Simultaneous Inference for Generalized Linear Mixed Models with Informative Dropout and Missing Covariates

by

Kunling Wu

M.Sc., Beijing Polytechnic University, China, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming
to the required standard

The University of British Columbia

December 2003

Library Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Kunling Wu — 19/12/2003

Name of Author *(please print)* — Date (dd/mm/yyyy)

Title of Thesis: Simultaneous Inference for Generalized Linear Mixed Models with Informative Dropout and Missing Covariates

Degree: Master of Science Year: 2003

Department of Statistics

The University of British Columbia
Vancouver, BC Canada

Abstract

Generalized linear mixed effects models (GLMMs) are popular in many longitudinal studies. In these studies, however, missing data problems arise frequently, which makes statistical analyses more complicated. In this thesis, we propose an exact method and an approximate method for GLMMs with informative dropouts *and* missing covariates, and provide a unified approach for simultaneous inference. Both methods are implemented by Monte Carlo EM algorithms. The approximate method is based on Taylor series expansion, and it avoids sampling the random effects in the E-step. Thus, the approximate method may be computationally more efficient when the dimension of random effects is not small. We also briefly discuss other methods for accelerating the EM algorithms. To illustrate the proposed methods, we analyze two real datasets, a AIDS 315 dataset and a dataset from a parent bereavement project, using these methods. A simulation study is conducted to evaluate the performance of the proposed methods under various situations.

Contents

List of Tables

List of Figures

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 UBC. Without his support, expertise and patience, this thesis would not have been completed. Also, I would like to thank my second reader, Dr. Paul Gustafson, for his invaluable comments and suggestions on this thesis.

Furthermore, I thank Dr. Nancy Heckman and Dr. Bertrand Clarke for their invaluable advice on my consulting projects, which will benefit me very much in the future. I thank all the faculty and staff in the Department of Statistics at UBC for providing such a nice academic environment. I also thank all the graduate students in the Department of Statistics for making my study at UBC so enjoyable.

Most importantly I would like to thank my parents for loving me and believing in me. My big thanks goes to my beloved husband, Weiliang Qiu, for his love, his constant support and encouragement, which push me to be the best at everything I do.

KUNLING WU

The University of British Columbia
December 2003

To my parents and husband.

Chapter 1

Introduction

1.1 Generalized Linear Mixed Effect Models

Longitudinal data or repeated measurement data occur frequently in many applications where repeated measurements are obtained for each individual. Statistical analysis of longitudinal data are reviewed in Diggle, Liang and Zeger (1994). One of the key advantages of a longitudinal study over a cross-sectional study is to separate variation over time within an individual from difference among individuals, while a cross-sectional study can not do this because it simply records one measurement for each individual. So the analysis of cross-sectional data may confound time effect and may give misleading results. For longitudinal data, it is important to recognize two sources of variations: intra-individual variation produced by different measurements within a given individual, and inter-individual variation among different individuals.

Generalized linear models (GLMs) such as logistic regression models, extend normal linear models to allow non-normal error distributions in the natural exponential family such as Poisson and binomial distributions. GLMs can handle not only continuous variables but also discrete variables, as long as the distribution of the variable belongs

to the natural exponential family. Therefore, GLMs provide a unified different approach for continuous and discrete responses and have wide applicability in practice. For example, in Agresti (1990), a sample of male residents of Framingham, Massachusetts, was collected according to their blood pressures. During a follow-up period, whether or not these male residents developed coronary heart diseases, was recorded and viewed as response. So the response variable is binary. To investigate the relationship between the blood pressure and the coronary heart disease, we can build a logistic regression model and then make statistical inferences based on this GLM. Generalized linear mixed models (GLMMs) are useful for longitudinal studies which extend GLMs by introducing random effects to account for correlation within the repeated measurements for a given individual. Such models can separate two kinds of variations, borrow information cross individuals and allow discrete and continuous responses. Therefore, GLMMs are very popular in the analysis of longitudinal data. A GLMM may be written as a hierarchical two-stage model. At the first stage, intra-individual variation is charactered by a GLM. In the second stage, inter-individual variation is represented through individual-specific regression parameters. Covariates are often introduced in the second stage to partially explain systematic variation.

There are two main approaches to estimate parameters in GLMMs: (i) an exact likelihood inference based on numerical integration (Booth and Hobert (1999)), and (ii) an approximate inference based on linearization procedures via Taylor series expansion (Breslow and Clayton (1993); Vonesh *et al.* (2002)). In the exact inference, marginal likelihood is obtained by integrating out random effects from the joint distribution of response and random effects. By maximizing the marginal likelihood, we obtain estimates of parameters of interest. However, the integration is usually intractable, one may use Monte Carlo approximations to evaluate it (Wei and Tanner (1990)). The exact likelihood

inference works well with a small dimension of random effects. However, computation may become quite demanding or unstable as the dimension of random effects increases. In such cases, we may consider the approximate inference which avoids this computation difficulty by integrating out the random effects. The strategy for the approximate inference is to iteratively solve LME models based on second-order Taylor series expansion around current estimates. If the number of measurements for each individual is large enough, approximate methods may give reasonable estimates for parameters. Otherwise, approximate maximum likelihood estimates may be inconsistent.

1.2 Missing Data Problems

Missing data are a serious problem in many applications and arise frequently in longitudinal studies. Two kinds of missing data often occur in a longitudinal study: (i) missing covariates due to different measurement schedules for covariates and response or other problems; and (ii) missing responses due to dropout or missing visits. For example, individuals may withdraw or die before the end of study or do not come to the study center for measurements at scheduled times for various reasons. Missing data problems make statistical analysis in longitudinal studies much more complicated, since standard complete-data methods are not directly applicable.

Commonly-used naive methods for missing data include the complete-case method, which deletes all incomplete observations, the mean imputation method, which substitutes missing values with the mean values of observed data, and the last-value-carried-forward method, which imputes a missing value by the immediate previous observed data. At the presence of missing data, the missing data mechanism must be taken into account in order to obtain valid statistical inference. Little and Rubin (1987) define three types

of missing data mechanisms. Missing data are missing completely at random (MCAR) if the probability of missingness is independent of both observed and unobserved data. For example, patients do not come to the study center because of reasons irrelevant to the treatment such as simply forgetting the appointment. 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 occasionally fail to visit the clinic because he/she is too old. Missing data are nonignorable or informative missing data (NIM) if the probability of missingness depends on unobserved data. For example, a patient fails to visit the clinic because he/she is too sick. If missing values are MCAR, the complete-case method will give unbiased, but inefficient estimates. If the missing data are not MCAR, the naive methods may give biased, even misleading results due to not taking missing data information into consideration. MCAR and MAR are called ignorably missing. We can ignore the missing data mechanism in likelihood inference when missing values are ignorably missing (Little and Rubin (1987)).

Little (1992, 1995) gave a review on missing covariates in regression and drop-out in repeated-measures studies. Ibrahim, Lipsitz, and Chen (1999) proposed a Monte-Carlo EM method for estimating parameters in GLMs with nonignorable missing covariates. Wu and Wu (2001) estimated parameters in nonlinear mixed effects models with missing covariates (MAR) by a three-step multiple imputation method. Wu and Carroll (1988) considered linear mixed effect models with informative dropout and assumed missingness depending on random effects. Ibrahim, Chen and Lipsitz (2001) developed a Monte Carlo EM algorithm for estimating parameters in GLMMs with informative dropout. However, little literature considers parameter estimation in GLMMs with informative dropout *and* missing covariates simultaneously.

1.3 Motivating Examples

1.3.1 Example 1

Our research is motivated by a longitudinal study from the AIDS Clinical Trial Group (ACTG) Protocol 315 (Wu and Ding (1999)). In this study, 46 HIV infected patients were treated with a potent antiviral drug. Plasma HIV-1 RNA (viral loads) were repeatedly quantified on days 2, 7, 10, 14, 21, 28, and weeks 8, 12, and 24 after initiation of treatment. After the antiviral treatment, the patients' viral loads will decay, and the decay rate may reflect the efficacy of the treatment. The Nucleic Acid Sequence-Based Amplification assay that is used to measure the viral load has a detection limit. If the viral load drops below the detection limit after the treatment, the viral load can not be measured, which may indicate that the treatment may be successful. To investigate the treatment effect, one approach is to define the response as whether the viral load is below the detection limit or not, which is thus a binary variable. In this study, patients drop out before the end of the study, and the dropout may be informative. Thus, the response contains non-ignorable missing values. Preliminary studies show that some baseline covariates such as CD4 cell counts, tumor necrosis factor (measured by TNF levels) and total complement levels (measured by CH50), may partially explain variation in the viral load trajectory. However, some of these covariates are also missing. Our objectives are to model the viral load trajectory and to identify covariates that may partially predict changes of viral loads, in the presence of informative dropouts and missing covariates.

1.3.2 Example 2

Our second example involves a longitudinal study from a parent bereavement project, which investigates the long-term mental outcomes of parents whose children died by

accident, suicide, or homicide. After their children's death, the parents usually experience a high level of mental distress. In this study, the mental distress of 239 parents were measured at baseline (*i.e.* 4 to 6 weeks after their children's death), and then at 4, 12, 24 and 60 months post-death. The Global Severity Index (GSI), which is the most sensitive indicator of mental distress, is used to measure the parents' distress levels. A high GSI score indicates a high level of mental distress. If the parents' adjustment to their children's death goes well, their GSI scores will decrease over time, at least lower than their baseline GSI scores. To examine how the parents' mental distress changes over time after their children's death, we treat whether or not a parent's GSI score after baseline is lower than his/her baseline value as response. Several other relevant factors were also obtained at baseline, including parents' gender, marital status, age, education, annual income, the cause of death, age and gender of the deceased child. These baseline factors may be important predictors of parents' distress and thus are viewed as covariates. Note that some baseline covariates such as income contain missing values, and some responses are also missing. Our objectives are to investigate the change of parent's distress levels over time and to determine which covariates affect the parent's mental distress.

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 GLMMs with informative dropout *and* missing covariates. To avoid computational difficulties when the dimension of random effects is not small, we propose an approximate inference method, which integrates out the random effects for more efficient computation.

The remainder of this thesis is organized as follows. Chapter 2 introduces GLMs

and GLMMs and reviews the literature about informative dropout and missing covariates. Chapter 3 discusses the exact inference method for estimation of GLMMs 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 two real data examples. Chapter 7 presents our simulation study. We conclude the thesis with a discussion in Chapter 8.

Chapter 2

Generalized Linear Mixed Models and Missing Data

2.1 Introduction

Before we present our methods for estimating parameters in GLMMs with informative dropout and missing covariates, we give a brief introduction to GLMs, GLMMs, and methods for the missing data problems in this chapter. In Section 2.2, we introduce GLMs and the methods of estimation for parameters in GLMs. Section 2.3 describes GLMMs, briefly discusses two main methods for estimating parameters in GLMMs and reviews the literature for GLMMs. In Section 2.4, we give a literature review about methods of handling informative dropout and missing covariates respectively.

2.2 Generalized Linear Models

A classical linear model is useful to model a continuous response under the assumption that the response follows a normal distribution and a linear relationship exists between

the mean of the response and covariates. However, in practise, some non-normal distributions such as binomial, Poisson, etc, may be better assumptions for some response variables such as discrete variables. For example, we may want to study whether developing a heart disease relates to the blood pressure level. Here, we treat the health status of patients' heart as our response. The response is thus a binary variable which takes values of 0 or 1, where 0 means that a patient has a heart disease and 1 means that a patient has no heart disease. Obviously, here the assumption of normality is completely unrealistic. Moreover, frequently the mean of the response can not be expressed as a linear form of the covariates. In those situations, we can not use standard linear models.

Generalized linear models (GLMs), which are a extension of classical linear models, can not only deal with variables whose distributions come from the exponential family but also allow nonlinear forms between the mean of responses and the covariates. Variables in the exponential family include continuous variables such as normal and exponential, and discrete variables such as binomial and Poisson. Due to the capability to handle continuous data as well as discrete data, GLMs unify different methodologies and thus have wide applicability in practice.

2.2.1 Model Specification

GLMs are specified by three components including a random component, a systematic component, and a link function.

Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^T$ be a vector of independent and identically distributed $(i.i.d)$ observations whose distribution belongs to the natural exponential family. Then the density function of each observation y_i can be expressed in the form

$$f(y_i; \beta_i) = \exp\{[y_i\theta_i - \varphi(\theta_i)]/a(\phi) + c(y_i, \phi)\}, \tag{2.1}$$

where $a(.)$, $\varphi(.)$ and $c(.)$ are specific functions, $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)^T$ is called the natural parameters, and ϕ is called the dispersion parameter. The random component of GLMs is specified by the above density function of the response variable.

The systematic component specifies the relation between covariates $\boldsymbol{x}_i$ to the linear predictor η_i by a linear form

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad i = 1, 2, \cdots, N, \tag{2.2}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters.

The mean $\mu_i = E(y_i)$ is related to the linear predictor through the link component of GLMs

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = g(\mu_i) \quad i = 1, 2, \cdots, N, \tag{2.3}$$

or

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \quad i = 1, 2, \cdots, N, \tag{2.4}$$

where $g(.)$ which is a monotone and differentiable function called the link function. In the exponential family, if a link function $g(.)$ satisfies $g(\mu_i) = \theta_i(\mu_i)$, then the link is called the canonical link or natural link. Binomial, Poisson and normal variables all have canonical link functions. A function $g(\mu_i) = \mu_i$ gives the identity link. For example, normal variables have the identity link function. In summary, GLMs allow for linear as well as non-linear models under a single framework. Moreover, GLMs make it possible to fit models where the underlying data are normal, Poisson, binomial, etc, by a suitable choice of the link functions.

2.2.2 Maximum Likelihood Estimation in GLMs

The principal method of estimation used in GLMs is the maximum likelihood method. In this section, we will briefly describe how to obtain the maximum likelihood esti-

mates of parameters in GLMs. We assume ϕ is known, then $c(y_i, \phi)$ is a constant in the log-likelihood function about $\boldsymbol{\theta}$ and thus is not ignored in the following log-likelihood function. For N independent observations, the log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{i=1}^{N} l_i(\theta_i|y_i) \\ &= \sum_{i=1}^{N} \log f(y_i|\theta_i;\phi) \\ &= \sum_{i=1}^{N} \frac{y_i\theta_i(\mu_i) - \varphi(\theta_i(\mu_i))}{a(\phi)}. \end{aligned} \tag{2.5}$$

Some useful Equations

Now we will derive some useful identities used in maximizing the likelihood function. The derivation of (2.5) with respect to θ_i gives

$$\frac{\partial l}{\partial \theta_i} = \frac{1}{a(\phi)}\left(y_i - \frac{\partial \varphi(\theta_i)}{\partial \theta_i}\right), \tag{2.6}$$

$$\frac{\partial^2 l}{\partial \theta_i^2} = -\frac{1}{a(\phi)}\frac{\partial^2 \varphi(\theta_i)}{\partial \theta_i^2}. \tag{2.7}$$

The following is two well-known likelihood results that we use here:

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta_i}\right) = 0, \tag{2.8}$$

$$\mathrm{var}\left(\frac{\partial l}{\partial \theta_i}\right) = -\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta_i^2}\right). \tag{2.9}$$

Substituting (2.6) and (2.7) to (2.8) and (2.9) respectively gives

$$\mathrm{E}(y_i) = \mu_i(\theta_i) = \frac{\partial \varphi(\theta_i)}{\partial \theta_i}, \tag{2.10}$$

and

$$\mathrm{var}(y_i) = \frac{1}{a(\phi)}\frac{\partial^2 \varphi(\theta_i)}{\partial \theta_i^2} = \frac{1}{a(\phi)}\frac{\partial \mu_i(\theta_i)}{\partial \theta_i} = \frac{1}{a(\phi)}V(\mu_i), \tag{2.11}$$

where $V(\mu_i) = \partial\mu_i(\theta_i)/\partial\theta_i$ is often called the variance function. Equation (2.11) indicates that the variance of the response depends on its mean. We differentiate two sides of equation (2.3) with respect to $\boldsymbol{\beta}$ and obtain

$$\boldsymbol{x}_i = \frac{\partial g(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}}.$$

Upon rearrangement, the above equation can be written as

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{V(\mu_i)\partial g(\mu_i)/\partial \mu_i}\boldsymbol{x}_i. \tag{2.12}$$

Maximum Likelihood Estimation

To obtain the maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$, we differentiate (2.5) with respect to $\boldsymbol{\beta}$, and then apply (2.10), (2.11) and (2.12) to get the following score function

$$\begin{aligned} S(\boldsymbol{\beta}) &= \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} \frac{\partial l_i(\theta_i|y_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{a(\phi)}\sum_{i=1}^{N}\frac{y_i - \mu_i}{V(\mu_i)\partial g(\mu_i)/\partial \mu_i}\boldsymbol{x}_i. \end{aligned} \tag{2.13}$$

Let

$$W = \text{diag}^{-1}\left\{V(\mu_1)(\partial g(\mu_1)/\partial \mu_1)^2, \cdots, V(\mu_N)(\partial g(\mu_N)/\partial \mu_N)^2\right\},$$

$$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N),$$

$$\Delta = \text{diag}\left\{\partial g(\mu_1)/\partial \mu_1, \partial g(\mu_2)/\partial \mu_2, \cdots, \partial g(\mu_N)/\partial \mu_N\right\},$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_N)^T.$$

Then (2.13) can be rewritten as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = S(\boldsymbol{\beta}) = \frac{1}{a(\phi)}XW\Delta(Y - \mu(\boldsymbol{\beta})). \tag{2.14}$$

MLEs can be obtained by solving the following score equation

$$S(\boldsymbol{\beta}) = \frac{1}{a(\phi)} XW\Delta(Y - \mu(\boldsymbol{\beta})) = 0. \tag{2.15}$$

The solution to the above equation (2.15) can be performed by Fisher scoring algorithm or Gauss-Newton algorithm. In the case of canonical links, both Fisher scoring and Newton-Raphson reduce to the iteratively re-weighted least squares algorithm.

Under the regularity condition, MLEs of parameters in GLMs have the asymptotic normality property

$$\hat{\boldsymbol{\beta}} \longrightarrow N(\boldsymbol{\beta}, a(\phi)(XWX^T)^{-1}).$$

As we see, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is equal to the inverse of the expected Fisher information matrix, which is

$$\mathrm{F}(\boldsymbol{\beta}) = -\mathrm{E}(\frac{\partial^2 l(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}) = \frac{1}{a(\phi)} XWX^T. \tag{2.16}$$

With a large sample size, we can apply this property to make inference about $\boldsymbol{\beta}$.

