  $(\phi_0, \phi_1, \phi_2, \phi_3, \phi_4) = (1, 0, 0, 0, 0)$ .

**Stopping rule.** The stopping rule for the EM algorithms in our examples is that the relative change in the parameter estimates from successive 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 a extremely small tolerance level, otherwise it may converge very slowly. The actual tolerance level we used in our examples is 0.05. The EM is stopped when each of the new parameters estimates falls within 5% difference from the corresponding estimates from the last EM iteration for two consecutive iterations.

**Running time.** For the data of Example 1, the algorithm for the exact method converged in about three hours on a SUN Sparc work-station (Ultra-60). For the data of Example 2, the algorithm for the exact method converged in about one hour, while the algorithm for the approximate method converged in about 15 minutes. Thus, the approximate method is computationally much more efficient.

**Sampling method.** In both examples, we use the multivariate rejection sampling method. Other sampling methods may also be applied and may be even more efficient.

# Chapter 7

## Simulation Study

### 7.1 Introduction

In order to evaluate the performance of the two proposed methods: the exact method (EX) and the approximate method (AP), we conduct a simulation study. In our simulations, we prepare EX and AP in terms of biases and mean-squared errors of their estimates under various situations. We also add the complete-data (CD) method in our comparisons. Section 7.2 describes the data generation models, and Section 7.3 compares the two methods of estimation in three different situations. We conclude our result in Section 7.4.

## 7.2 Design of the Simulation Study

### 7.2.1 Models

We generate the responses variable  $y_{ij}$  from the following NLME

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{1i}} + P_{2i}e^{-\lambda_{2i}t_{2i}}) + \varepsilon_{ij}, \quad (7.1)$$

$$\begin{aligned} \log_{10}(P_{1i}) &= \beta_1 + b_{1i}, & \lambda_{1i} &= \beta_2 + b_{2i}, \\ \log_{10}(P_{2i}) &= \beta_3 + b_{3i}, & \lambda_{2i} &= \beta_4 + b_{4i}, \end{aligned} \quad (7.2)$$

where  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$  are the model parameters which are of interest,  $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$  are assumed to be *i.i.d* with a normal distribution  $N(0, D)$ , and  $D$  is a diagonal matrix with rank 4. The true values of  $\beta$  is taken as  $(10, 40, 8, 4)^T$ . The number of individuals is  $N = 48$ . The choice of  $t_{ij}$ ,  $n_{ij}$ , and  $\sigma^2$  will be reported in later sections with the results.

To evaluate the proposed methods, we generate some missing values of responses  $y_{ij}$ 's as follows. The model for missing responses is

$$\log \left( \frac{P(R_{ij} = 1 | \phi, \mathbf{b}_i)}{1 - P(R_{ij} = 1 | \phi, \mathbf{b}_i)} \right) = \phi_0 + \phi_1 b_{1i} + \phi_2 b_{2i} + \phi_3 b_{3i} + \phi_4 b_{4i}, \quad (7.3)$$

where  $\phi = (\phi_0, \phi_1, \phi_2, \phi_3, \phi_4)^T$  and  $r_{ij}$  is the missing responses indicator. The above missingness model suggest that the missingness of the responses depends on the random effects of each individual, and thus the responses is nonignorable missing. We generate missing responses based on model (7.3). We choose appropriate values of  $\phi$  to mimic certain missing rate, and we generate binary data  $r_{ij}$  based on values of  $\phi$  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.



### 7.2.2 Comparison Criteria

We compare EX and AP in terms of 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. The bias for  $\beta_j$ , the  $j$ th component of  $\beta$ , is defined as

$$\text{bias}_j = \hat{\beta}_j - \beta_j, \quad (7.4)$$

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

$$\text{MSE}_j = \text{bias}_j^2 + s_j^2, \quad (7.5)$$

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

$$\text{bias}_j / \beta_j \times 100\%, \quad (7.6)$$

and the percent relative root MSE is

$$\sqrt{\text{MSE}_j} / |\beta_j| \times 100\%. \quad (7.7)$$

To show the difference between EX and AP, we also calculate the mean and standard error of the absolute differences between the estimates from EX and AP.