2.2.3 Quasi-Likelihood Approach

Based on the fact that the just first two moments of variables are mainly involved in the score function, Wedderburn (1974) proposed the quasi-likelihood method for estimating parameters in GLMs. The advantages of this method are that we do not need to make specific distribution assumptions, and that its estimators own the similar asymptotic properties as MLEs.

2.3 Generalized Linear Mixed Models

2.3.1 Generalized Linear Mixed Models

A GLMM is a extension of a GLM to longitudinal data by introducing random effects to account for correlation within repeated measurements for a given individual. It can separate the inter-individual variation and the intra-individual variation and borrow strength across individuals. Thus, a GLMM is very popular in the analysis of longitudinal data. A GLMM may be written as a hierarchical two-stage model. In the first stage, the intra-individual variation is specified by a generalized linear regression model. In the second stage, the inter-individual variation is represented through individual-specific regression parameters.

Let y_{ij} denote the jth observation on individual i, $i = 1, 2, \cdots, N$; $j = 1, 2, \cdots, n_i$. Then there are a total of $\sum_{i=1}^{N} n_i$ observations.

- stage 1 (intra-individual variation)

  Let $\boldsymbol{b}_i$ be the random effects associated with individual i. We assume that conditioning on $\boldsymbol{b}_i$, observations $y_{i1}, y_{i2}, \cdots, y_{in_i}$ are independent and each has the density function from the natural exponential family.

$$f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = \exp\{[y_{ij}\theta_{ij} - \varphi(\theta_{ij})]/a(\phi) + c(y_{ij}, \phi)\}, \tag{2.17}$$

$$E(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = \mu_{ij} = g(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i), \tag{2.18}$$

  where ϕ is a dispersion parameter, Here we assume that ϕ is known. The function $g(.)$ is the link function, $\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$ is the linear predictor, and $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are two vectors of covariates such as time, baseline value, etc.

- stage 2 (inter-individual variation)

$$\boldsymbol{\eta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b}_i, \tag{2.19}$$

$$\boldsymbol{b}_i \sim N(0, D), \tag{2.20}$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{in_i})^T$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i})$, and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \cdots, \boldsymbol{z}_{in_i})$, and $\boldsymbol{\beta}$ is a vector of fixed parameters. We assume that the random effects, $\boldsymbol{b}_i$'s, are *i.i.d.* The covariance matrix D in (2.20) quantifies the random inter-individual variation.

2.3.2 Maximum Likelihood Estimation

Let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$. From the preceding section, the joint density of $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)$ and $\boldsymbol{b} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_N)$ can be written as

$$f(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\beta}, D) = \prod_{i=1}^{N}\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|D). \tag{2.21}$$

Since random effects $\boldsymbol{b}$ are unobservable variables, we integrate out random effects and obtain the marginal distribution for $\boldsymbol{y}$

$$f(\boldsymbol{y}|\boldsymbol{\beta}, D) = \prod_{i=1}^{N}\int\prod_{j=1}^{n_i}\{f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|D)\}\, d\boldsymbol{b}_i. \tag{2.22}$$

Thus, the corresponding log-likelihood is

$$l(\boldsymbol{\beta}, D|\boldsymbol{y}) = \sum_{i=1}^{N}\log\left(\int\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i\right). \tag{2.23}$$

If the above log-likelihood has a closed form, we can obtain the MLEs of parameters in GLMMs by solving the score equation as usual. However, usually integration with respect to random effects is intractable such that we can not get the closed form for the log-likelihood. This problem results in two main approaches of estimation for parameters in GLMMs including an exact likelihood inference method based on numerical

integration and an approximate inference method based on linearization procedures via Taylor series expansion. In the exact inference, when integration becomes intractable due to moderate to large random effects, one may solve this problem by implementing the Monte Carlo EM algorithm. The exact inference method works very well with a small dimension of random effects. However, the computation may become quite demanding or unstable as the dimension of random effects increases, while the approximate inference method avoids the computation problem by integrating out the random effects. The strategy for the approximate inference method is to iteratively solve LME models based on first-order or second-order Taylor series expansion around current estimates. If the number of the intra-individual measurements (measurements for each individual) is large enough, the approximate method may give reasonable estimates for parameters. Otherwise, approximate MLEs may be inconsistent.

2.3.3 Literature for Generalized Linear Mixed Models

McCulloch (1997) derived a Monte Carlo Newton-Raphson algorithm and combined it with a simulated maximum likelihood method to come up with a hybrid method for GLMMs. His simulation study showed that the Monte Carlo EM algorithm, the Monte Carlo Newton-Raphson algorithm and the hybrid method worked well in calculating MLEs for GLMMs, and the hybrid method gave more precise estimators. Booth and Hobert (1999) proposed two new implementations for the maximum likelihood fitting in GLMMs. Both methods are carried out by the Monte Carlo EM algorithm. The main difference is that the first method uses a rejection sampling to generate samples from the exact conditional distribution of the random effects, while the second one uses a multivariate t importance sampling. Breslow (1993) proposed a penalized quasi-likelihood (PQL) method and a marginal quasi-likelihood (MQL) method, and demonstrated their

suitability for inference in GLMMs by simulation and application in several examples. In the simulation study, PQL and MQL made correct inferences on regression coefficients, but underestimated parameters a bit (in absolute value). Vonesh (2002) proposed a conditional second-order generalized estimation equation (CGEE2) to estimate the parameters in GLMMs and also showed that the efficiency of estimators was improved due to the involvement of the second-order moment.

2.4 Literature for Missing Data

We frequently encounter the missing data (response *or/and* covariate) problem in practise. However, ignoring missing data or using overly simple methods to handle missing data often leads to invalid inference. Thus, it is very important to find appropriate approaches to deal with missing data in our hand. Various strategies for considering the missing data mechanism have been proposed in the recent literature.

2.4.1 Literature of Informative Dropout

Wu and Carroll (1988) considered linear mixed effect models with informative dropout under the assumption that the informative dropout could be modeled by a probit model which included the random effects as its covariates. Diggle and Kenward (1994) derived a likelihood method to get MLEs in a multivariate linear model with informative dropout modeled by a logistic regression model which included the response as covariate. Computation of the likelihood was speeded up by using the probit approximation to the logit transformation. Their simulation work has shown that considering the informative dropout mechanism in the statistical inference reduces the bias caused by the ordinary least square (OLS) estimator or by only considering the informative dropout as MAR.

Little (1995) gave a review on modeling the dropout mechanism in repeated-measures studies. Regarding how to factor the dropout mechanism, models handling dropout were classified into selection models and pattern-mixture models. The main difference between two types of models is that we need to specify the form of missing data mechanism in the selection models while pattern-mixture models do not require that. He classified NIM into nonignorable Outcome-Based missing data where the dropout depends on missing values, and Random-effect-Based missing data where the dropout depends on future values. He also suggested to examine the sensitivity of results to the choice of missing data mechanism when we almost know nothing about the missing data mechanism. Ibrahim, Chen and Lipsitz (2001) developed a Monte Carlo EM algorithm to obtain MLEs in GLMMs with informative dropout and nonmontone missing data patterns. Moreover, they proposed that the missing data mechanism may be modeled by a logistic regression or a sequence of one-dimensional conditional distributions which may reduce the number of nuisance parameters.

2.4.2 Literature of Missing Covariates

Little (1992) defined three special types of patterns of missing covariates: (i) univariate missing data where only one covariate values are missing, (ii) monotone or nested missing data where the $(j+1)$th covariate x_{j+1} is observed for every case in which the jth ($j = 1, 2, \cdots, p$) covariate x_j is observed and (iii) a special pattern where two covariates can not be observed at the same time. He reviewed the methods of estimation in the regression models with missing covariates. The six reviewed statistical methods dealing with missing covariates are compared in this paper, including complete-case methods, available-case methods, least squares on imputed data, maximum likelihood, Bayesian methods and multiple imputation. He suggested that the maximum likelihood, Bayesian methods and

multiple imputation would be a better choice for dealing with missing covariate problems. Moreover, he preferred the maximum likelihood in a large sample and Bayesian methods or multiple imputation in a small sample. Ibrahim (1990) analyzed the problem of missing covariates (MAR) in GLMs with discrete covariates and applied the EM algorithm to obtain MLEs under the assumption that the missing covariates came from a discrete distribution. The asymptotic variance of MLEs was estimated by computing the observed information matrix via Louis's method. Ibrahim, Lipsitz, and Chen (1999) proposed a Monte-Carlo EM algorithm for estimating parameters in GLMs with nonignorable missing covariates. In this paper, they assumed a multinomial model for the missing data mechanism and a sequence of one-dimensional conditional distribution for unobserved covariates. Wu and Wu (2001) estimated parameters in nonlinear mixed effect 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 thus obtained the overall inference based on B analysis results.

Chapter 3

Exact Inference for GLMMs with Informative Dropout and Missing Covariates

3.1 Introduction

In this chapter, we develop an exact inference method based on numerical integration to obtain MLEs for parameters in GLMMs with informative dropout and missing covariates. The proposed exact method is implemented by a Monte Carlo EM algorithm, which need to generate samples for missing values and random effects by Gibbs sampler in each EM step. In Section 3.2, we give a description of GLMMs with informative dropout and missing covariates, considered in this thesis. Section 3.3 describes a Monte Carlo EM algorithm for implementing the exact inference method. A detailed description of our sampling methods is provided in Section 3.4. In Section 3.5, we present a PX-EM algorithm, which may boost the convergence rate of the standard EM algorithm. Computation issues regarding our algorithm are discussed in Section 3.6.

3.2 Models and Likelihood

We assume that data are collected from N individuals. Let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$, where y_{ij} is the outcome for individual i at time t_{ij}, $j = 1, 2, \cdots, n_i$, $i = 1, 2, \cdots, N$. The response $\boldsymbol{y}_i$ may contain missing values due to dropouts. So we write $\boldsymbol{y}_i = (\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i})$, where $\boldsymbol{y}_{obs,i}$ corresponds to the observed components of $\boldsymbol{y}_i$, and $\boldsymbol{y}_{mis,i}$ contains the missing components of $\boldsymbol{y}_i$. Let $\boldsymbol{r}_i = (r_{i1}, \cdots, r_{in_i})^T$ be a vector of missing *response* indicators such that $r_{ij} = 1$ if y_{ij} is missing for individual i at time t_{ij}, and $r_{ij} = 0$ if y_{ij} is observed for individual i at time t_{ij}. Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{if})^T$ be a $(f \times 1)$ vector of time-independent covariates for individual i. Since the time-independent covariates may also contain missing values, we write $\boldsymbol{x}_i = (\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i})$, where $\boldsymbol{x}_{obs,i}$ = the observed part of $\boldsymbol{x}_i$ and $\boldsymbol{x}_{mis,i}$ = the missing part of $\boldsymbol{x}_i$. Let $\boldsymbol{s}_i = (s_{i1}, \cdots, s_{if})^T$ be a vector of missing *covariate* indicators such that $s_{ik} = 1$ if x_{ik} is missing and $s_{ik} = 0$ if x_{ik} is observed, $k = 1, \cdots, f$.

Let $f(.)$ denote a generic density function. If the response and all covariates are completely observed for each individual, the corresponding GLMM can be written as a hierarchical two-stage model as follows.

$$f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = \exp\left[\{y_{ij}\theta_{ij}(\mu_{ij}) - \varphi(\theta_{ij}(\mu_{ij}))\}/a(\phi) + c(y_{ij}, \phi)\right], \tag{3.1}$$

$$\begin{aligned} \eta_{ij} &= g(\mu_{ij}) = A_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\overset{i.i.d}{\sim} N(0, D), \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, N, \end{aligned} \tag{3.2}$$

where $E(y_{ij}|\boldsymbol{b}_i) = \mu_{ij}$ and ϕ is the dispersion parameter (here we assume that ϕ is known). The function $g(.)$ is a link function, η_{ij} is the linear predictor, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a vector of fixed effects, and $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iq})^T$ is a vector of random effects. The covariate $A_{ij}^T = (\boldsymbol{x}_i^T, \boldsymbol{t}_{ij}^T)$ is a $(1 \times p)$ vector, where $\boldsymbol{t}_{ij}$ is a vector of time-dependent covariates. Usually, the covariate vector $\boldsymbol{z}_{ij}$ is a subset of A_{ij}. The $q \times q$ matrix D

quantifies the random inter-individual covariance. By integrating out the unobservable random effects $\boldsymbol{b}_i$, we obtain the following complete-data marginal distribution

$$f(\boldsymbol{y}_1 \cdots, \boldsymbol{y}_N | \boldsymbol{x}_1 \cdots, \boldsymbol{x}_N, \boldsymbol{\beta}, D) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \{ f(y_{ij} | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i) f(\boldsymbol{b}_i | D) \} d\boldsymbol{b}_i. \tag{3.3}$$

In the presence of missing values in the response and covariates, the complete-data marginal distribution becomes more complicated. When the missing responses are informative, we have to take into account the missing data mechanism, i.e., the distribution of the missing data indicators $\boldsymbol{r}_i$. Otherwise, the estimates of parameters may be biased. In this thesis, we make the following assumptions: (i) The missing covariates are MAR, i.e, the missing covariate mechanism does not depend on any unobserved values, but may depend on observed values. In other words, the density function for the missing covariate indicator $\boldsymbol{s}_i$ satisfies $f(\boldsymbol{s}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}) = f(\boldsymbol{s}_i | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a vector of parameters. (ii) The missing responses are informative, i.e, the missing response mechanism may depend on the unobserved values. We denote $f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi})$ as the density function of the missing response indicator, where $\boldsymbol{\psi}$ is a vector of parameters. (iii) Let $f(\boldsymbol{x}_i | \boldsymbol{\alpha})$ to be the density function for covariates $\boldsymbol{x}_i$, where $\boldsymbol{\alpha}$ is a vector of parameters. Modeling strategies for specifying the missing data mechanism $f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi})$ and the covariate model $f(\boldsymbol{x}_i | \boldsymbol{\alpha})$ are explored in Chapter 5. By integrating out $\boldsymbol{y}_{mis,i}$ and $\boldsymbol{x}_{mis,i}$, we obtain the marginal distribution for the observed data $(\boldsymbol{y}_{obs}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{s})$.

$$\begin{aligned} f(\boldsymbol{y}_{obs}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{s} | \boldsymbol{\beta}, D, \boldsymbol{\psi}, \boldsymbol{\delta}, \boldsymbol{\alpha}) = & \prod_{i=1}^{N} \int \int \int \prod_{j=1}^{n_i} \{ f(y_{ij} | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i) f(\boldsymbol{b}_i | D) f(\boldsymbol{x}_i | \boldsymbol{\alpha}) \\ & f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i,) f(\boldsymbol{s}_i | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\delta}) \} d\boldsymbol{b}_i d\boldsymbol{x}_{mis,i} d\boldsymbol{y}_{mis,i}, \end{aligned} \tag{3.4}$$

where $\boldsymbol{y}_{obs} = (\boldsymbol{y}_{obs,1}, \cdots, \boldsymbol{y}_{obs,N}), \boldsymbol{x}_{obs} = (\boldsymbol{x}_{obs,1}, \cdots, \boldsymbol{x}_{obs,N}), \boldsymbol{r} = (\boldsymbol{r}_1, \cdots, \boldsymbol{r}_N)$ and $\boldsymbol{s} = (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_N)$. Rubin (1976) showed that the missing data mechanism can be ignored from likelihood inference if the data are MAR. Since we assume that the missing covariates

are MAR, ignoring the missing covariates mechanism leads to the the following observed data log-likelihood:

$$l(\boldsymbol{\beta}, D, \boldsymbol{\psi}, \boldsymbol{\alpha} | \boldsymbol{y}_{obs}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{s}) = \sum_{i=1}^{N} \log \Big\{ \int \int \int \prod_{j=1}^{n_i} (f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i) f(\boldsymbol{x}_i|\boldsymbol{\alpha}) \\ f(\boldsymbol{b}_i|D) f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}) d\boldsymbol{b}_i d\boldsymbol{x}_{mis,i} d\boldsymbol{y}_{mis,i} \Big\}. \tag{3.5}$$

Maximizing the above log-likelihood gives us the MLEs for parameters in the GLMM. However, the intractable integration in (3.5) makes the observed data log-likelihood difficult to maximize. In this thesis, we propose an exact inference via Markov Chain Monte Carlo techniques and an approximate inference method via Taylor series expansion. In next section, we describe the Monte Carlo EM algorithm in details, which implements the exact inference method. The approximate inference method will be illustrated in Chapter 4.

3.3 Monte Carlo EM Algorithm

The EM algorithm (Dempster, Laid, 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 effect models. 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 a M-step that updates the parameter estimates by maximizing the expectation of the conditional log-likelihood. This iterative computation between the E-step and M-step till convergence leads to the MLEs.

If we treat $(\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i, \boldsymbol{r}_i) \equiv (\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i)$ as the "complete data",

the complete data density for individual i is given by

$$\begin{aligned} & f(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i | \boldsymbol{\beta}, D, \boldsymbol{\psi}, \boldsymbol{\alpha}) \\ &= f(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i) f(\boldsymbol{x}_i | \boldsymbol{\alpha}) f(\boldsymbol{b}_i | D) f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}). \end{aligned}$$

This leads to the complete data log-likelihood

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{N} l_i(\boldsymbol{\gamma}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i) \\ &= \sum_{i=1}^{N} \Big[\log\{f(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)\} + \log\{f(\boldsymbol{x}_i | \boldsymbol{\alpha})\} \\ &\quad + \log\{f(\boldsymbol{b}_i | D)\} + \log\{f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi})\} \Big], \end{aligned} \tag{3.6}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, D)$ and $l_i(\boldsymbol{\gamma}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i)$ is the contribution to the complete data log-likelihood from the ith individual. Note that we are mainly interested in estimating the parameters $(\boldsymbol{\beta}, D)$, and treat $(\boldsymbol{\alpha}, \boldsymbol{\psi})$ as nuisance parameters.

Ibrahim *et al.* (2001) proposed a Monte Carlo EM algorithm for estimating parameters in GLMMs with informative dropout without missing covariates. Here we extend their method to GLMMs with informative dropout *and* missing covariates for simultaneous inference.

3.3.1 E-step

Let $\boldsymbol{\gamma}^{(t)}$ be the current parameter estimates. Then the conditional expectation of the complete-data log-likelihood given the observed data for individual i at the $(t+1)$st EM

iteration is given by

$$
\begin{aligned}
& Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \\
=& E\left(l_i(\boldsymbol{\gamma}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i)|\mathbf{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}\right) \\
=& \int\int\int \Big[\log\{f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)\} + \log\{f(\boldsymbol{x}_i|\boldsymbol{\alpha})\} \\
& + \log\{f(\boldsymbol{b}_i|D)\} + \log\{f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi})\}\Big] \\
& \times f(\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) d\boldsymbol{b}_i d\boldsymbol{y}_{mis,i} d\boldsymbol{x}_{mis,i}, \\
\equiv& I_1 + I_2 + I_3 + I_4.
\end{aligned}
\tag{3.7}
$$

In general, the above integration is intractable and does not have a closed form expression. However, this integral can be evaluated by using Monte Carlo approximations (Wei and Tanner (1990)). Specifically, a sample of size m_i $\{(\boldsymbol{y}_{mis,i}^{(1)}, \boldsymbol{x}_{mis,i}^{(1)}, \boldsymbol{b}_i^{(1)}), \cdots, (\boldsymbol{y}_{mis,i}^{(m_i)}, \boldsymbol{x}_{mis,i}^{(m_i)}, \boldsymbol{b}_i^{(m_i)})\}$ can be drawn from $f(\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ via Gibbs sampler along with the adaptive rejection algorithm (Gilks and Wild, 1992). Then we may approximate $Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ by

$$
\begin{aligned}
Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \approx & \frac{1}{m_i}\sum_{j=1}^{m_i} \log\{f(\mathbf{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}|\boldsymbol{\beta}, \boldsymbol{b}_i^{(j)}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)})\} \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \log\{f(\boldsymbol{x}_{obs,i}, x_{mis,i}^{(j)}|\boldsymbol{\alpha})\} \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \log\{f(\boldsymbol{b}_i^{(j)}|D)\} \\
& + \frac{1}{m_i}\sum_{j=1}^{m_i} \log\{f(\boldsymbol{r}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}, \boldsymbol{\psi})\}.
\end{aligned}
\tag{3.8}
$$

For simplicity, we may take m_i constant in each iteration. However, increasing m_i 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

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) =& \sum_{i=1}^{N} Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \\ =& \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}|\boldsymbol{\beta}, \boldsymbol{b}_i^{(j)}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)})\} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\mathbf{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)}|\boldsymbol{\alpha})\} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{b}_i^{(j)}|D)\} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{r}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}, \boldsymbol{\psi})\} \\ \equiv& Q^{(1)}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha}|\boldsymbol{\gamma}^{(t)}) + Q^{(3)}(D|\boldsymbol{\gamma}^{(t)}) + Q^{(4)}(\boldsymbol{\psi}|\boldsymbol{\gamma}^{(t)}). \end{aligned} \tag{3.9}$$

3.3.2 M-step

We can obtain the updated estimates $\boldsymbol{\gamma}^{(t+1)}$ at the $(t+1)$st iteration by maximizing $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$. Assuming that the parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, D and $\boldsymbol{\psi}$ are all distinct, we can update $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, D and $\boldsymbol{\psi}$ by maximizing $Q^{(1)}$, $Q^{(2)}$, $Q^{(3)}$ and $Q^{(4)}$ separately at the M-step. The maximizer $\boldsymbol{\beta}^{(t+1)}$ for $Q^{(1)}$ may be computed via iteratively re-weighted least squares where the missing values are replaced by their simulated values $\{\boldsymbol{y}_{mis,i}^{(j)}, \boldsymbol{x}_{mis,i}^{(j)}, \boldsymbol{b}_i^{(j)}\}$.

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} =& \arg\max_{\boldsymbol{\beta}}\{Q^{(1)}(\boldsymbol{\beta}, |\boldsymbol{\gamma}^{(t)})\} \\ =& \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i}\{f(\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}|\boldsymbol{\beta}, \boldsymbol{b}_i^{(j)}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)})\}. \end{aligned} \tag{3.10}$$

The maximizer $D^{(t+1)}$ for $Q^{(3)}$ can be written as follows:

$$\begin{aligned} D^{(t+1)} =& \arg\max_{D}\{Q^{(3)}(D, |\boldsymbol{\gamma}^{(t)})\} \\ =& \arg\max_{D} \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{b}_i^{(j)}|D)\} \end{aligned} \tag{3.11}$$

To update $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$, one can use standard methods for commonly used models such as multivariate normal models and logistic regression models.

$$\begin{aligned}\boldsymbol{\alpha}^{(t+1)} =& \arg\max_{\boldsymbol{\alpha}}\{Q^{(2)}(\boldsymbol{\alpha}, |\boldsymbol{\gamma}^{(t)})\} \\ =& \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)}|\boldsymbol{\alpha})\},\end{aligned} \tag{3.12}$$

$$\begin{aligned}\boldsymbol{\psi}^{(t+1)} =& \arg\max_{\boldsymbol{\psi}}\{Q^{(4)}(\boldsymbol{\psi}, |\boldsymbol{\gamma}^{(t)})\} \\ =& \arg\max_{\boldsymbol{\psi}} \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{r}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}, \boldsymbol{\psi})\}.\end{aligned} \tag{3.13}$$

To obtain the MLEs $\hat{\boldsymbol{\gamma}}$, we may start with any reasonable values for $\boldsymbol{\gamma}$, which can be obtained by the complete-case method or other naive methods , and then iterate between E-step and M-step until convergence is reached.

3.3.3 Variance Estimation

The asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be obtained by the method of Louis (1982). Specifically, note that the observed information matrix equals the conditional expected complete information minus the missing information, that is,

$$I_{obs}(\hat{\boldsymbol{\gamma}}) = I_{com}(\hat{\boldsymbol{\gamma}}) - I_{mis|obs}(\hat{\boldsymbol{\gamma}}). \tag{3.14}$$

Let

$$S_{ij}(\boldsymbol{\gamma}) = \frac{\partial l_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}},$$

$$\dot{Q}(\boldsymbol{\gamma}|\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N} \dot{Q}_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}) = \sum_{i=1}^{N}\sum_{k=1}^{m_i} \frac{1}{m_i} S_{ij}(\boldsymbol{\gamma}),$$

and

$$\ddot{Q}(\boldsymbol{\gamma}|\hat{\boldsymbol{\gamma}}) = \frac{\partial^2 Q(\boldsymbol{\gamma}|\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} = \sum_{i=1}^{N}\sum_{k=1}^{m_i} \frac{1}{m_i} \frac{\partial S_{ij}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}.$$

Since $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi}$ and D are distinct, matrices $\ddot{Q}(\boldsymbol{\gamma}|\hat{\boldsymbol{\gamma}})$, $\dot{Q}(\boldsymbol{\gamma}|\hat{\boldsymbol{\gamma}})$ and $S_{ij}(\hat{\boldsymbol{\gamma}})$ are block diagonal. Then, based on (3.14), the asymptotic observed information matrix gives

$$I_{obs}(\hat{\boldsymbol{\gamma}}) = -\ddot{Q}(\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{\gamma}}) - \left\{ \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{1}{m_i} S_{ij}(\hat{\boldsymbol{\gamma}}) S_{ij}^T(\hat{\boldsymbol{\gamma}}) - \sum_{i=1}^{N} \dot{Q}_i(\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{\gamma}}) \dot{Q}_i^T(\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{\gamma}}) \right\}. \tag{3.15}$$

Thus, the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be approximated by

$$\text{cov}(\hat{\boldsymbol{\gamma}}) = I_{obs}^{-1}(\hat{\boldsymbol{\gamma}}). \tag{3.16}$$

3.4 Sampling Methods

3.4.1 Gibbs Sampler

As we can see from the preceding section, generating samples from the conditional distribution $f(\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ is crucial for implementing the E-step of the Monte-Carlo EM algorithm. Gibbs sampler is a popular method to generate samples from a complicated multidimensional distribution by sampling from each of the full conditional distributions in turn. Here we use the Gibbs sampler to simulate the "missing values" as follows. Set initial values $\boldsymbol{y}_{mis,i}^{(0)}$, $\boldsymbol{x}_{mis,i}^{(0)}$ and $\boldsymbol{b}_i^{(0)}$. Supposed that the current generated values are $\boldsymbol{y}_{mis,i}^{(k)}$, $\boldsymbol{x}_{mis,i}^{(k)}$ and $\boldsymbol{b}_i^{(k)}$.

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

$$f(\boldsymbol{y}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(k)}, \boldsymbol{b}_i^{(k)}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}),$$

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

$$f(\boldsymbol{x}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(k+1)}, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i^{(k)}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}), \text{ and}$$

Step 3. draw a sample $\{\boldsymbol{b}_i^{(k+1)}\}$ from $f(\boldsymbol{b}_i | \boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(k+1)}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(k+1)}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$.

After a sufficiently large burn-in of d iterations, the sampled values will achieve a steady state, that is, $\{(\boldsymbol{y}_{mis,i}^{(k+1)}, \boldsymbol{x}_{mis,i}^{(k+1)}, \boldsymbol{b}_i^{(k+1)}), k = d+1, \cdots, B\}$ can be treated as samples from the multidimensional density function $f(\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$.

3.4.2 Adaptive Rejection Algorithm

Gilk and Wilks (1992) proposed an adaptive rejection algorithm for effectively sampling any univariate log-concave density function. In the current situation, we can write

$$f(\boldsymbol{y}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) \propto f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)}),$$

$$f(\boldsymbol{x}_{mis,i} | \boldsymbol{y}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) \propto f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}) f(\boldsymbol{x}_i | \boldsymbol{\alpha}^{(t)}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)}),$$

$$f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) \propto f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)}) f(\boldsymbol{b}_i | D^{(t)}).$$

Density functions $f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})$ and $f(\boldsymbol{b}_i | D^{(t)})$ often come from the exponential family, and thus are log-concave in each component of $\boldsymbol{y}_{mis,i}$, $\boldsymbol{x}_{mis,i}$, and $\boldsymbol{b}_i$ respectively. If $f(\boldsymbol{r}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)})$ is log-concave in each component of $\boldsymbol{y}_{mis,i}$ and $f(\boldsymbol{x}_i | D^{(t)})$ is log-concave in each component of $\boldsymbol{x}_{mis,i}$, then the products of log-concave functions, $f(\boldsymbol{y}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$, $f(\boldsymbol{x}_{mis,i} | \boldsymbol{y}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$, and $f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$, are log-concave. So we can use the adaptive rejection algorithm to generate samples from $f(\boldsymbol{y}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$, $f(\boldsymbol{x}_{mis,i} | \boldsymbol{y}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ and $f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ respectively. Note that the adaptive rejection sampling can only be applied to the univariate case, but $\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}$ and $\boldsymbol{b}_i$ are often multidimensional. Thus, to implement the Gibbs sampler described earlier, we need to modify the sampling scheme to incorporate multidimensional variables, as described below.

For example, suppose that $\boldsymbol{y}_{mis,i}$ is a multivariate of dimension l, that is, $\boldsymbol{y}_{mis,i} = (y_{mis,1,i}, \cdots, y_{mis,l,i})^T$. Since the function $f(\boldsymbol{y}_{mis,i} | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ is log-concave with

respect to each component of $\boldsymbol{y}_{mis,i}$, and

$$f(y_{mis,\tilde{k},i}|\boldsymbol{y}_{obs,i}, \{y_{mis,h,i}, h \neq \tilde{k}\}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) \propto f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}),$$

where $\tilde{k} = 1, \cdots, l$, the function $f(y_{mis,\tilde{k},i}|\boldsymbol{y}_{obs,i}, \{y_{mis,h,i}, h \neq \tilde{k}\}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ is log-concave with respect to $y_{mis,\tilde{k},i}$. Thus, another Gibbs sampler, together with the adaptive rejection sampling, can be used for generating a sample from $f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Specifically, we can proceed as follows.

Step 1. use the adaptive rejection sampling method to generate $y_{mis,1,i}^{(k+1)}$ from

$$f(y_{mis,1,i}|\boldsymbol{y}_{obs,i}, \{y_{mis,h,i}^{(k)}, h \geq 2\}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)});$$

Step 2. use the adaptive rejection sampling method to generate $y_{mis,2,i}^{(k+1)}$ from

$$f(y_{mis,2,i}|\boldsymbol{y}_{obs,i}, \{y_{mis,1,i}^{(k+1)}, y_{mis,h,i}^{(k)}, h \geq 3\}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)});$$

.....

Step l. use the adaptive rejection sampling method to generate $y_{mis,l,i}^{(k+1)}$ from

$$f(y_{mis,l,i}|\boldsymbol{y}_{obs,i}, \{y_{mis,h,i}^{(k+1)}, h \neq l_i\}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}).$$

After a burn-in period, the vector $\left\{y_{mis,1,i}^{(k+1)}, \cdots, y_{mis,l}^{(k+1)}\right\}$ may be treated as a sample from $f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Samples from $f(\boldsymbol{x}_{mis,i}|\boldsymbol{y}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$, and $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ can be obtained in a similar way.

3.4.3 Rejection Sampling

When the density functions do not satisfy the log-concave property, the usual rejection sampling method can be used for generating the desired samples. For example, suppose that we want to sample from $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, r_i, \boldsymbol{\gamma}^{(t)})$, which can be written as $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) = \mathrm{c}f(\boldsymbol{b}_i|D^{(t)})f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})$, where c is a constant, then the usual rejection sampling method can be described as follows.

Step 1. generate a random value $\boldsymbol{b}^*$ from $f(\boldsymbol{b}_i|D^{(t)})$, and draw a sample U from the Uniform(0,1),

Step 2. accept $\boldsymbol{b}^*$ as a sample from $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ if $U \leq f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})/\tau$ where $\tau = \sup_{\boldsymbol{u}}\{f(\boldsymbol{y}_i|\boldsymbol{u}, \boldsymbol{x}_i, \boldsymbol{\beta}^{(t)})\}$. Otherwise, reject $\boldsymbol{b}^*$ and go to step 1.

Samples from $f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ and $f(\boldsymbol{x}_{mis,i}|\boldsymbol{y}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ can be obtained in a similar way.

3.4.4 Sampling Method for Binary Variables

If the missing variables are binary variables, then we may use an easier way to generate the desired sample. Here, we take the missing response $\boldsymbol{y}_{mis,i}$ as an example. Suppose that the response is binary and we want to draw samples from $f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. For simplicity, here we assume that $\boldsymbol{y}_{mis,i}$ is univariate. The corresponding sampling procedure is described as follows.

Step 1. draw a sample U from the Uniform(0,1),

Step 2. take 0 as a sample from $f(\boldsymbol{y}_{mis,i}|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{b}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$ if $U \leq f(0|\boldsymbol{y}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Otherwise, take 1 as a sample.

3.5 PX-EM Algorithm

Although the EM algorithm is a very popular tool for estimation due to its easy implementation and stable convergence, it may converge quite slowly in some applications such as ours. To speed up the convergence, many acceleration methods have been proposed (e.g. Liu and Rubin, 1994a, Meng and Van Dyk, 1997), A particularly useful method is

the PX-EM algorithm (Liu *et al*, 1998) , which speeds up the EM algorithm by introducing additional working parameters to the model. For the PX-EM, the E-step is usually the same as the standard EM, while the M-step needs to maximize the expected log-likelihood over the original parameters *and* the working parameters. Thus, the PX-EM algorithm can be obtained by simple modification of the standard EM. Liu *et al* (1998) showed that the PX-EM algorithm may dramatically accelerate the rate of convergence without loss of the simplicity and stability of the standard EM. Next, we show how to apply the PX-EM algorithm to GLMMs with informative dropouts and missing covariates. We may expand the GLMM (3.1)-(3.2) by introducing additional working parameters as follows.

$$f(y_{ij}|\boldsymbol{\beta}^*, \boldsymbol{b}_i) = \exp\left[\{y_{ij}\theta_{ij}(\mu_{ij}) - \varphi(\theta_{ij}(\mu_{ij}))\}/a(\phi) + c(y_{ij}, \phi)\right], \tag{3.17}$$

$$\begin{aligned} \eta_{ij} &= g(\mu_{ij}) = A_{ij}^T\boldsymbol{\beta}^* + \boldsymbol{z}_{ij}^T\Lambda\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\overset{i.i.d}{\sim} N(0, D^*), \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, N, \end{aligned} \tag{3.18}$$

where Λ is a $q \times q$ matrix, called working parameters. The PX-EM algorithm is the standard EM applied to the expanded models (3.17)-(3.18) rather than the original models (3.1)-(3.2). Specifically, the E-step and the M-step are described as follows.

E-step: Let $\boldsymbol{\Theta}^{(t)} = (\boldsymbol{\beta}^{*(t)}, \boldsymbol{\alpha}^{*(t)}, \boldsymbol{\psi}^{*(t)}, D^{*(t)}, \Lambda^{(t)}) \equiv (\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\psi}^{(t)}, D^{(t)}, I_{q\times q})$ be the current estimates of the expanded parameters. The E-step of PX-EM is obtained by simply adding the working parameters Λ to $Q(.)$, i.e, the E-step of the standard EM in Section 3.2.1. Then the conditional expectation of the complete-data log-likelihood given the observed data for the model (3.17)-(3.18) can be written as

$$
\begin{aligned}
Q^*(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &= \sum_{i=1}^{N} Q_i^*(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}|\boldsymbol{\beta}^*, \Lambda \boldsymbol{b}_i^{(j)}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)})\} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\mathbf{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(j)}|\boldsymbol{\alpha}^*)\} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{b}_i^{(j)}|D^*)\} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{1}{m_i} \log\{f(\boldsymbol{r}_i|\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i}^{(j)}, \boldsymbol{\psi}^*)\} \\
&\equiv Q^{*(1)}(\boldsymbol{\beta}^*, \Lambda|\boldsymbol{\Theta}^{(t)}) + Q^{*(2)}(\boldsymbol{\alpha}^*|\boldsymbol{\Theta}^{(t)}) + Q^{*(3)}(D^*|\boldsymbol{\Theta}^{(t)}) + Q^{*(4)}(\boldsymbol{\psi}^*|\boldsymbol{\Theta}^{(t)}).
\end{aligned}
\tag{3.19}
$$

Obviously, everything in this E-step is the same as the E-step of the standard EM in Section 3.2.1, except the extra working parameters Λ in (3.18).

M-step: By the same standard maximization procedures as the M-step in Section 3.2.2, we maximize $Q^{*(1)}$, $Q^{*(2)}$, $Q^{*(3)}$ and $Q^{*(4)}$ separately to update the estimates of the expanded parameters to $\boldsymbol{\beta}^{*(t+1)}$, $\boldsymbol{\alpha}^{*(t+1)}$, $\boldsymbol{\psi}^{*(t+1)}$, $D^{*(t+1)}$ and $\Lambda^{(t+1)}$. The only difference in this step between the PX-EM and the EM is that the PX-EM maximizes $Q^{*(1)}$ over $\boldsymbol{\beta}^*$ *and* Λ, while the EM does this only over $\boldsymbol{\beta}$. The reduction to the original parameters in the models (3.1) – (3.2) gives

$$
\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{*(t+1)},\ \boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{*(t+1)},\ \boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{*(t+1)},\ D^{(t+1)} = \Lambda^{(t+1)} D^{*(t+1)} \Lambda^{(t+1)^T}.
$$

This iterative calculation between the E-step and M-step until convergence leads to the MLEs of parameters in the original models (3.1) – (3.2).