## 7.3 Simulation Results

### 7.3.1 Comparison of Methods with Varying Missing Rates

To check the impact of the missing rates on estimation by EX and AP, we estimate the parameters based on three missing rate respectively. A missing rate of 10% and

20%. Specifically, in our model (7.3), we set  $\phi = (-2.6, 0.1, -0.1, -0.1, 0.1)^T$  to get roughly an 10% missing rate, and set  $\phi = (-1.4, 0.1, -0.1, -0.1, 0.1)^T$  to get an average of 20% missing rate. Since the random-effects based probability of a missing observation is a constant for a particular subject, there are situations where a subject has a high missing probability and the responses are missing for ALL visits. To overcome this difficulty, we always keep the responses for the first two visits. That is, every subject has at least two observations. Specifically, we choose  $t_i = (0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0)$ , the relative scheduled time points, for each subject, and generate  $r_{ij}$  ( $j = 1, \dots, n_i = 9$ ). If  $r_{ij} = 1$ , we manually remove the  $j$ th value from the generated responses  $y_{ij}$  ( $j = 1, \dots, n_i = 9$ ) for the  $i$ th subject. We choose  $D = \text{diag}(4, 25, 4, 1)$  and  $\sigma = 0.5$  in these simulations. To check the effect of each factor, we change a factor each time and compare the estimates with the original ones. Therefore, in the following tables, one half of each table is exactly the same.

Table 7.1 shows average simulation results based on 150 simulations. We see that the exact method performs better than the approximate method in the sense that the exact method yields smaller relative MSE and smaller bias. CD may yield very biased estimates, especially for the most important parameter  $\beta_2$ , the initial viral decay rate. All methods performs better when the missing rate is lower. The absolute differences between EX and AP are similar under different missing rates.

Table 7.1: Simulation results for varying missing rates.

Missingness rate (%)	True values	%bias			%√MSE			EX-AP	
		EX	AP	CD	EX	AP	CD	Mean	SD
10	$\beta_1 = 10$	3.5	3.2	2.9	4.7	12.11	20.10	0.05	0.02
	$\beta_2 = 40$	-3.1	-5.5	-2.5	9.5	27.82	21.77	0.97	0.64
	$\beta_3 = 8$	0.22	-0.5	0.76	4.1	5.64	11.76	0.06	0.03
	$\beta_4 = 4$	0.33	-2.8	1.10	6.7	7.50	9.01	0.03	0.01
20	$\beta_1 = 10$	4.71	3.66	-10.53	11.03	21.22	25.03	0.08	0.06
	$\beta_2 = 40$	6.11	-10.29	30.33	12.76	21.83	40.93	1.33	0.84
	$\beta_3 = 8$	3.23	-3.01	-6.11	5.32	12.78	17.19	0.06	0.05
	$\beta_4 = 4$	-2.77	5.75	-3.75	5.66	22.50	24.71	0.07	0.03

### 7.3.2 Comparison of Methods with Different Random Effects Covariances

To see how the variability of  $\mathbf{b}_i$  affects the estimates from the three methods, we consider two variance-covariance matrices for  $\mathbf{b}_i$ :

$$\text{Var}(\mathbf{b}_i) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \text{Var}(\mathbf{b}_i) = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Table 7.2 shows average simulation results based on methods EX, AP, and CD. We choose  $t_i = (0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0)$ , the relative scheduled time points, for each subject, and generate  $r_{ij}$  ( $j = 1, \dots, n_i = 9$ ). If  $r_{ij} = 1$ , we manually remove the  $j$ th value from the generated responses  $y_{ij}$  ( $j = 1, \dots, n_i = 9$ ) for the  $i$ th subject. We choose  $\phi = (-2.0, 0.1, -0.1, -0.1, 0.1)^T$  and  $\sigma = 0.5$  in these simulations. The response missing rate is roughly 10%. We may conclude that, for most of the parameters, the exact method performs better than the approximate method in the

Table 7.2: Simulation results for different covariance matrices for random effects.