3.6 Convergence

When carrying out the Monte Carlo EM algorithm, Monte Carlo samples for the "missing data" are drawn at each iteration to approximate true values. Consequently, Monte Carlo errors are introduced. One way to reduce the Monte Carlo errors is to increase the Monte Carlo sample size m_i. However, the computation is intensive for a large m_i. Because the estimate $\boldsymbol{\gamma}^{(t)}$ in the initial EM steps is often far from the true values of the parameters, Monte Carlo samples of a large size m_i may be wasted. Thus, we usually use a small m_i at initial iterations, and then increase m_i with the iteration, as suggested by Booth and Hobert (1999).

After an initial burn-in period, the Gibbs sampler converges to a stationary state and thus produces draws from the conditional density function $f(\boldsymbol{y}_{mis,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{b}_i | \boldsymbol{y}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Obviously, the determination of the burn-in period is very important. We way use common diagnostic methods to determine the burn-in period, such as time series plots.

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 next chapter, we propose an approximate inference method which may avoid these convergence difficulties and may be much more efficient.

Chapter 4

Approximate Inference for GLMMs with Informative Dropout and Missing Covariates

4.1 Introduction

In Chapter 3, we have described the exact inference method implemented by the Monte Carlo EM algorithm. However, the exact method may be computationally intensive and may even offer potentially computational difficulties such as slow or non-convergence. Moreover, when the dimension of random effects is not small, sampling random effects may result in inefficient and computationally unstable Gibbs samplers, which may lead to a high degree of autocorrelation and a lack of convergence. In the presence of missing response and missing covariates, these problems become more serious. In this chapter, we propose an approximate inference method which is not only much more efficient, but also avoids potential computational difficulties. This approximate method is obtained by Taylor series expansion and it avoids sampling the random effects in the E-step by

integrating them out. Pinheiro *et al.* (2001), in a different context, have showed that the convergence rate of the EM algorithm can be greatly improved by integrating out the random effects in the E-step.

The outline of this chapter is as follows. In Section 4.2, we present the approximate inference method for GLMMs without missing values, and then extend this method to GLMMs with informative dropout and missing covariates implemented by the Monte Carlo EM algorithm in Section 4.3. In Section 4.4, we briefly describe the sampling methods used in Section 4.3. We conclude this chapter with a discussion about the PX-EM algorithm, an extension of the standard EM algorithm.

4.2 Approximate Inference without Missing Values

As described in Chapter 3, a GLMM is written as

$$f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = \exp\left[\{y_{ij}\theta_{ij}(\mu_{ij}) - \varphi(\theta_{ij}(\mu_{ij}))\}/a(\phi) + c(y_{ij}, \phi)\right], \tag{4.1}$$

$$\begin{aligned} \eta_{ij} &= g(\mu_{ij}) = A_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\overset{i.i.d}{\sim} N(0, D), \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, N, \end{aligned} \tag{4.2}$$

where the notation is the same as (3.1)-(3.2) in Chapter 3. Denoting the observation vector as $\boldsymbol{y}_i = (y_{i1} \cdots, y_{in_i})^T$, and design matrices as $A_i = (A_{i1}, \cdots, A_{ik})^T$ and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \cdots, \boldsymbol{z}_{iq})^T$, then the conditional mean and covariance of $\boldsymbol{y}_i$ satisfy $E(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{\mu}_i = g^{-1}(A_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b}_i)$ and $\text{cov}(\boldsymbol{y}_i|\boldsymbol{b}_i) = C_i^{\mu} = \text{diag}\{V(\mu_{ij})/a(\phi), j = 1, \cdots, n_i\}$ respectively.

The above GLMM yields a marginal log-likelihood function by integrating out the

random effects

$$
\begin{aligned}
& l(\boldsymbol{\beta}, D|\boldsymbol{y}_1 \cdots, \boldsymbol{y}_N) \\
&= \log \left\{ \int \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i \right\} \\
&= \log \left[\int \exp \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} \Big\{ \log(f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)) + \log(f(\boldsymbol{b}_i|D)) \Big\} \right\} d\boldsymbol{b}_1 \cdots d\boldsymbol{b}_N \right].
\end{aligned}
\tag{4.3}
$$

To estimate the parameters in the GLMM, we need to maximize the log-likelihood function (4.3). However, in most cases, the integral in (4.3) is intractable. Evaluating the integral in (4.3) by Monte Carlo methods for the exact inference method may offer potential computational problems, as noted earlier. Here, we consider a much more efficient approximate inference method based on Taylor series expansion. The following approximation is based on a second-order Taylor series expansion about the current parameter estimates $\hat{\boldsymbol{\theta}}$, which is equivalent to the Laplace's approximation (see [20] [26]),

$$
\int e^{k(\boldsymbol{\theta})} d\boldsymbol{\theta} \approx (2\pi)^{\frac{q}{2}} \left| - \frac{\partial^2 k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right|^{-\frac{1}{2}} e^{k(\hat{\boldsymbol{\theta}})}, \tag{4.4}
$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $k(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is a maximizer to $k(\boldsymbol{\theta})$. Applying the Laplace's approximation to the log-likelihood (4.3) yields

$$
l(\boldsymbol{\beta}, D|\boldsymbol{y}_1 \cdots, \boldsymbol{y}_N) \approx \frac{mq}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log \left| \frac{\partial^2 k_i(\boldsymbol{b}_i)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T} \Big|_{\boldsymbol{b}_i = \boldsymbol{b}_i^0} \right| + \sum_{i=1}^{N} k_i(\boldsymbol{b}_i^0), \tag{4.5}
$$

where $k_i(\boldsymbol{b}_i) = \sum_{j=1}^{n_i} \Big\{ \log(f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)) + \log(f(\boldsymbol{b}_i|D)) \Big\}$, and $\boldsymbol{b}_i^0$ maximizes the function $k_i(\boldsymbol{b}_i)$.

Maximizing the approximate log-likelihood function (4.5) with respect to $\boldsymbol{\beta}$ and D, and maximizing $k_i(\boldsymbol{b}_i)$ with respect to $\boldsymbol{b}_i$ are equivalent to jointly solving the following score equations (see [20] [26]),

$$
\begin{cases}
\sum_{i=1}^{N} {A_i}^T {W_i}^{-1} B_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \\
a(\phi) {\boldsymbol{z}_i}^T {W_i}^{-1} B_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = D^{-1} \boldsymbol{b}_i, \quad i = 1, \cdots, N,
\end{cases}
\tag{4.6}
$$

where B_i is a $n_i \times n_i$ diagonal matrix with diagonal terms $\partial g(\mu_{ij})/\partial \mu_{ij}$ and $W_i = B_i C_i^{\mu} B_i$. It can be shown that the solution to (4.6) via Fisher scoring is equivalent to iteratively solving the following linear equations (see [20] [26]),

$$\begin{pmatrix} A^T W^{-1} A & A^T W^{-1} Z \\ Z^T W^{-1} A & Z^T W^{-1} Z + D_d{}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} A^T W^{-1} \tilde{\boldsymbol{y}} \\ Z^T W^{-1} \tilde{\boldsymbol{y}} \end{pmatrix}, \tag{4.7}$$

where $\tilde{\boldsymbol{y}}_i = A_i{}^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_i^T \hat{\boldsymbol{b}}_i + B_i(\boldsymbol{y}_i - g^{-1}(A_i^T \hat{\boldsymbol{\beta}} + \boldsymbol{z}_i^T \hat{\boldsymbol{b}}_i))$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$ are the current estimates. Row vectors are $\tilde{\boldsymbol{y}}^T = (\tilde{\boldsymbol{y}}_1^T, \cdots, \tilde{\boldsymbol{y}}_N^T)$ and $\boldsymbol{b}^T = (\boldsymbol{b}_1^T, \cdots, \boldsymbol{b}_N^T)$, and matrices are $A^T = (A_1^T \cdots, A_N^T)$, $Z^T = \text{diag}(\boldsymbol{z}_1^T, \cdots, \boldsymbol{z}_N^T)$, $D_d = \text{diag}\{D, \cdots, D\}$ and $W = \text{diag}\{W_1, \cdots, W_N\}$. Solving the linear equations (4.7) is equivalent to solving the following linear mixed effect (LME) model (see [3][26]),

$$\tilde{\boldsymbol{y}}_i = A_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, N, \tag{4.8}$$

where $\boldsymbol{\epsilon}_i$'s are independent with a normal distribution $N(0, W_i)$, $\boldsymbol{b}_i$'s are independent and have an identical normal distribution $N(0, D)$, and $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent. From (4.8) we can derive

$$\boldsymbol{b}_i | \tilde{\boldsymbol{y}}_i \sim N\left(\Sigma_i \boldsymbol{z}_i W_i^{-1} (\tilde{\boldsymbol{y}}_i - A_i \boldsymbol{\beta}), \Sigma_i\right), \tag{4.9}$$

where $\Sigma_i = (\boldsymbol{z}_i W_i{}^{-1} \boldsymbol{z}_i^T + D^{-1})^{-1}$, and

$$\tilde{\boldsymbol{y}}_i \sim N\left(A_i^T \boldsymbol{\beta}, \boldsymbol{z}_i^T D \boldsymbol{z}_i + W_i\right). \tag{4.10}$$

In summary, approximate estimates for GLMMs can be obtained by iteratively solving the liner mixed effect model (4.8), which can be easily handled by standard software packages such as Splus and SAS.

4.3 Approximate Inference with Missing Values

In the previous section, we discuss an approximate inference method for the GLMM without missing values. However, in our GLMM (4.1) – (4.2), the response $\boldsymbol{y}_i$ is non-ignorably missing and the covariate $\boldsymbol{x}_i$ is ignorably missing. In this section we consider a similar method for GLMMs with missing values. Note that missing values in GLMM (4.1) – (4.2) correspond to missing responses in $\tilde{\boldsymbol{y}}_i$ and missing covariates in $\boldsymbol{x}_i$ in the LME model (4.8) respectively. Here we write $\tilde{\boldsymbol{y}}_i = (\tilde{\boldsymbol{y}}_{obs,i}, \tilde{\boldsymbol{y}}_{mis,i})$, where $\tilde{\boldsymbol{y}}_{obs,i}$ contains the observed components of $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}_{mis,i}$ contains the missing components of $\tilde{\boldsymbol{y}}_i$. Note that $\tilde{\boldsymbol{y}}_{obs,i}$ and $\tilde{\boldsymbol{y}}_{mis,i}$ are appropriate functions of the missing and observed components of $\boldsymbol{y}_i$ respectively. So the missing response indicator for $\tilde{\boldsymbol{y}}_i$ is the same as the missing response indicator for $\boldsymbol{y}_i$. For LME models with non-ignorable missing responses, Ibrahim *et al.* (2001) derived a much more efficient Monte Carlo EM algorithm by integrating out the random effects in the E-step. Here we extend their approaches to the GLMM with informative dropout and missing covariates by iteratively solving the LME model (4.8) with non-ignorable missing responses and ignorable missing covariates. Since sampling random effects is avoided in the E-step, the rate of convergence of the EM algorithm may be greatly improved. The E-step and M-step are described in details as follows.

E-step: Let $\boldsymbol{\gamma}^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\psi}^{(t)}, D^{(t)})$ be the current parameter estimates. The response in the LME model (4.8) can be written as $\tilde{\boldsymbol{y}}_i = A_i^T\boldsymbol{\beta}^{(t)} + \boldsymbol{z}_i^T\boldsymbol{b}_i^{(t)} + B_i^{(t)}(\boldsymbol{y}_i - g^{-1}(A_i^T\boldsymbol{\beta}^{(t)}+\boldsymbol{z}_i^T\boldsymbol{b}_i^{(t)}))$ where $\boldsymbol{b}_i^{(t)} = \Sigma_i^{(t)}\boldsymbol{z}_i{W_i^{(t)}}^{-1}(\tilde{\boldsymbol{y}}_i - A_i^T\boldsymbol{\beta}^{(t)})$, $B_i^{(t)} = B_i|_{\mu_{ij}=g^{-1}(A_{ij}^T\boldsymbol{\beta}^{(t)}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i^{(t)})}$ and $W_i^{(t)} = B_i^{(t)}C_i^{\mu}B_i^{(t)}\Big|_{\mu_{ij}=g^{-1}(A_{ij}^T\boldsymbol{\beta}^{(t)}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i^{(t)})}$.

As in the previous section, the contribution of individual i in the $(t+1)$st iteration

is given by

$$
\begin{aligned}
& Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \\
\approx & \int\int\int \Big\{ \log\big(f(\tilde{\boldsymbol{y}}_i|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_i)\big) + \log\big(f(\boldsymbol{b}_i|D)\big) \\
& \qquad + \log\big(f(\boldsymbol{x}_i|\boldsymbol{\alpha})\big) + \log\big(f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i,\boldsymbol{x}_i,\boldsymbol{\psi})\big)\Big\} \\
& \qquad \times f(\tilde{\boldsymbol{y}}_{mis,i},\boldsymbol{x}_{mis,i},\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_{obs,i},\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)})d\boldsymbol{b}_i d\tilde{\boldsymbol{y}}_{mis,i} d\boldsymbol{x}_{mis,i} \\
= & \int\int\Big[\int \big\{ \log(f(\tilde{\boldsymbol{y}}_i|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_i)) + \log(f(\boldsymbol{b}_i|D)) \\
& \qquad + \log(f(\boldsymbol{x}_i|\boldsymbol{\alpha})) + \log(f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i,\boldsymbol{x}_i,\boldsymbol{\psi}))\big\} \\
& \qquad f(\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_i,\boldsymbol{x}_i,\boldsymbol{\gamma}^{(t)})d\boldsymbol{b}_i\Big] \\
& \qquad f(\tilde{\boldsymbol{y}}_{mis,i},\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i},\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)})d\tilde{\boldsymbol{y}}_{mis,i}d\boldsymbol{x}_{mis,i} \\
\equiv & \; I_1 + I_2 + I_3 + I_4, \qquad (4.11)
\end{aligned}
$$

where $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_i)$ is the normal density with mean $A_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^T\boldsymbol{b}_i$ and covariance $W_i^{(t)}$. Equation (4.9) implies $\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_i,\boldsymbol{x}_i,\boldsymbol{\gamma}^{(t)} \sim N(\boldsymbol{b}_i^{(t)},\Sigma_i^{(t)})$, where $\Sigma_i^{(t)} = (\boldsymbol{z}_i {W_i^{(t)}}^{-1}\boldsymbol{z}_i^T + {D^{(t)}}^{-1})^{-1}$.

After some algebra, we can integrate out the random effects $\boldsymbol{b}_i$ from (4.11) and obtain the following results

$$
\begin{aligned}
I_1 = & -\frac{1}{2}\log|W_i^{(t)}| - \frac{1}{2}\int\int\Big\{\int(\tilde{\boldsymbol{y}}_i - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{b}_i)^T {W_i^{(t)}}^{-1}(\tilde{\boldsymbol{y}}_i - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{b}_i) \\
& f(\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_i,\boldsymbol{x}_i,\boldsymbol{\gamma}^{(t)})d\boldsymbol{b}_i\Big\} f(\tilde{\boldsymbol{y}}_{mis,i},\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i},\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)})d\tilde{\boldsymbol{y}}_{mis,i}d\boldsymbol{x}_{mis,i} \\
= & -\frac{1}{2}\log|W_i^{(t)}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{z}_i {W_i^{(t)}}^{-1}\boldsymbol{z}_i^T\Sigma_i^{(t)}) \\
& -\frac{1}{2}\int\int(\tilde{\boldsymbol{y}}_i - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{b}_i^{(t)})^T {W_i^{(t)}}^{-1}(\tilde{\boldsymbol{y}}_i - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{b}_i^{(t)}) \\
& \qquad f(\tilde{\boldsymbol{y}}_{mis,i},\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i},\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)})d\tilde{\boldsymbol{y}}_{mis,i}d\boldsymbol{x}_{mis,i}.
\end{aligned}
$$

$$I_2 = -\frac{1}{2}\log|D| - \frac{1}{2}\int\int\Big\{\int(\boldsymbol{b}_i^T D\boldsymbol{b}_i) f(\boldsymbol{b}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)})d\boldsymbol{b}_i\Big\}$$
$$f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})d\boldsymbol{y}_{mis,i}d\boldsymbol{x}_{mis,i}$$
$$= -\frac{1}{2}\log|D| - \frac{1}{2}\mathrm{tr}(D^{-1}\Sigma_i^{(t)})$$
$$-\frac{1}{2}\int\int(\boldsymbol{b}_i^{(t)T} D\boldsymbol{b}_i^{(t)}) f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})d\boldsymbol{y}_{mis,i}d\boldsymbol{x}_{mis,i}.$$

Since $f(\boldsymbol{x}_i|\boldsymbol{\alpha})$ and $f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, \mathbf{x}_i, \boldsymbol{\psi})$ do not depend on $\boldsymbol{b}_i$, we have

$$I_3 = \int\int f(\boldsymbol{x}_i|\boldsymbol{\alpha}) f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})d\boldsymbol{y}_{mis,i}d\boldsymbol{x}_{mis,i},$$

and

$$I_4 = \int\int f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i, \boldsymbol{\psi})(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})d\boldsymbol{y}_{mis,i}d\boldsymbol{x}_{mis,i}.$$

As we can see, integrals I_1, I_2, I_3 and I_4 do not involve the random effects $\boldsymbol{b}_i$. Thus we only need to generate random samples from $f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. This leads to a much more efficient EM than that for the exact method.

Suppose that $\{(\tilde{\boldsymbol{y}}_{mis,i}^{(1)}, \boldsymbol{x}_{mis,i}^{(1)}), \cdots, (\tilde{\boldsymbol{y}}_{mis,i}^{(m_i)}, \boldsymbol{x}_{mis,i}^{(m_i)})\}$ is a sample of size m_i generated from $f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Let $\boldsymbol{x}_i^{(k)} = (\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(k)})$, $\tilde{\boldsymbol{y}}_i^{(k)} = (\tilde{\boldsymbol{y}}_{obs,i}, \tilde{\boldsymbol{y}}_{mis,i}^{(k)})$. Then $\tilde{\boldsymbol{b}}_i^{(tk)} = \boldsymbol{\Sigma}_i^{(t)}\boldsymbol{z}_i W_i^{(t)^{-1}}(\tilde{\boldsymbol{y}}_i{}^{(k)} - A_i{}^T\boldsymbol{\beta}^{(t)})$, $k = 1, 2, \cdots, m_i$. Thus $Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ can be approximated as

$$\begin{aligned} Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \approx\ & \Big[-\frac{1}{2}\log|W_i^{(t)}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{z}_i W_i^{(t)^{-1}}\boldsymbol{z}_i^T\Sigma_i^{(t)}) \\ & -\frac{1}{2m_i}\sum_{k=1}^{m_i}(\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\tilde{\boldsymbol{b}}_i^{(tk)})^T W_i^{(t)^{-1}}(\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\tilde{\boldsymbol{b}}_i^{(tk)})\Big] \\ & +\Big[-\frac{1}{2}\log|D| - \frac{1}{2}\mathrm{tr}(D^{-1}\Sigma_i^{(t)}) - \frac{1}{2m_i}\sum_{k=1}^{m_i}(\tilde{\boldsymbol{b}}_i^{(tk)T} D^{-1}\tilde{\boldsymbol{b}}_i^{(tk)})\Big] \\ & +\Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{x}_i^{(k)}|\boldsymbol{\alpha})\Big] + \Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i^{(k)}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\psi})\Big]. \end{aligned} \tag{4.12}$$

Therefore, the E-step for all individuals at the $(t+1)$st iteration can be written as

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) &= \sum_{i=1}^{N} Q_i(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) \\ &\approx \sum_{i=1}^{N}\Big[-\frac{1}{2}\log|W_i^{(t)}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{z}_i {W_i^{(t)}}^{-1}\boldsymbol{z}_i^T\Sigma_i^t) \\ &\quad -\frac{1}{2m_i}\sum_{k=1}^{m_i}(\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\tilde{\boldsymbol{b}}_i^{(tk)})^T {W_i^{(t)}}^{-1}(\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\tilde{\boldsymbol{b}}_i^{(tk)})\Big] \\ &\quad +\sum_{i=1}^{N}\Big[-\frac{1}{2}\log|D| - \frac{1}{2}\mathrm{tr}(D^{-1}\Sigma_i^{(t)}) - \frac{1}{2m_i}\sum_{k=1}^{m_i}(\tilde{\boldsymbol{b}}_i^{(tk)T}D^{-1}\tilde{\boldsymbol{b}}_i^{(tk)})\Big] \\ &\quad +\sum_{i=1}^{N}\Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{x}_i^{(k)}|\boldsymbol{\alpha})\Big] + \sum_{i=1}^{N}\Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i^{(k)}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\psi})\Big] \\ &\equiv Q^{(1)}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(t)}) + Q^{(2)}(D|\boldsymbol{\gamma}^{(t)}) + Q^{(3)}(\boldsymbol{\alpha}|\boldsymbol{\gamma}^{(t)}) + Q^{(4)}(\boldsymbol{\psi}|\boldsymbol{\gamma}^{(t)}). \end{aligned} \tag{4.13}$$

M-step: Since the parameters in $\boldsymbol{\gamma}$ are all distinct, we can maximize $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ by maximizing $Q^{(1)}$, $Q^{(2)}$, $Q^{(3)}$ and $Q^{(4)}$ separately, leading to the updated estimate $\boldsymbol{\gamma}^{(t+1)}$. These maximizations can be accomplished by standard complete data optimization methods.

The covariance matrix for the parameter estimates, $\hat{\gamma}$, can again be obtained using Louis's method (1982), as in Chapter 3.

4.4 Strategies for Sampling the Missing Values

To implement the E-step of the EM algorithm, we need to generate random samples for the missing response $\tilde{\boldsymbol{y}}_{mis,i}$ and missing covariates $\boldsymbol{x}_{mis,i}$ from the joint density function $f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. As in Chapter 3, we can use Gibber sampler to draw the desired samples. The procedure is described as follows. Suppose that the current samples for missing values are $\tilde{\boldsymbol{y}}_{mis,i}^{(k)}$ and $\boldsymbol{x}_{mis,i}^{(k)}$.

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

$$f(\tilde{\boldsymbol{y}}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}^{(k)}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)});$$

Step 2. Draw a sample fror missing covarites $\{\boldsymbol{x}_{mis,i}^{(k+1)}\}$ from

$$f(\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \tilde{\boldsymbol{y}}_{mis,i}^{(k+1)}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}).$$

After a burn-in period, the sampled vaules $\{\tilde{\boldsymbol{y}}_{mis,i}^{(k+1)}, \boldsymbol{x}_{mis,i}^{(k+1)}\}$ can be treated as the true sample from the density function $f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)})$. Note that

$$f(\tilde{\boldsymbol{y}}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\gamma}^{(t)}) \propto f(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}) f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}),$$

where $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)})$ is a normal density function with mean $A_i^T\boldsymbol{\beta}^{(t)}$ and covariance $W_i^{(t)} + \boldsymbol{z}_i^T D^{(t)} \boldsymbol{z}_i$. If the density function $f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, x_i, \boldsymbol{\gamma}^{(t)})$ is log-concave, we can use the adaptive rejection algorithm to draw sample in Step 1; otherwise, we may consider the general rejection sampling method. Similarly, since

$$f(\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma}^{(t)}) \propto f(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}) f(\boldsymbol{x}_i|\boldsymbol{\gamma}^{(t)}) f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}),$$

as in Chapter 3, samples from $f(\boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma}^{(t)})$ in Step 2, can be obtained using the adaptive rejection algorithm or the rejection sampling method, depending on whether both $f(\boldsymbol{x}_i|\boldsymbol{\gamma}^{(t)})$ and $f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)})$ are log-concave.

4.5 PX-EM

The EM algorithm described in Section 4.3 may still be quite slow. To improve the speed of the EM algorithm, in this section we again consider the PX-EM for the approximate method, which is obtained by applying the standard EM algorithm to an expanded model. Specifically, we introduce a $q \times q$ working parameter matrix V to the LME model (4.8) and obtained the following expanded LME model

$$\tilde{\boldsymbol{y}}_i = A_i^T\boldsymbol{\beta}^* + \boldsymbol{z}_i^T V \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, N, \tag{4.14}$$

where the $\boldsymbol{\epsilon}_i$'s are independent error terms with an identical distribution $N(0, W_i)$, $\boldsymbol{b}_i \overset{i.i.d}{\sim} N(0, D^*)$, and $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent. Let $\Theta = (\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\psi}^*, D^*, \boldsymbol{V})$, where $\boldsymbol{\psi}^*$ is a vector of parameters for the dropout model and $\boldsymbol{\alpha}^*$ is a vector of parameters for the covariate model. Note that model (4.14) is reduced to model (4.8) when $V = I_{q\times q}$.

E-step: Let $\Theta^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\psi}^{(t)}, \boldsymbol{D}^{(t)}, I_{q\times q})$ be the current parameter estimates. Then in the E-step the conditional expectation of the complete-data log-likelihood given the observed data for the expanded model (4.14) can be written as

$$
\begin{aligned}
Q^*(\Theta|\boldsymbol{\Theta}^{(t)}) =& \sum_{i=1}^{N} Q_i^*(\Theta|\Theta^{(t)}) \\
\approx& \sum_{i=1}^{N} \Big[-\frac{1}{2}\log|W_i^{(t)}| - \frac{1}{2}\mathrm{tr}(V^T \boldsymbol{z}_i {W_i^{(t)}}^{-1} \boldsymbol{z}_i^T V \Sigma_i^{(t)}) \\
& - \frac{1}{2m_i}\sum_{k=1}^{m_i} (\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta}^* - \boldsymbol{z}_i^T V \tilde{\boldsymbol{b}}_i^{(tk)})^T {W_i^{(t)}}^{-1} (\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta}^* - \boldsymbol{z}_i^T V \tilde{\boldsymbol{b}}_i^{(tk)}) \Big] \\
& + \sum_{i=1}^{N} \Big[-\frac{1}{2}\log|D^*| - \frac{1}{2}\mathrm{tr}({D^*}^{-1}\Sigma_i^{(t)}) - \frac{1}{2m_i}\sum_{k=1}^{m_i} (\tilde{\boldsymbol{b}}_i^{(tk)T} {D^*}^{-1} \tilde{\boldsymbol{b}}_i^{(tk)}) \Big] \\
& + \sum_{i=1}^{N} \Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{x}_i^{(k)}|\boldsymbol{\alpha}^*) \Big] + \sum_{i=1}^{N} \Big[\frac{1}{m_i}\sum_{k=1}^{m_i} f(\boldsymbol{r}_i|\tilde{\boldsymbol{y}}_i^{(k)}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\psi}^*) \Big] \\
\equiv& Q^{*(1)}(\boldsymbol{\beta}^*, V|\Theta^{(t)}) + Q^{*(2)}(\boldsymbol{\alpha}^*|\Theta^{(t)}) + Q^{*(3)}(D^*|\Theta^{(t)}) + Q^{*(4)}(\boldsymbol{\psi}^*|\Theta^{(t)}),
\end{aligned}
\tag{4.15}
$$

where $W_i^{(t)} = \mathrm{diag}\left\{ V_i(\mu_{ij})(\partial g(\mu_{ij})/\partial\mu_{ij})^2/a(\phi)|_{\mu_{ij}=A_{ij}^T\boldsymbol{\beta}^{(t)}+\boldsymbol{z}_{ij}^T\tilde{\boldsymbol{b}}_i^{(t)}} \right\}$, $\Sigma_i^{(t)} = (\boldsymbol{z}_i {W_i^{(t)}}^{-1}\boldsymbol{z}_i^T + {D^{(t)}}^{-1})^{-1}$, $\boldsymbol{x}_i^{(k)} = (\boldsymbol{x}_{mis,i}^{(k)}, \boldsymbol{x}_{obs,i})$, $\tilde{\boldsymbol{y}}_i^{(k)} = (\tilde{\boldsymbol{y}}_{mis,i}^{(k)}, \tilde{\boldsymbol{y}}_{obs,i})$, $\tilde{\boldsymbol{b}}_i^{(tk)} = \Sigma_i^{(t)}\boldsymbol{z}_i {W_i^{(t)}}^{-1}(\tilde{\boldsymbol{y}}_i^{(k)} - A_i^T\boldsymbol{\beta}^{(t)})$. The sample of size m_i $\left\{ (\tilde{\boldsymbol{y}}_{mis,i}^{(1)}, \boldsymbol{x}_{mis,i}^{(1)}), \cdots, (\tilde{\boldsymbol{y}}_{mis,i}^{(m_i)}, \boldsymbol{x}_{mis,i}^{(m_i)}) \right\}$ is drawn from $f(\tilde{\boldsymbol{y}}_{mis,i}, \boldsymbol{x}_{mis,i}|\tilde{\boldsymbol{y}}_{obs,i}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \Theta^{(t)})$ by Gibbs sampler along with the adaptive rejection algorithm. Again, everything in this E-step is the same as the E-step of the standard EM in Section 4.3, except the extra working parameters Λ in (4.15).

M-step: In the M-step, we maximize $Q^{*(1)}$, $Q^{*(2)}$, $Q^{*(3)}$, and $Q^{*(4)}$ separately to

update the parameter estimates to $\boldsymbol{\beta}^{*(t+1)}$, $\boldsymbol{\alpha}^{*(t+1)}$, $\boldsymbol{\psi}^{*(t+1)}$, $\boldsymbol{D}^{*(t+1)}$ and $\boldsymbol{V}^{(t+1)}$. Then the estimates of original parameters are given by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{*(t+1)}, \boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{*(t+1)}, \boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{*(t+1)}, \boldsymbol{D}^{(t+1)} = \boldsymbol{V}^{(t+1)}\boldsymbol{D}^{*(t+1)}\boldsymbol{V}^{(t+1)^T}.$$

Chapter 5

Covariate Models and Dropout Models

5.1 Introduction

In the previous chapters, we have discussed how to estimate the parameters in GLMMs with informative dropout and missing covariates. As we note earlier, to provide a valid inference, we need to specify a dropout model for the missing response, and a covariate model for the time-independent covariates, and then incorporate them into our analyses. In this chapter, we describe how to specify these models. Sections 5.2 and 5.3 introduce dropout models and covariate models respectively. In Section 5.4, we discuss sensitivity analyses for the dropout model and covariate model.

5.2 Dropout Models

A dropout model is the distribution for the missing response indicators r_{ij}. The parameters in the dropout model are treated as nuisance parameters and are usually not of

inference interest. Thus, we should 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 GLMM unidentifiable. Therefore, one should be very cautious about adding extra nuisance parameters.

Since the missing response indicators r_{ij}, are binary, a natural model for the r_{ij}'s is a logistic regression model as follows. We may assume

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\psi}) = \prod_{j=1}^{n_i} \pi_{ij}^{r_{ij}}(1-\pi_{ij})^{1-r_{ij}}, \tag{5.1}$$

i.e. an independent assumption for the r_{ij}'s, and

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = h(\boldsymbol{\psi}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_{ij}), \tag{5.2}$$

where $\pi_{ij} = \Pr(r_{ij} = 1)$ and $h(.)$ is an often linear function of $\boldsymbol{y}_i$, $\boldsymbol{x}_i$ and $\boldsymbol{t}_{ij}$. To determine a suitable function $h(.)$, one can consider standard model selection techniques, such as the likelihood ratio test, AIC/BIC, or consider simple and reasonable linear functions. For example, if we believe that the current missing response indicator only depends on the current or previous response values, then it may be reasonable to assume $h(\boldsymbol{\psi}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_{ij}) = \psi_0 + \psi_1 y_{ij} + \psi_2 y_{i,j-1}$. Note that the independence model (5.1) is simple and may not contain too many nuisance parameters, but it fails to incorporate possible correlation among the r_{ij}'s.

To incorporate possible correlation among the r_{ij}'s, we may adapt the model considered in Ibrahim *et al.* (2001),

$$\begin{aligned} f(\boldsymbol{r}_i, \boldsymbol{\psi}) =& f(r_{i1}|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{\psi}_1) f(r_{i2}|r_{i1}, \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{\psi}_2) \\ & \cdots \times f(r_{ij}|r_{i1}, \cdots, r_{i,(j-1)}, \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{\psi}_j) \\ & \cdots \times f(r_{in_i}|r_{i1}, \cdots, r_{i,(n_i-1)}, \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{\psi}_{n_i}), \end{aligned} \tag{5.3}$$

where the $\boldsymbol{\psi}_j$'s are the parameters for the jth one-dimensional conditional distribution, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_M)$ and $M = \max_i\{n_i\}$. We assume that the $\boldsymbol{\psi}_j$'s are distinct.

The one-dimensional conditional distributions in the product of (5.3) may be chosen to be logistic regression models. Again, one can choose parsimonious one-dimensional distributions in (5.3) by standard model selection techniques. Lipsitz and Ibrahim (1996) noted that model (5.3) approximates a joint log-linear model, a natural model for binary variables.

5.3 Covariate Models

When some covariates are missing, we need a distributional assumption 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 (see [15]). 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),

$$\begin{aligned} f(\boldsymbol{x}_i, \boldsymbol{\alpha}) =& f(x_{ic}|x_{i1}, \cdots, x_{i,c-1}, \boldsymbol{\alpha}_c) \\ & \times f(x_{i,c-1}|x_{i1}, \cdots, x_{i,c-2}, \boldsymbol{\alpha}_{c-1}) \\ & \cdots \times f(x_{i1}, \boldsymbol{\alpha}_1), \end{aligned} \tag{5.4}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_c)$ and $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c$ are distinct. The index c is the number of covariates that include missing values. Note that we do not need to make distributional assumptions for the completely observed covariates, which are conditioned on and are suppressed in the expressions. Note also that this modeling scheme allows the missing covariates to be continuous, categorical and mixed. For example, suppose that x_1 is

continuous and x_2 is binary. By the above modeling strategy, we may specify a normal distribution for x_1 and a logistic regression model for x_2 conditional on x_1.

5.4 Sensitivity Analyses

Since both the dropout model and the covariate model are not verifiable based on the observed data, it is important to conduct sensitivity analyses. That is, we should 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 draw a relatively reliable conclusion. Otherwise, the results may depend on the assumed models and the conclusions may not be reliable.

Chapter 6

Real Data Examples

6.1 Introduction

In previous chapters, we have described an exact method and an approximate method for GLMMs with informative dropouts and missing covariates. In this chapter, we will discuss application of these methods to two real datasets. In Section 6.2, we consider a dataset from the AIDS Clinical Trial Group (ACTG) Protocol 315 and investigate the viral load trajectory after an antiviral treatment. In Section 6.3, we consider a dataset from a parent bereavement project to study the pattern of change of parents' mental distress over time after their children's death. In Section 6.4, we discuss computation issues in the analyses of our examples.

6.2 Example 1

6.2.1 Data Description

Our research is motivated by a longitudinal study from the AIDS Clinical Trial Group (ACTG) Protocol 315 (Wu and Ding, 1999). In this study, 46 HIV infected patients were treated with a potent antiviral drug, a combination of ritonavir, 3TC, and AZT. Plasma HIV-1 RNA (viral load) was repeatedly quantified on days 2, 7, 10, 14, 21, 28, and weeks 8, 12, and 24 after initiation of treatment. After the antiviral treatment, the patients' 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 start to rebound (rise). The Nucleic Acid Sequence-Based Amplification assay that is used to measure the viral load has a detection limit of 100 RNA copies per ml plasma. If the viral load drops below the detection limit after the treatment, the viral load can not be measured, which may indicate that the treatment is successful. Note that patients with a viral load below the detection limit at early stage may have viral rebound and may have a viral load dropping again after rebound. Figure 6.1 shows the viral load trajectories for six randomly selected patients. To investigate the treatment effect, one approach is to define the response as whether the viral load is below the detection limit or not, which is thus a binary variable. In this study, some patients drop out before the end of the study, and the dropout may be informative. Thus, the response contains non-ignorable missing values. We summarize our data in Table 6.1. As we see from Table 6.1, 8.9% of our responses are missing due to patients' dropout.

Preliminary studies show that some baseline covariates such as CD4 cell counts, tumor necrosis factor (measured by TNF levels) and total complement levels (measured by CH50), may partially explain variation in the viral load trajectory. However, some of

Table 6.1: Summary statistics

Variable	Sample mean	Sample standard deviation	Percentage of missing values
Response	0.1	0.3	8.9%
CD4	175.4	87.5	0%
CH50	242.3	49.6	15.2%
TNF	60.0	29.0	8.7%
# of patients: $N = 46$.			
# of observations per patient: $n_i = 7$ or 8.			

these covariates are also missing, since in the multi-center study some baseline covariates may not be measured at some centers. As indicated in Table 6.1, the baseline CH50 contains approximately 15.2% missing values, the TNF level contains roughly 8.7% missing values and the CD4 cell count is completely observed.

Our objectives are to model the viral load trajectory and to identify covariates that may partially predict changes of viral loads, in the presence of informative dropouts and missing covariates.