Covariance Matrix	True values	%bias			%√MSE			EX-AP	
		EX	AP	CD	EX	AP	CD	Mean	SD
Var( $\mathbf{b}_i$ ) =diag(1,9,1,1)	$\beta_1 = 10$	2.7	2.4	2.1	3.9	5.6	9.1	0.02	0.01
	$\beta_2 = 40$	2.3	-9.0	5.0	6.0	11.2	15.7	0.23	0.14
	$\beta_3 = 8$	0.4	3.1	2.1	2.1	2.4	12.6	0.02	0.01
	$\beta_4 = 4$	1.5	2.5	5.0	1.9	5.5	6.5	0.03	0.02
Var( $\mathbf{b}_i$ ) =diag(4,25,4,1)	$\beta_1 = 10$	3.5	3.2	2.9	4.7	12.11	20.10	0.05	0.02
	$\beta_2 = 40$	-3.1	-5.5	-2.5	9.5	27.82	21.77	0.97	0.64
	$\beta_3 = 8$	0.22	-0.5	0.76	4.1	5.64	11.76	0.06	0.03
	$\beta_4 = 4$	0.33	-2.8	1.10	6.7	7.50	9.01	0.03	0.01

The lower half of this table is identical to the upper half of Table (7.1)

sense that the exact method yields smaller relative MSEs and smaller bias. Both methods perform better when the variances of the random effects are smaller. For  $\beta_1$ , the estimate under approximate method has slightly smaller bias. The absolute differences between EX and AP have smaller mean and standard deviation when the variances of  $\mathbf{b}$  are smaller.

### 7.3.3 Comparison of Methods with Varying Intra-individual Measurements

To examine how the number of intra-individual measurements affect our estimates, we consider the two methods of estimation under two maximum number of measurements,  $n_i = 9$  and  $n_i = 15$ . For the case in which maximum number of measurements is 9, we choose  $t_i = (0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0)$  as the relative scheduled time points, for each subject. For the case in which the maximum number of measurements is 5, we choose  $t_i = (0.01, 0.02, 0.04, 0.07, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)$  as the relative scheduled time points for each subject. In both cases,

Table 7.3: Simulation results for varying intra-individual measurements.

# of obs. per patient	True values	%bias			%√MSE			EX-AP	
		EX	AP	CD	EX	AP	CD	Mean	SD
2 to 9	$\beta_1 = 10$	3.5	3.2	2.9	4.7	12.11	20.10	0.05	0.02
	$\beta_2 = 40$	-3.1	-5.5	-2.5	9.5	27.82	21.77	0.97	0.64
	$\beta_3 = 8$	0.22	-0.5	0.76	4.1	5.64	11.76	0.06	0.03
	$\beta_4 = 4$	0.33	-2.8	1.10	6.7	7.50	9.01	0.03	0.01
2 to 15	$\beta_1 = 10$	2.1	3.0	2.8	3.5	10.9	11.4	0.03	0.01
	$\beta_2 = 40$	-1.9	-3.0	5.7	7.5	11.9	9.6	0.22	0.12
	$\beta_3 = 8$	-0.10	0.1	1.3	2.7	4.9	9.5	0.03	0.02
	$\beta_4 = 4$	0.30	-2.0	0.88	3.6	2.7	6.1	0.05	0.02

The upper half of this table is identical to the upper half of Table (7.1)

we always keep the responses values for the first two visits of each patients, but the responses from the third visit may be missing. Therefore, the actual number of observations for each patient may vary from 2 to 9 or 2 to 15 respectively. Additionally, We use the missing rate of 10% (roughly),  $\text{Var}(\mathbf{b}_i) = \text{diag}(4, 25, 4, 1)$ , and  $\sigma = 0.5$  in these simulations.

The results of 100 simulations are summarized in Table 7.3. We see that, when there are more intra-individual measurements, all methods performs better. Again, the exact method performs better than the approximate method, no matter what the number of intra-individual measurements are. CD performs worse than EX and AP. The absolute differences between EX and AP have smaller mean and standard deviation when there are more observations for each subjects.

#### 7.3.4 Comparison of Methods with Different Variances

To investigate the impact of intra-individual variability on EX and AP, we estimate the parameters based on two data generation strategy, with  $\sigma = 0.5$  and 1 respectively.

Table 7.4: Simulation results for varying variances.

$\sigma$	True values	%bias			%√MSE			EX-AP	
		EX	AP	CD	EX	AP	CD	Mean	SD
$\sigma = 0.5$	$\beta_1 = 10$	3.5	3.2	2.9	4.7	12.11	20.10	0.05	0.02
	$\beta_2 = 40$	-3.1	-5.5	-2.5	9.5	27.82	21.77	0.97	0.64
	$\beta_3 = 8$	0.22	-0.5	0.76	4.1	5.64	11.76	0.06	0.03
	$\beta_4 = 4$	0.33	-2.8	1.10	6.7	7.50	9.01	0.03	0.01
$\sigma = 1$	$\beta_1 = 10$	3.4	3.3	3.1	6.3	14.1	19.18	0.06	0.04
	$\beta_2 = 40$	-2.7	-5.1	6.9	8.8	27.3	27.10	0.78	0.54
	$\beta_3 = 8$	0.31	-0.6	1.2	6.35	9.8	9.28	0.05	0.08
	$\beta_4 = 4$	0.23	-2.7	1.9	7.15	9.7	17.3	0.03	0.03

The lower half of this table is identical to the upper half of Table (7.1)

Table 7.4 shows the average simulation results based on 150 simulations. In these simulations, we choose  $t_i = (0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0)$  as the relative scheduled time points for each subject, and generate  $r_{ij}$  ( $j = 1, \dots, n_i = 9$ ). If  $r_{ij} = 1$ , we manually remove the  $j$ th value from the generated responses  $y_{ij}$  ( $j = 1, \dots, n_i = 9$ ) for the  $i$ th subject. We choose  $\text{Var}(\mathbf{b}_i) = \text{diag}(4, 25, 4, 1)$  and the missing rate is 10% roughly. We find that, EX performs better than AP and yields less biased estimates. Complete-case methods tend to over estimate the initial decay rate, and yields large estimation variances. The absolute differences between EX and AP are relatively similar.

## 7.4 Conclusions

Based on the simulation results in the preceding sections, we may conclude as follows.

- EX and AP results are quite close relatively.

- For most cases, EX performs better than AP in the sense that EX yields smaller bias and smaller relative MSEs. However, in some cases, the estimates from AP may have smaller bias. CD often performs the worst in sense of large MSE.
- Both methods perform better when the missing rates are lower.
- Both methods perform better when the variances of random effects are smaller.
- Both methods perform better when the number of observations for each subject is larger.
- The Complete-case method tends to overestimate the initial decay rate and the corresponding estimates are positive biased, especially when the missing rate is large.

Note that, in our simulation studies, AP is computationally more efficient than EX. The running time of AP is about 1/4 of that of the EX, and it does not have convergence problems. In practice, the EX is preferable when the computation load is not too heavy. When EX is too slow, AP is preferable.

## Chapter 8

# Conclusion and Discussion

In this thesis, we have proposed two methods to estimate the parameters for NLMEs with random effects based informative dropout and missing covariates. The proposed methods include an exact method and an approximate method, both are implemented by a Monte Carlo EM algorithm. For the exact method, sampling the random effects may offer potential computational difficulties such as slow or non-convergence, especially when the dimension of random effects is large. To overcome this difficulty, we proposed an approximate method which integrates out the random effects in the E-step and thus avoids sampling the random effects in the Monte Carlo EM. Pinheiro and Wu (2001) show that convergence rate of the EM algorithm can be improved greatly by integrating out the random effects.

We also conducted a simulation study to compare the performance of the exact method and the approximate method. In our simulations, the exact method gives somewhat more reliable results than the approximate method in the sense that it provides smaller mean squared errors on the parameter estimates. Our simulations also suggest that the proportion of missing values, the variances of random effects, the number of intra-individual measurements, and the intra-individual variabilities



may affect the performance of the exact method and the approximate method.

The exact method was applied to an HIV dataset with missing covariates and informative dropouts. We find that patients who have higher baseline CD4 cell counts and slower initial viral decay rates may be more likely to dropout. Thus, ignoring dropouts may lead to over-optimistic assessment of the antiviral treatment. We also applied the exact and the approximate method on the second HIV dataset. We obtain similar conclusions as the first dataset. We find that the approximate method may under-estimate the initial viral decay rate, and thus may be somewhat conservative in assessing the treatment effect. It may be caused by the lack of adequacy of our approximate model. We also notice that the estimation of standard errors in approximate method is not reliable, and Jackknife standard error is recommended. But, in practice, the approximate method is computationally much more efficient than the exact method.