6.2.2 Models

Let y_{ij} be the viral load status for patient i at the jth visit, $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, n_i$, where $N = 46$ and $n_i = 7$ or 8. If the viral load for patient i at the jth visit is below the detection limit, $y_{ij} = 1$; otherwise, $y_{ij} = 0$. Naturally, we consider a logistic regression model for the binary response. To take into account the inter-patient variation and the intra-patient correlation, we add a random effect term b_i to the logistic regression model and obtain the following GLMM.

$$\begin{aligned} \text{logit}\left\{\Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)\right\} &= \log\left\{\frac{\Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)}{1 - \Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)}\right\} \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 t_{ij} + b_i, \end{aligned} \tag{6.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$, x_{i1} is the baseline CD4 cell count for patient i, x_{i2} is the baseline CH50 for patient i, x_{i3} is the baseline TNF for patient i, and t_{ij} is the jth measurement time for patient i. The regression coefficients, β_1, β_2, β_3, and β_4, represent the fixed effects associated with the baseline CD4, CH50, TNF, and time respectively, and b_i represents the random effect associated with each patient. We assume that $b_1, \cdots, b_N$ are independent and follow an identical normal distribution $N(0, \sigma^2)$ with σ^2 unknown.

In this study, the baseline CH50 and TNF contain missing values and the CD4 cell count is completely observed. For this example, it appears reasonable to assume that the missingness of baseline covariates is MAR, *i.e.*, the missingness may depend on the observed values but not on the missing values. To make a valid likelihood inference, we need to specify a model for the covariates which contain missing values. Since CH50 and TNF are continuous and each approximately has a normal marginal distribution, the joint distribution of CH50 and TNF (*i.e.*, x_{i2} and x_{i3}) may be written as a product of two one-dimensional normal distributions.

$$f(x_{i2}, x_{i3}|x_{i1}, \boldsymbol{\alpha}) = f(x_{i3}|x_{i1}, x_{i2}) f(x_{i2}|x_{i1}), \tag{6.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_7)^T$, $f(x_{i2}|x_{i1})$ is the density function of $N(\alpha_1 + \alpha_2 x_{i1}, \alpha_3)$, and $f(x_{i3}|x_{i1}, x_{i2})$ is the density function of $N(\alpha_4 + \alpha_5 x_{i1} + \alpha_6 x_{i2}, \alpha_7)$.

As noted earlier, 8.9% of the responses y_{ij} are missing due to patients' dropout. The dropout may be due to drug intolerance or drug resistance, so we assume that the response is non-ignorably missing or the dropout is informative, *i.e.*, the missingness of responses may depend on the missing values. When the missing data (response) are non-ignorable, we must model the missing data mechanism in order to obtain valid statistical results. To incorporate the missing data mechanism, we need to specify a distribution for the missing response indicator. The missing response indicator is defined as $r_{ij} = 1$ if y_{ij} is missing; $r_{ij} = 0$ if y_{ij} is observed. Here, we use a logistic regression model for the

missing data indicator, which includes the current response y_{ij}, CD4 cell count x_{i1} and time t_{ij} as covariates and is chosen based on the likelihood ratio test.

$$\text{logit}\{\Pr(r_{ij}=1|\boldsymbol{\phi})\} = \log\left\{\frac{\Pr(r_{ij}=1|\boldsymbol{\phi})}{1-\Pr(r_{ij}=1|\boldsymbol{\phi})}\right\} = \phi_0 + \phi_1 y_{ij} + \phi_2 x_{i1} + \phi_3 t_{ij}, \quad (6.3)$$

where $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2, \phi_3)^T$. Thus, in model (6.3), we link the missingness of the response to the values being missing and therefore allow the response to be non-ignorably missing. For simplicity, we focus on the following independent model

$$f(\boldsymbol{r}|\boldsymbol{\phi}) = \prod_{i=1}^{N}\prod_{j=1}^{n_i} \Pr(r_{ij}=1|\boldsymbol{\phi})^{r_{ij}}\{1-\Pr(r_{ij}=1|\boldsymbol{\phi})\}^{1-r_{ij}}.$$

More complicated models without assuming r_{ij}'s being independent are possible as well, which contain more nuisance parameters and may be unidentifiable.

6.2.3 Analysis and Results

In this section, we analyze the ACTG protocol 315 dataset using our proposed methods. Note that before our analysis, covariates CD4, CH50 and TNF were standardized to avoid extremely small estimates. We consider the following methods to estimate the parameters in model (6.1) – (6.3) with informative dropouts and missing covariates:

(i) the exact method using the Monte Carlo EM algorithm,

(ii) the approximate method using the Monte Carlo EM algorithm.

Table 6.2 shows maximum likelihood estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$, along with their standard errors and p-values, based on the above two methods. Compared with the approximate method, the exact method sometimes gave somewhat larger estimates and smaller standard errors. For example, CD4 cell count is marginally significant (p-value 0.107) based on the approximate method, but highly significant (p-value 0.025)

Table 6.2: Estimates for the AIDS data

Methods		Parameters				
		β_0	β_1	β_2	β_3	β_4
Exact method	Estimate	-4.811	0.898	0.634	-0.745	5.869
	SE	0.998	0.400	0.456	0.538	1.361
	p-value	< 0.001	0.025	0.164	0.166	< 0.001
Approximate method	Estimate	-3.879	0.861	0.395	-0.291	6.240
	SE	0.789	0.534	0.598	0.600	0.880
	p-value	< 0.001	0.107	0.508	0.627	< 0.001

* SE refers to the standard error.

based on the exact method. Since the approximate method integrates out the random effects instead of sampling the random effects, it should be faster than the exact method. However, in this example, the EM convergence for the exact method is obtained in 21 iterations, while the EM convergence for the approximate method is reached in 55 iterations. This is probably because only one random effect is included in model (6.1).

From Table 6.2, both methods indicate that the time covariate is highly significant, suggesting a strong relationship between the viral load trajectory and time. The estimated coefficient for the time covariate, $\hat{\beta}_4$, is 5.869 based on the exact method. This means that the estimated odds of patients' viral loads dropping below the detection limit is $\exp(5.869) = 354$ times higher when time increases by one unit (6 months). The exact method suggests that CD4 cell count may be associated with patients' viral loads. The estimated coefficient for CD4, $\hat{\beta}_1$, is 0.898. This means that the estimated odds of patients' viral loads dropping below the detection limit, is $\exp(0.898) = 2.45$ times higher when CD4 increases by one unit (262 cell counts). Based on the p-values, the baseline CH50 and TNF do not appear to have significant effects on patients' viral loads, using either of the two methods of estimation.

As discussed in previous chapters, the PX-EM algorithm should have a higher convergence rate than EM. This is confirmed in this example. The number of iterations to convergence for the exact method is 21 by the EM algorithm, while the number of iterations to convergence for the exact method is 10 by the PX-EM algorithm.

6.2.4 Sensitivity Analysis

To check the sensitivity of our results to the choice of the covariate model, we re-analyze the dataset using the following alternative covariate models, which are obtained based on a standard model selection method – the likelihood ratio test.

(i) Alternative Covariate Model 1 (CM1): Model (6.2) with $\alpha_2 = \alpha_5 = 0$. The two conditional distribution in the covariate model become $x_{i2}|x_{i1} \sim N(\alpha_1, \alpha_3)$ and $x_{i3}|x_{i1}, x_{i2} \sim N(\alpha_4 + \alpha_6 x_{i2}, \alpha_4)$, *i.e.*, x_{i2} and x_{i3} are independent of x_{i1}.

(ii) Alternative Covariate Model 2 (CM2): Model (6.2) with $\alpha_2 = \alpha_5 = \alpha_6 = 0$. The two conditional distributions in the covariate model become $x_{i2}|x_{i1} \sim N(\alpha_1, \alpha_3)$ and $x_{i3}|x_{i1}, x_{i2} \sim N(\alpha_4, \alpha_7)$, *i.e.*, x_{i1}, x_{i2} and x_{i3} are independent.

The estimates based on the original covariate model (6.2), and the alternative covariate models CM1 and CM2 are shown in Table 6.3. As we can see from Table 6.3, the three covariate models gave similar results. This suggests that parameter estimates and their standard errors may not depend on the covariate models.

Similarly, we need to assess sensitivity of our results to the dropout models. We performed sensitivity analyses based on the following choices of the dropout models.

(i) Alternative Dropout Model 1 (DM1):

$$\text{logit}\left\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\right\} = \phi_0 + \phi_1 y_{i,j-1} + \phi_2 y_{ij} + \phi_3 x_{i1} + \phi_4 t_{ij};$$

Table 6.3: Sensitivity analysis for covariate models

Covariate Models		Parameters				
		β_0	β_1	β_2	β_3	β_4
Original Model (6.2)	Estimate	-4.811	0.898	0.634	-0.745	5.869
	SE	0.998	0.400	0.456	0.538	1.361
	p-value	< 0.001	0.025	0.164	0.166	< 0.001
CM1	Estimate	-4.731	0.906	0.594	-0.767	5.754
	SE	1.007	0.387	0.433	0.520	1.375
	p-value	< 0.001	0.019	0.170	0.140	< 0.001
CM2	Estimate	-4.893	0.921	0.676	-0.763	5.920
	SE	0.999	0.424	0.455	0.543	1.356
	p-value	< 0.001	0.030	0.137	0.160	< 0.001

* SE refers to the standard error.

(ii) Alternative Dropout Model 2 (DM2):

$$\text{logit}\left\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\right\} = \phi_0 + \phi_1 y_{i,j-1} + \phi_2 y_{ij};$$

(iii) Alternative Dropout Model 3 (DM3):

$$\text{logit}\left\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\right\} = \phi_0 + \phi_2 x_{i1} + \phi_3 t_{ij}.$$

Note that DM3 corresponds to an ignorable missing data mechanism (*i.e.*, MAR). Table 6.4 gives the estimates, standard errors, and p-values based on the original dropout model and the alternative dropout models DM1, DM2 and DM3 respectively. As we can see from Table 6.4, all these dropout models, except dropout model DM2, gave similar results. Dropout model DM2, which excludes CD4 and time, produced slightly smaller absolute values of estimates and standard errors. But this does not affect our conclusion. Thus, our results are not very sensitive to the choice of the dropout models.

Table 6.4: Sensitivity analysis for dropout models

Dropout Models		Parameters				
		β_0	β_1	β_2	β_3	β_4
Original Model (6.3)	Estimate	-4.811	0.898	0.634	-0.745	5.869
	SE	0.998	0.400	0.456	0.538	1.361
	p-value	< 0.001	0.025	0.164	0.166	< 0.001
DM1	Estimate	-4.911	1.047	0.611	-0.735	6.656
	SE	1.027	0.417	0.500	0.613	1.341
	p-value	< 0.001	0.012	0.221	0.230	< 0.001
DM2	Estimate	-3.761	0.772	0.428	-0.323	6.011
	SE	0.571	0.299	0.311	0.324	0.898
	p-value	< 0.001	0.010	0.168	0.318	< 0.001
DM3	Estimate	-4.867	0.949	0.630	-0.753	6.153
	SE	1.001	0.410	0.466	0.549	1.345
	p-value	< 0.0001	0.021	0.177	0.171	< 0.001

* SE refers to the standard error.

6.2.5 Conclusion

Based on our analyses, we conclude that a patient's viral load tends to more likely drop below the detection limit if he/she stays longer in the study, and a patient with a higher baseline CD4 cell count is more likely to have his/her viral load below the detection limit over the time course. In this example, the exact method converged much faster than the approximate method, and gave smaller standard errors. Also, different covariate models and different dropout models lead to essentially the same results, so our conclusions may be robust.

6.3 Example 2

6.3.1 Data Description

Our second example involves a longitudinal study from a parent bereavement project, which investigates long-term mental outcomes of parents whose children died by accident, suicide, or homicide. After their children's death, the parents usually experience a high level of mental distress. In this study, the mental distress of 239 parents were calculated based on their answers in the questionnaire, at baseline (*i.e.* 4 to 6 weeks after their children's death), and then at 4, 12, 24 and 60 months post-death. The Global Severity Index (GSI), which is the most sensitive indicator of mental distress, is used to measure the parents' distress levels. A high GSI score indicates a high level of mental distress. If the parents' adjustment to their children's death goes well, their GSI scores will decrease over time, at least lower than their baseline GSI scores. Figure 6.2 shows the profiles of GSI scores for six randomly selected parents from 239 parents enrolled in this study. To investigate how the parents' mental distress changes over time after their children's death, we treat whether or not a parent's GSI score after baseline is lower than his/her baseline value as response. Several other relevant factors were also obtained at baseline, including parents' gender, marital status, age, education, annual income, cause of death, age, and gender of the deceased child. These baseline factors may be important predictors of parents' distress and thus are viewed as covariates. Summary statistics for the response, education and income are given in Table 6.5.

Since the response is binary, we consider a GLMM model for analysis. Note that some baseline covariates such as income contain missing values, and some responses are also missing since some parents may be not in a good mood at the scheduled time. As indicated in Table 6.5, 4.2% of incomes are missing and 19.7% of responses are missing.

Table 6.5: Summary statistics

Variable	Sample mean	Sample standard deviation	Percentage of missing values
Response	0.7	0.5	19.7%
Education	13.7	2.3	0%
Income	4.67	1.9	4.2%
# of parents: $N = 239$.			
# of observations per parent: $n_i = 4$.			

Our objectives are to investigate the change of parent's distress levels over time and to determine which covariates affect the parent's mental distress. We will use the methods developed here for GLMM models with informative dropouts and missing covariates.

6.3.2 Models

To get a parsimonious model, we used standard model selection techniques such as the likelihood ratio test to determine which covariates should be included in our model. Since some covariates and responses contain missing values, model selection were carried out based on the complete-case method. Based on these analyses, we include income, education, and time as covariates in the model. Note that education does not have missing values, while income contains 4.2% missing values.

We denote y_{ij} as the response for parent i at the jth time after baseline, $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, n_i$. Here $N = 239$ and $n_i = 4$. If GSI for parent i at the jth time is lower than his/her baseline GSI, $y_{ij} = 1$; otherwise, $y_{ij} = 0$. We consider the following GLMM to investigate the effects of covariates on the mental distress.

$$\begin{aligned} \text{logit}\left\{\Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)\right\} &= \log\left\{\frac{\Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)}{1 - \Pr(y_{ij} = 1|\boldsymbol{\beta}, b_i)}\right\} \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 t_{ij} + b_i, \end{aligned} \tag{6.4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, x_{i1} is the education level for parent i, x_{i2} is the annual family income for parent i, and t_{ij} is the jth scheduled time for parent i. The regression coefficients, β_1, β_2, and β_3, represent the fixed effects associated with the parents' education level, income, and time respectively, and b_i represents the random effect associated with each parent. Here, we assume that $b_1, \cdots, b_N$ are independent and have an identical normal distribution $N(0, \sigma^2)$ with σ^2 unknown.

Note that income x_{i2} contains approximately 4.2% missing values. Here, we assume that the missing income is MAR. To incorporate missing covariates, we need to specify a model for income. We assume the following covariate distribution

$$x_{i2}|x_{i1} \sim N(\alpha_1 + \alpha_2 x_{i1}, \alpha_3). \tag{6.5}$$

Note also that 19.7% of responses y_{ij} are missing. The response (*i.e.*, GSI status) is missing probably due to parents' high stress, so we assume that our response is non-ignorably missing, *i.e.*, the missingness of responses may depend on the missing values. To incorporate the missing data mechanism in our analysis, we specify a distribution for the missing response indicator. Recall that the missing response indicator is defined as $r_{ij} = 1$ if y_{ij} is missing; $r_{ij} = 0$ if y_{ij} is observed. Here, we use a logistic model for the missing response mechanism, which includes the current response value y_{ij}, the immediate previous response value $y_{i,j-1}$, education x_{i1} , and time t_{ij} as covariates.

$$\text{logit}\left\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\right\} = \log\left\{\frac{\Pr(r_{ij} = 1|\boldsymbol{\phi})}{1 - \Pr(r_{ij} = 1|\boldsymbol{\phi})}\right\} = \phi_0 + \phi_1 y_{i,j-1} + \phi_2 y_{ij} + \phi_3 x_{i1} + \phi_4 t_{ij}, \tag{6.6}$$

where $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2, \phi_3)^T$. We assume that the r_{ij}'s are independent for all i and j. Note that the covariates in model (6.6) are selected based on the likelihood ratio test.

Table 6.6: Estimates for the Parent Bereavement data

Methods		Parameters			
		β_0	β_1	β_2	β_3
Exact	Estimate	-1.882	0.182	0.083	0.345
method	SE	0.966	0.070	0.165	0.258
	p-value	0.051	0.010	0.612	0.181
Approximate	Estimate	-1.579	0.139	0.058	-0.239
method	SE	0.898	0.065	0.152	0.193
	p-value	0.079	0.033	0.704	0.216

* SE refers to the standard error.

6.3.3 Analysis and Results

We consider the following methods to estimate the parameters in models (6.4)-(6.6).

(i) the exact method using the Monte Carlo EM algorithm,

(ii) the approximate method using the Monte Carlo EM algorithm.

Estimates of $\boldsymbol{\beta}$, along with their standard errors and p-values, are shown in Table 6.6. Compared with the exact method, the approximate method resulted in smaller absolute values of estimates and smaller standard errors. Especially for the estimate of β_3, the exact and approximate methods gave opposite results. As discussed in previous chapters, the approximate method should have a faster convergence rate, since it avoids sampling the random effect in each EM iteration. However, for this example, the number of iterations to convergence for the approximate method is 24, larger than the number of iterations to convergence for the exact method, which is 13. The PX-EM algorithm improved the convergence speed a bit in this example. The number of iterations to convergence for the exact method based on PX-EM is 9, smaller than 13.

Table 6.6 shows that education is significant based on the exact method and the approximate method. The estimate for education $\hat{\beta}_1$ based on the exact method is 0.182,

Table 6.7: Sensitivity analysis for covariate models

Covariate Models		Parameters			
		β_0	β_1	β_2	β_3
Original model (6.5)	Estimate	-1.882	0.182	0.083	0.345
	SE	0.966	0.070	0.165	0.258
	p-value	0.051	0.010	0.612	0.181
CM1	Estimate	-1.969	0.189	0.107	0.312
	SE	1.043	0.076	0.175	0.255
	p-value	0.059	0.013	0.542	0.222

*SE refers to the standard error.

which suggests that the estimated odds of having a lower distress than the baseline value is $\exp(0.182) = 1.2$ times higher, when parents increase their education level by one unit. Based on both the exact method and the approximate method, income and time do not have significant effects on change of parents' mental distress.

6.3.4 Sensitivity Analysis

To check the sensitivity of the above results to the covariate models, we consider the following alternative covariate model

(i) Alternative Covariate Model 1 (CM1): Model (6.5) with $\alpha_2 = 0$. That is, $x_{i2}|x_{i1} \sim N(\alpha_1, \alpha_3)$, *i.e.*, x_{i2} is independent of x_{i1}.

Table 6.7 shows that results based on the original covariate model and the alternative covariate model are quite similar. This suggests that the results may be robust to the covariate models.

We also check the sensitivity of our results to the dropout models. We consider the following alternative dropout models.

Table 6.8: Sensitivity analysis for dropout models

Dropout Model		Parameters			
		β_0	β_1	β_2	β_3
Original model (6.6)	Estimate	-1.882	0.182	0.083	0.345
	SE	0.966	0.070	0.165	0.258
	p-value	0.051	0.010	0.612	0.181
DM1	Estimate	-1.592	0.161	0.063	0.545
	SE	0.808	0.059	0.136	0.273
	p-value	0.049	0.006	0.644	0.046
DM2	Estimate	-1.958	0.193	0.087	0.239
	SE	1.019	0.074	0.169	0.254
	p-value	0.055	0.010	0.604	0.347
DM3	Estimate	-1.460	0.167	0.094	0.939
	SE	1.006	0.073	0.172	0.282
	p-value	0.146	0.022	0.585	< 0.001

* SE refers to the standard error.

(i) Alternative Dropout Model 1 (DM1):

$$\text{logit}\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\} = \phi_0 + \phi_1 y_{ij} + \phi_2 x_{i1} + \phi_3 t_{ij};$$

(ii) Alternative Dropout Model 2 (DM2):

$$\text{logit}\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\} = \phi_0 + \phi_1 y_{i,j-1} + \phi_2 y_{ij};$$

(iii) Alternative Dropout Model 3 (DM3):

$$\text{logit}\{\Pr(r_{ij} = 1|\boldsymbol{\phi})\} = \phi_0 + \phi_1 x_{i1} + \phi_2 t_{ij}.$$

Note that DM3 suggests that the missing responses may be MAR. The comparison of estimates based on the original dropout model and the above alternative dropout models is given in Table 6.8. As we can see from Table 6.8, whether y_{ij} and $y_{i,j-1}$ are included in the dropout model affects our inference on the coefficient of the time covariate (*i.e.*, β_3). For the dropout model DM3, which excludes y_{ij} and $y_{i,j-1}$ as covariates, we obtain a

highly significant p-value (< 0.001) for β_3. For the dropout model DM1, which excludes $y_{i,j-1}$, we get a marginally significant β_3. Other dropout models lead to insignificant β_3. That is, the conclusion about β_3 is sensitive to the choices of the dropout models. However, estimates of other parameters and their standard errors are quite robust to the different dropout models.

6.3.5 Conclusion

Our analyses suggest that parents with a higher education level are more likely to have a lower level of mental distress, *i.e.*, they may have a good adjustment to their children's death. Possibly due to the low dimension of random effects and the small number of intra-parent measurements, the approximate method in this example did not improve the convergence rate. Unlike in Example 1, the approximate method in this example gave smaller standard errors than the exact method. For this example, sensitivity analyses suggest that our conclusions about the time covariate may not be reliable, *i.e.*, they may depend on the choices of dropout models, but our conclusions about other covariates are reliable.

6.4 Computation Issues

Starting values. For the EM algorithms in our examples, the starting values for $\boldsymbol{\beta}$ were obtained based on the logistic regressions using the completely observed cases, the starting values for $\boldsymbol{\alpha}$ were obtained based on linear regression models using the completely observed cases, and the starting values for $\boldsymbol{\phi}$ were obtained based on logistic regressions using the last-value-carried-forward method.

Convergence of the Gibbs sampler. We checked the convergence of the Gibbs

sampler used in each Monte-Carlo EM by examining the time series and autocorrelation function plots. For example, Figure 6.3 shows the time series and autocorrelation function plots for generating missing CH50 in the first example. From Figure 6.3, we notice that the Gibbs sampler converged quickly and the autocorrelations between successive generated samples are negligible. We also drew the time series plot and autocorrelation function for the random effect b_{46} associated with patient 46 in the first example, shown in Figure 6.4. It shows that the Markov chain converged quickly, but the autocorrelations are negligible after lag 6. Time series and autocorrelation function plots for other random effects and other missing covariates show similar behaviors. Therefore, for each EM iteration, we discarded the first 200 samples as the burn-in, and then we took one sample from every 10 simulated samples until 500 samples were obtained.

Stopping rule. The stopping rule for the EM and PX-EM algorithms in our examples is that the relative change in the parameter values from successive iterations is smaller than a given tolerance level (e.g. 0.01). However, due to Monte Carlo errors induced by the Gibbs sampler, it is difficult to converge for a extremely small tolerance level, otherwise it may converge very slowly.

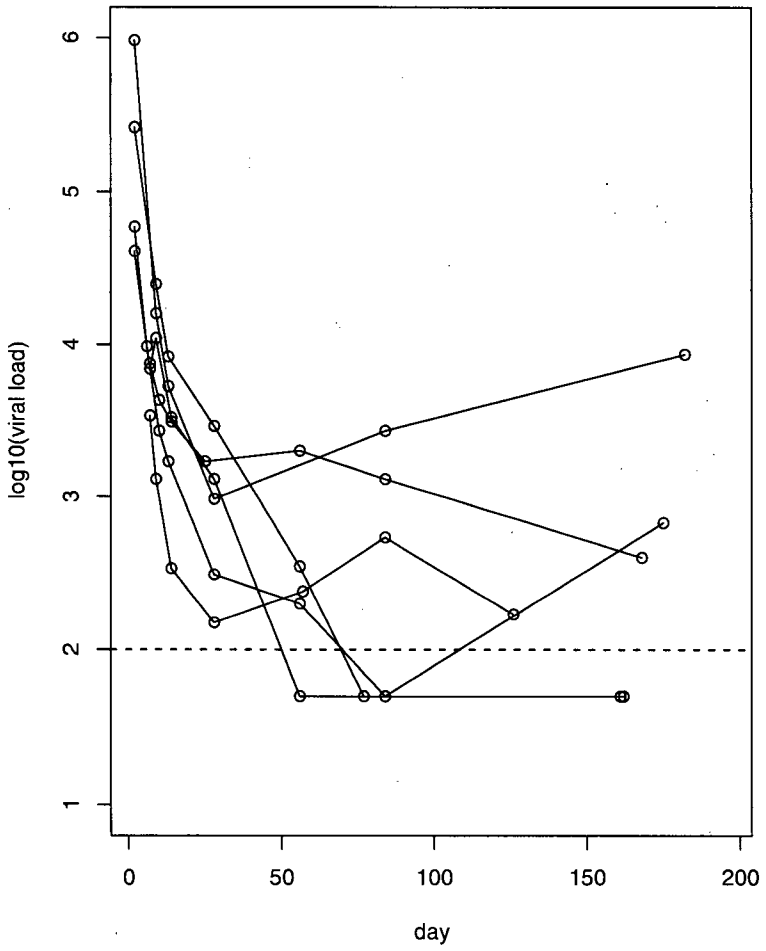

Figure 6.1: Viral loads ($\log_{10}$ scale) for six randomly selected patients. The open dots are the observed values and the dashed line indicates the detection limit of viral loads. The viral loads below the detection limit are substituted with $\log_{10}(50)$.

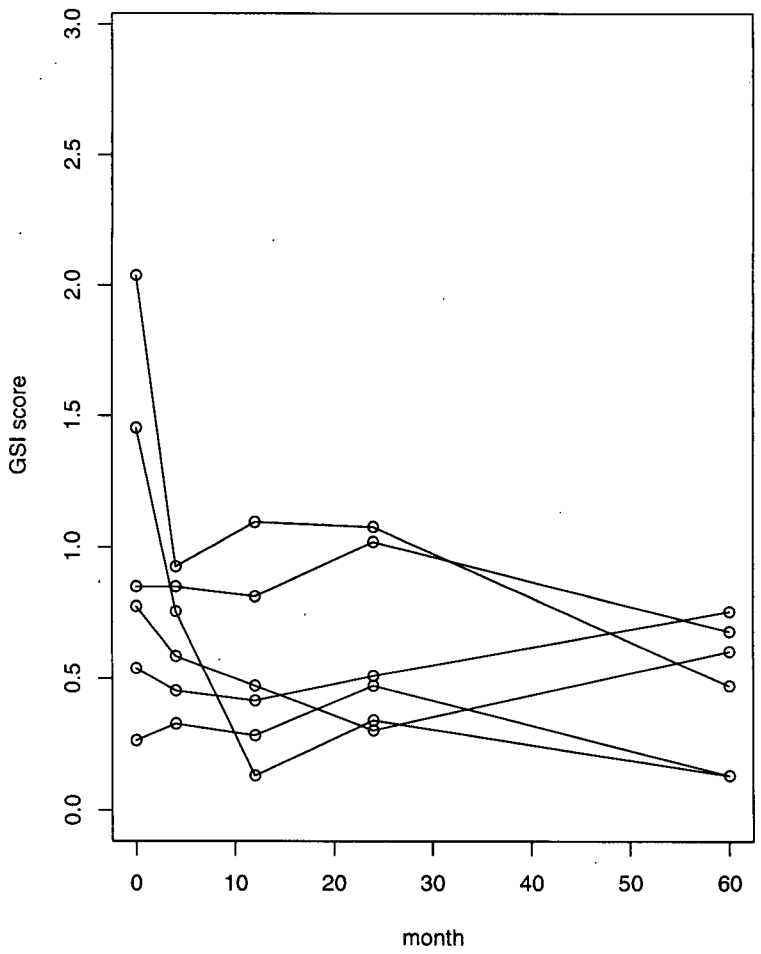

Figure 6.2: GSI scores for six randomly selected parents. The open dots are the observed values and the GSI scores at time 0 are the baseline values.

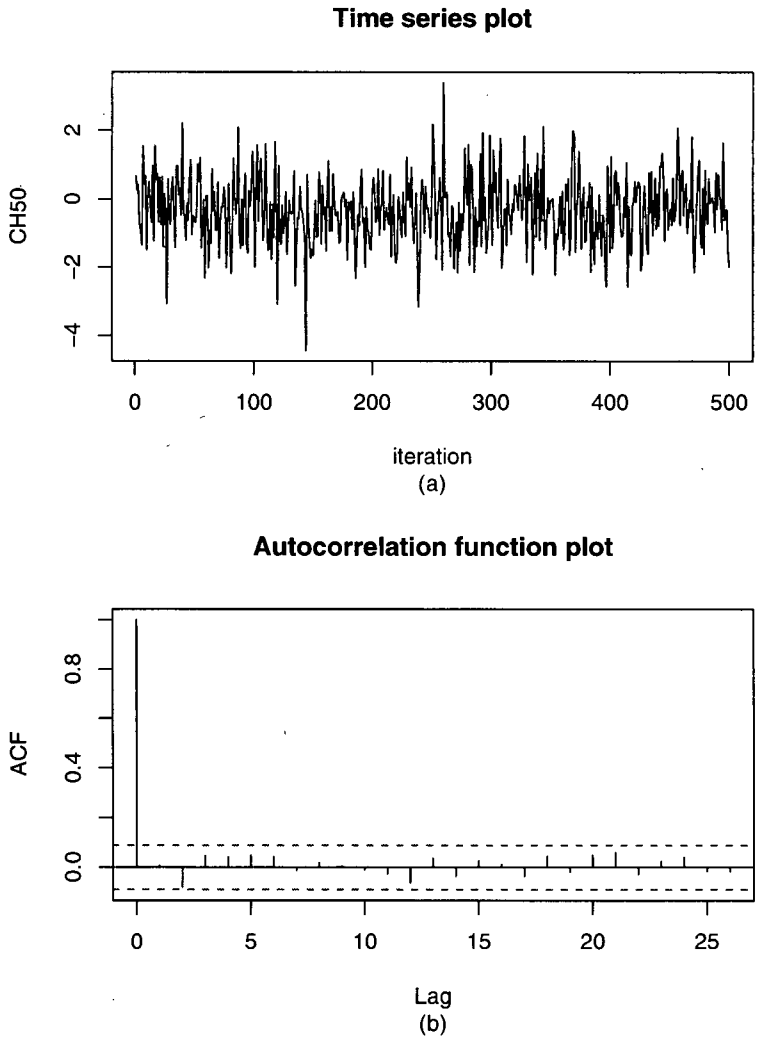

Figure 6.3: (a) Time series and (b) autocorrelation function plots for CH50.

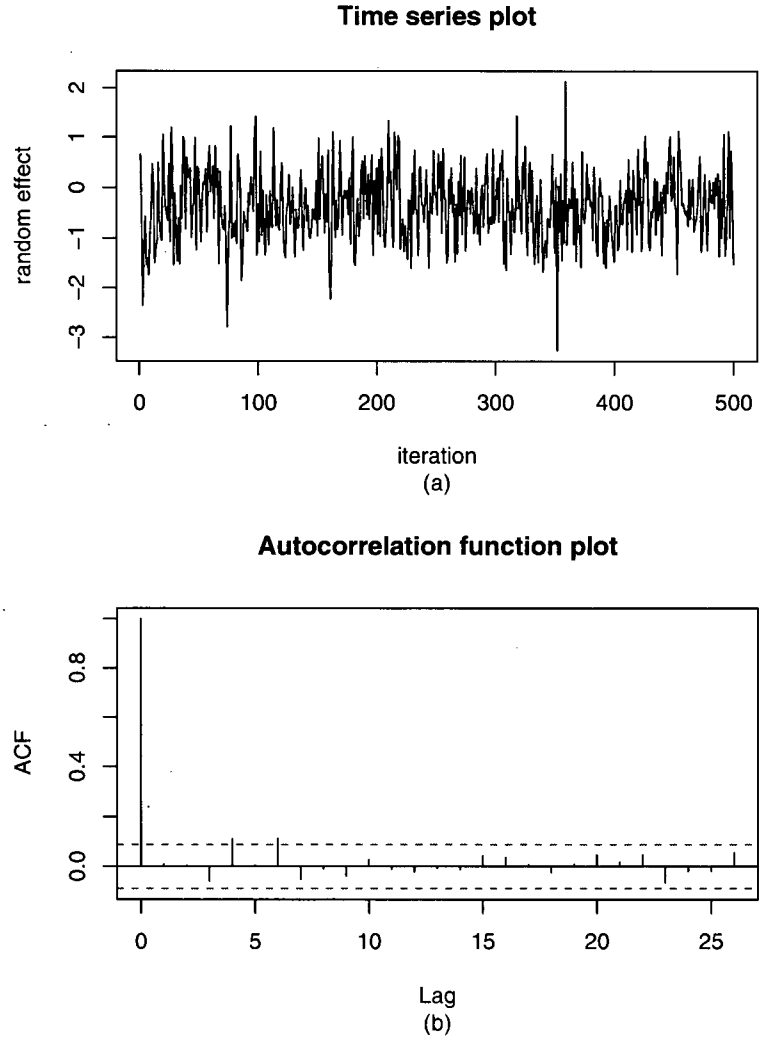

Figure 6.4: (a) Time series and (b) autocorrelation function plots for b_{46} associated with patient 46.

Chapter 7

Simulation Study

7.1 Introduction

To evaluate the performance of the two proposed methods: the exact method (EX) and the approximate method (AP), we conduct a simulation study in this chapter. In our simulations, we compare EX and AP in terms of biases and mean-squared errors of their estimates. Section 7.2 gives a description of data generation models in our simulations. In Section 7.3, we compare two methods of estimation in four different situations, and examine the effects of missing rates, variance of random effects, sample size, and number of intra-individual measurements. We conclude this chapter in Section 7.4.

7.2 Description of the Simulation Study

7.2.1 Models

We generate the response variable y_{ij} from the following GLMM

$$\begin{aligned} \text{logit}\,\{\Pr(y_{ij}=1|\boldsymbol{\beta},b_i)\} =& \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 t_{ij}+b_i, \\ & i=1,\cdots,N, \quad j=1,\cdots,n_i, \end{aligned} \tag{7.1}$$

where $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\beta_3)$, the random effects b_i's are assumed to be $i.i.d$ with a normal distribution $N(0,\sigma^2)$. The true values of $\boldsymbol{\beta}$ and σ^2 are $\boldsymbol{\beta}=(-3,0.5,-0.3,4)$ and $\sigma^2=0.3$. The number of individuals (sample size) is $N=50$, and the number of intra-individual measurements is $n_i=10$. The n_i time points for each individual are 2, 7, 9, 14, 20, 28, 40, 56, 70, and 84.

The covariates x_{i1} and x_{i2} are continuous variables. Covariate variable x_{i1} is generated from $N(1,0.1)$ and covariate x_{i2} is generated from the following model

$$x_{i2}|x_{i1}\sim N(\alpha_1+\alpha_2 x_{i1},\alpha_3), \tag{7.2}$$

where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\alpha_3)$ and the true value of $\boldsymbol{\alpha}$ is $\boldsymbol{\alpha}=(-1.5,1,0.2)$. In our simulation study, the missing covariate mechanism is assumed to be MAR. For each generated data set, we keep x_{i1} completely observed and delete those values of x_{i2} with probability 0.8 which correspond to the largest values of x_{i1}.

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

$$\text{logit}\,\{\Pr(r_{ij}=1|\boldsymbol{\phi})\}=\phi_0+\phi_1 y_{ij}, \tag{7.3}$$

where $\boldsymbol{\phi}=(\phi_0,\phi_1)$ and the missing response indicator r_{ij} is a binary variable. The above model suggests that the missingness of the response depends on the missing values, and

thus the response is non-ignorably missing (NIM). We generate missing responses based on the model (7.3). That is, if $r_{ij} = 1$, then y_{ij} is deleted; if $r_{ij} = 0$, y_{ij} is retained. Note that different values of $\boldsymbol{\phi}$ will lead to different missing rates of responses. We will discuss EX and AP in two different values of $\boldsymbol{\phi}$ in Section 7.3.1.

7.2.2 Bias and Mean-Squared Error

We examine the convergence of Monte Carlo Markov Chains by their time series plots and autocorrelations function plots. Time series plots and autocorrelation function plots have shown that Markov Chains converged very fast, usually in 100 or 200 iterations, and autocorrelations are negligible after lag 2. Figure 7.1 and Figure 7.2 show typical time series plots and typical autocorrelation plots for the missing covariates and random effects from our simulated data sets. Thus, to ensure the convergence, we conservatively discard the first 500 samples, and then take one sample every 10 samples until we obtained the desired number of samples. We run $B = 100$ replicates in each simulation, and compare EX and AP in terms of biases and mean square errors (MSEs). Here, bias and MSE are reported in terms of percent relative bias and percent relative root mean-squared error. The bias for β_j, the jth component of $\boldsymbol{\beta}$, is defined as

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

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

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

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

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

and the percent relative $\sqrt{\text{MSE}}$ ($\%\sqrt{\text{MSE}}$) is

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

In our simulations, we consider (i) two missing rates: 20% and 40%, (ii) two different variances of b_i: $\sigma^2 = 0.3$ and $\sigma^2 = 1$, (iii) two different sample sizes: $N = 50$ and $N = 100$, (iv) three different numbers of intra-individual measurements: $n_i = 5$, $n_i = 10$ and $n_i = 20$. In the above situations, we compare estimates based on EX and AP, and investigate how the missing rate, the variance of b_i, the sample size, and the number of intra-individual measurements affect estimation of the parameters.