Since the parametric models for the missing responses are not testable based on the observed data, it is important to conduct sensitivity analyses. Based on our sensitivity analyses, we find that the estimates are robust to different dropout models.

Finally, we give an outline for possible future work.

- (1) For simplicity, in our examples and simulations, we only include random effects in the dropout models. It is conceivable that the dropout probability may also depend on covariates and the responses. In the future, we may study dropout models which simultaneously consider random effects, covariates, and responses.
- (2) In our study, we only consider nonlinear mixed effect models for normal data. Generally, our proposed methods may be extended to other models, such as generalized linear mixed effects models (GLMMs) and generalized nonlinear

mixed effects models (GNLMMs).

- (3) We have only considered baseline covariates. Our proposed methods may be extended to time-dependent covariates.
- (4) Multivariate rejection sampling methods were used in our 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.
- (5) In our approximate method, we used first order Taylor expansion to linearize the model. In general, it is not necessarily a good approximation. In the future, we may investigate better approximation, such as higher order Taylor expansion, Laplace approximation, etc.
- (6) Investigation of accuracy of standard error estimation. In our work, the estimation of SEs maybe unreliable, especially in the approximate method.
- (7) In our work, we consider  $\mathbf{e}_i | \boldsymbol{\beta}_i \sim N(0, \sigma^2 I)$ . In the future, we may consider more complicated relationships on the error terms. We may consider random effects model for longitudinal data with  $e_{ij}$  being an AR(1) time series.

# References

- 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.
- Davidian, M., and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurements Data*. Chapman & Hall.
- 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., Heagerty, P., Liang, K, and Zeger, S. (2002) *Analysis of Longitudinal Data, 2nd Edition*. Oxford.
- 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.
- 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.
- Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.
- 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. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765-769.
- 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.
- Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (2001). Missing responses in generalized linear models when the missing data mechanism is nonignorable. *Biometrics* **88**, 551-564.
- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673-687.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

- Little, R. J. A. (1992). Regression with missing  $X$ 's: a review. *Journal of the American Statistical Association* **87**, 1227-1237.
- 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 Schlucher, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497-512.
- Liu, C., Rubin, D. B., and Wu, Y. N. (1998), Parameter expansion for EM acceleration: the PX-EM algorithm, *Biometrika* **85**, 755-770.
- 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.
- McLachlan, G. J. and Krishnan, T. (1997). *The EM-algorithm and Extension*. John Wiley & Sons.
- Meng, X. L. and van Dyk, D. A. (1997). The EM algorithm - an old folk-song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Ser. B.* **59**, 511-567.
- 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.
- Pinhairo, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects models. *Journal of Computational and Graphical Statistics* **4**, 12-35.

- Pinheiro, J. C. and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate  $t$ -distribution. *Journal of Computational and Graphical Statistics* **10**, 249-276.
- Robert, C. P. and Casella G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* **97**, 40-52.
- 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., and Chinchilli, V. M. (1996). *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, 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* **55**, 410-418.
- Wu, H. and Ding, A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410-418.
- Wu, H. and Wu, L. (2001). A multiple imputation method for missing covariates in nonlinear mixed-effect models, with application to HIV dynamics. *Statistics in Medicine* **20**, 1755-1769.
- Wu, K. (2003). Simultaneous inference for generalized linear mixed models with informative dropout and missing covariates. *Master Thesis, the University of British Columbia*.
- 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, a). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates, *Journal of the American Statistical Association* **99**, 700-709.
- Wu, L. (2004, b). Exact and approximate simultaneous inferences for nonlinear mixed-effects models with dropouts and missing covariates, *Technical report, the Department of Statistics, the University of British Columbia*.
- Wu, L. and Wu, H. (2002). Missing time-dependent covariates in HIV viral dynamic models. *Journal of the Royal Statistical Society, Ser. C* **51**, 297 - 318.