7.3 Simulation Results

7.3.1 Comparison of Methods with Varying Missing Rates

To see the impact of the missing rates on estimation by EX and AP, we estimate the parameters based on two missing rates respectively. A 20% missing rate and a 40% missing rate are considered. In our case, if the true values of $\boldsymbol{\phi}$ are $\boldsymbol{\phi} = (-1.8, 1)$, the missing response mechanism (7.3) leads to an average of 20% missing rate for the response; if $\boldsymbol{\phi} = (-0.8, 1)$, the missing response mechanism (7.3) leads to an average of 40% missing rate. Regarding the covariate x_2 with missing values, we take the same missing rate as the response.

Table 7.1 shows average simulation results from 100 simulated data sets based on methods EX and AP. EX and AP yield comparable results for the two missing rates. In the 20% missing rate case, compared with AP, EX gives smaller biases, but slightly larger mean-squared errors. In the 40% missing rate case, EX produces slightly larger biases *and* mean-squared errors than AP. As we can see from Table 7.1, the missing rate

Table 7.1: Simulation results with varying missing rates

Missingness rate (%)	Parameter (true values)	%bias		%$\sqrt{\text{MSE}}$	
		EX	AP	EX	AP
20	$\beta_0 = -3$	6	1	29	28
	$\beta_1 = 0.5$	-6	-8	115	112
	$\beta_2 = -0.3$	-3	-5	144	141
	$\beta_3 = 4$	2	-2	12	11
40	$\beta_0 = -3$	22	14	46	40
	$\beta_1 = 0.5$	44	39	164	148
	$\beta_2 = -0.3$	50	48	153	148
	$\beta_3 = 4$	3	-4	17	15

greatly affects biases and mean-squared errors of estimates from two methods, especially estimates from EX, that is, the absolute values of biases and the mean-squared errors increase with the missing rate.

7.3.2 Comparison of Methods with Different Variances

To investigate how the variability of b_i affects the estimates from two methods, we consider two sets of values of σ^2: a small variance $\sigma^2 = 0.3$ and a moderate variance $\sigma^2 = 1$ at the same missing rate 20%.

We summarize the simulation results of estimation from EX and AP in Table 7.2. EX produces slightly lager mean-squared errors of estimates than AP in both cases. However, the performance of EX is still quite close to AP. We also note that the mean-squared errors of estimates based on EX and AP increase as σ^2 increases. That is, the variability of random effects affects estimation of EX and AP.

Table 7.2: Simulation results with varying variances

Variance	Parameter (true values)	%bias		%$\sqrt{\text{MSE}}$	
		EX	AP	EX	AP
Small	$\beta_0 = -3$	6	1	29	28
Variance	$\beta_1 = 0.5$	-6	-8	115	112
$\sigma^2 = 0.3$	$\beta_2 = -0.3$	-3	-5	144	141
	$\beta_3 = 4$	2	-2	12	11
Moderate	$\beta_0 = -3$	4	-5	36	36
Variance	$\beta_1 = 0.5$	-3	-7	156	150
$\sigma^2 = 1$	$\beta_2 = -0.3$	9	7	176	169
	$\beta_3 = 4$	2	-6	12	12

7.3.3 Comparison of Methods with Different Sample sizes

To examine the effect of the sample size on estimation, we estimate the parameters based on EX and AP with two different sample sizes: $N = 50$ and $N = 100$, with a 20% missing rate.

The average simulation results from EX and AP are shown in Table 7.3. We note that, as the sample size increases from 50 to 100, AP becomes more reliable in the sense that AP provides somewhat smaller biases and mean-squared errors than EX. However, AP does not outperform EX much. Moreover, both AP and EX yield smaller mean-squared errors for larger sample sizes.

7.3.4 Comparison of Methods with Varying Intra-individual Measurements

To see how the number of intra-individual measurements affects our estimates, we consider the two methods of estimation under three different numbers of intra-individual measurements : $n_i = 5$, $n_i = 10$ and $n_i = 20$. If $n_i = 5$, the time points for each individ-

Table 7.3: Simulation results with varying sample sizes

Number of individuals	Parameter (true values)	%bias		$\%\sqrt{\text{MSE}}$	
		EX	AP	EX	AP
N=50	$\beta_0 = -3$	6	1	29	28
	$\beta_1 = 0.5$	-6	-8	115	112
	$\beta_2 = -0.3$	-3	-5	144	141
	$\beta_3 = 4$	2	-2	12	11
N=100	$\beta_0 = -3$	9	4	21	20
	$\beta_1 = 0.5$	8	4	82	79
	$\beta_2 = -0.3$	11	9	101	98
	$\beta_3 = 4$	2	-2	9	8

ual are 2, 9, 20, 40, 70; if $n_i = 10$, the time points for each individual are 2, 7, 9, 14, 20, 28, 40, 56, 70 and 84; if $n_i = 20$, the time points for each individual are 2, 4, 7, 9, 12, 14, 17, 20, 24, 28, 33, 40, 46, 53, 56, 60, 66, 70, 76 and 84.

The simulation results are indicated in Table 7.4. Both EX and AP produce smaller mean-squared errors as the number of intra-individual measurements increases (*i.e.*, as n_i increases). Compared with EX, AP provides slightly smaller mean-squared errors in the three cases. But, the results from EX and AP are still quite close and become closer as n_i gets larger.

7.3.5 Conclusion

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

- Estimates based on EX and AP get worse in terms of biases and mean square errors as the missing rate gets larger.

- The mean-squared errors of estimates from both EX and AP increase as the vari-

Table 7.4: Simulation results with varying intra-individual measurements

Number of intra-individual measurements	Parameter (true values)	%bias		%$\sqrt{\text{MSE}}$	
		EX	AP	EX	AP
$n_i = 5$	$\beta_0 = -3$	11	4	40	37
	$\beta_1 = 0.5$	9	1	150	148
	$\beta_2 = -0.3$	-0.1	-7	200	193
	$\beta_3 = 4$	7	2	19	16
$n_i = 10$	$\beta_0 = -3$	6	1	29	28
	$\beta_1 = 0.5$	-6	-8	115	112
	$\beta_2 = -0.3$	-3	-5	144	141
	$\beta_3 = 4$	2	-2	12	11
$n_i = 20$	$\beta_0 = -3$	5	1	22	21
	$\beta_1 = 0.5$	-4	-4	91	90
	$\beta_2 = -0.3$	2	4	117	118
	$\beta_3 = 4$	1	-3	8	8

ability of random effects σ^2 increases.

- Increasing the sample size reduces the mean-squared errors of estimates for both EX and AP.
- Increasing the number of intra-individual measurements reduces biases and mean-squared errors of estimates for both EX and AP.
- AP yields somewhat smaller mean-squared errors than EX and thus provides more stable results. This is probably because sampling the random effect in the EX, may lead to unstable Gibbs samplers and thus induce more Monte Carlo errors.

Note that the convergence rate of EX is approximately as fast as that of AP in our simulations probably due to the fact that only one random effect is included in our GLMMs.

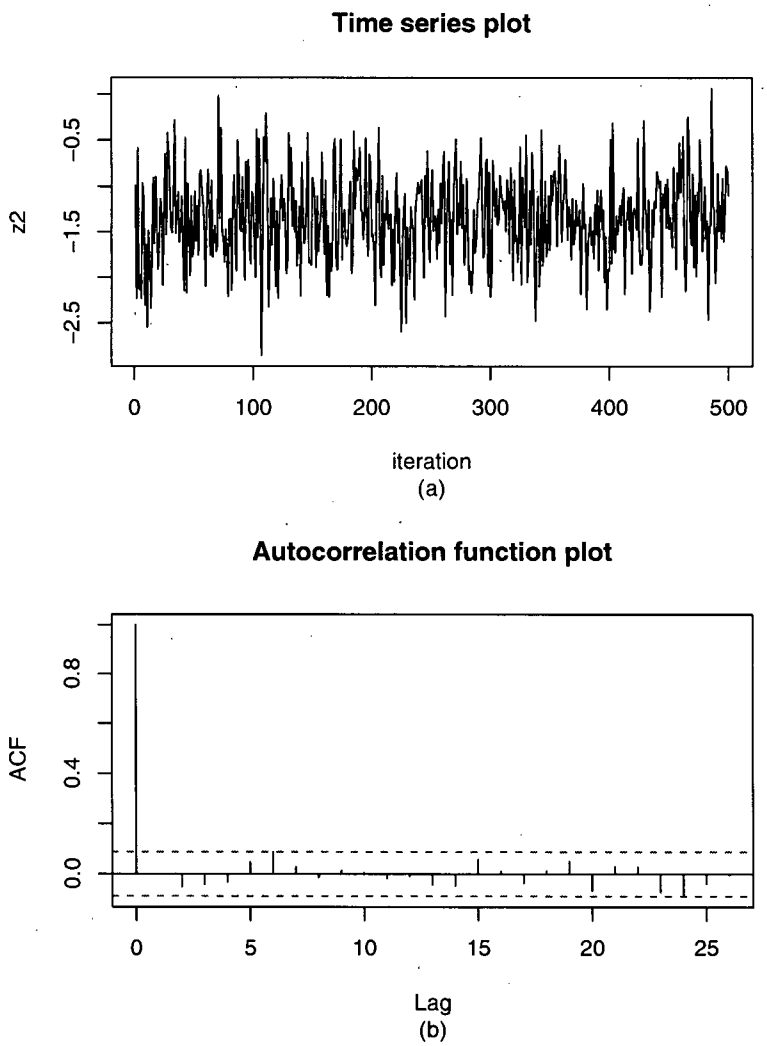

Figure 7.1: (a) Time series and (b) autocorrelation function plots for z_2.

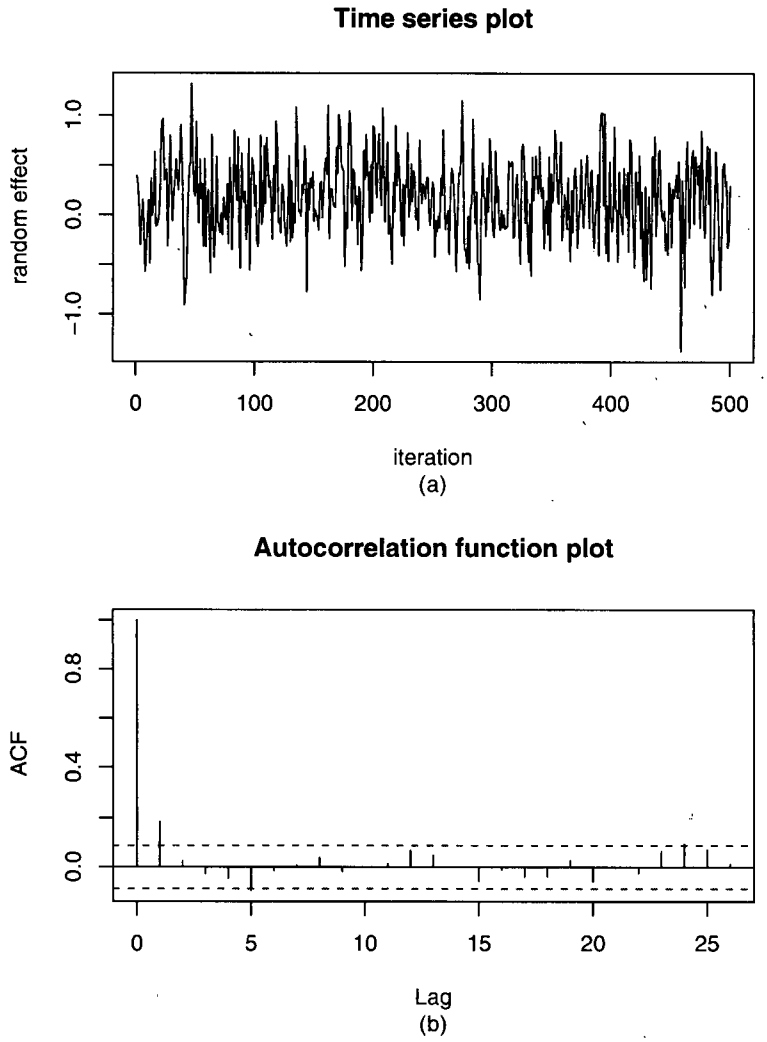

Figure 7.2: (a) Time series and (b) autocorrelation function plots for b_{18} associated with individual 18.

Chapter 8

Conclusion and Discussion

In this thesis, we have proposed two methods to estimate the parameters for GLMMs with informative dropouts and missing covariates. These include an exact method and an approximate method, which are implemented by the Monte Carlo EM algorithm. For the exact method, the conditional expectation in the E-step of the Monte-Carlo EM is evaluated by Monte Carlo approximations (Wei and Tanner, 1990), which generate random samples for the unobservable random effects, missing covariates, and missing responses. However, sampling the random effects may offer potential computational difficulties such as very slow or non-convergence, especially when the dimension of the random effects is not small. To overcome this difficulty, in the more efficient approximate method, we integrate out the random effect in the E-step and thus avoid sampling the random effects in the Monte Carlo EM. Pinheiro and Wu (2001) have shown that the convergence rate of the EM algorithm can be greatly improved by integrating out the random effects.

To further speed up the Monte Carlo EM, we also applied a PX-EM algorithm, which accelerates the EM algorithm by introducing additional working parameters to the model. Based on our two examples, the PX-EM algorithm is much faster than the

standard EM algorithm.

We conducted a simulation study to compare the performance of the exact method and the approximate method. In our simulations, in general, the approximate method gives somewhat more stable results than the exact method in the sense that it provides smaller mean-squared errors. As the number of intra-individual measurements or the sample size increases, the performance of the approximate method and the exact method becomes similar. Our simulations also suggest that the proportion of missing values, the variance of random effects, the sample size, the number of intra-individual measurements, may affect the performance of the exact method and the approximate method.

The proposed methods were applied to an AIDS dataset to evaluate an antiviral treatment. The results of our analyses based on the exact and approximate methods suggest that the viral loads of HIV patients tend to decrease with time, and that patients with higher CD4 cell counts are more likely to have their viral loads suppressed below the detection limit. We also applied our methods to a data set from a parent bereavement project to investigate the change of parents' mental distress after their children's death and to determine which factors influence parents' mental distress. We conclude that parents with a higher education level are more likely to have a better adjustment to their children's death.

Note that we have assumed parametric models for the missing covariates and missing response indicators. So it is important to conduct sensitivity analyses of our results to these parametric models. Based on our sensitivity analyses, except for β_3 in the second example, the results of the two examples are quite robust to the choices of the covariate model and the dropout model. Thus these results except for β_3 in the second example may be reliable.

Finally, we give an outline for possible future work.

(i) For simplicity, in our examples and simulations, we include only one random effect in the GLMMs to demonstrate our methods. In the future, we will study models with more random effects and further investigate the computational advantages/disadvantages of the proposed methods, via simulations.

(ii) In our examples and simulations, we only consider mixed effect logistic regression models with informative dropouts and missing covariates. Generally, our proposed methods can be applied to other GLMMs, such as mixed-effect Poisson models, and nonlinear mixed effect models with informative dropouts and missing covariates.

(iii) In the thesis, we assume that covariates with missing values are time-independent. When some covariates with missing values are time-dependent, similar methods can be proposed.

(iv) We have assumed that the missing responses depend on the values being missing. We could also apply our methods to shared-parameter models, in which the missingness of responses is assumed to depend on the unobservable random effects.

Bibliography

[1] Agresti, A. *Categorical Data Analysis.* New York: John Wiley, 1990.

[2] Booth, J. G. and Hobert, J. P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Soc. B*, 61:265–285, 1999.

[3] Breslow, N. E. and Clayton, D. G. Imcomplete data in generalized linear models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[4] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood estimation from incomplete data via the the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Soc. B*, 39:1–38, 1977.

[5] Diggle, P. and Kenward, M. G. Informaive Drop-out in Longitudinal Data Analysis. *Apllied Statistics*, 43:49–93, 1994.

[6] Diggle, P. J., Liang, K. Y., and Zeger, S. L. *Analysis of Longitudinal Data.* Oxford: Oxford University Press, 1994.

[7] Gilks, W. R. and Wild, P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[8] Ibrahim, J. G. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85:765–769, 1990.

[9] Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88:551–564, 2001.

[10] Ibrahim, J. G. and Lipsitz, S. R. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Soc. B*, 61:173–190, 1999.

[11] Lipsitz, S. R. and Ibrahim, J. G. A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83:916–922, 1996.

[12] Little, R. J. A. Regression with missing X's: A review. *Journal of the American Statistical Association*, 87:1227–1237, 1992.

[13] Little, R. J. A. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

[14] Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data.* New York: John Wiley, 1987.

[15] Little, R. J. A. and Schlucher, M. D. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72:497–512, 1985.

[16] Liu, C. and Rubin, D. B. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648, 1994a.

[17] Liu, C., Rubin, D. B., and Wu, Y. N. Parameter expansion to accerlate EM: The PX-EM algorithm. *Biometrika*, 85:755–770, 1998.

[18] Louis, T. A. Finding the observed information matix when using the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 44:226–233, 1982.

[19] McCulloch, C. E. Maximum likelihood algorithm for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.

[20] McCullock, C. E. and Searle, S. R. *Generalized, Linear, and Mixed Models.* New York: Wiley, 2001.

[21] Meng, X. L. and Van Dyk. The EM algorithm - an old folk song sung to a fast new tune (with Discussion). *Journal of the Royal Statistical Society, Soc. B*, 59:511–567, 1997.

[22] Pinheiro, J. C. and Wu, Y. Efficient algorithms for bobust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, 10:249–276, 2001.

[23] Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, 97:271–283, 2002.

[24] Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447, 1974.

[25] Wei, G. C. and Tanner, M. A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

[26] Wolfinger, R. Laplace's approximation for nonlinear mixed models. *Journal of the American Statistical Association*, 80:791–795, 1993.

[27] Wu, H. and Ding, A. Population HIV-1 Dynamics in Vivo: Applicable Models and Inferential Tools for Virological Data from AIDS Clinical Trials. *Biometrics*, 55:410–418, 1999.

[28] Wu, H. and Wu, L. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. *Statistics in medicine*, 20:1755–1769, 2001.

[29] Wu, M. C. and Carroll, R. J. Estimation and comparision of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44:175–188, 1988.