A Continuous Latent Factor Model for Non-ignorable Missing Data in Longitudinal Studies

by

Jun Zhang

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved November 2013 by the
Graduate Supervisory Committee:

Mark Reiser, Co-Chair
Jarrett Barber, Co-Chair
Ming-Hung Kao
Jeffrey Wilson
Robert D. St Louis

ARIZONA STATE UNIVERSITY

December 2013

ABSTRACT

Many longitudinal studies, especially in clinical trials, suffer from missing data issues. Most estimation procedures assume that the missing values are ignorable or missing at random (MAR). However, this assumption leads to unrealistic simplification and is implausible for many cases. For example, an investigator is examining the effect of treatment on depression. Subjects are scheduled with doctors on a regular basis and asked questions about recent emotional situations. Patients who are experiencing severe depression are more likely to miss an appointment and leave the data missing for that particular visit. Data that are not missing at random may produce bias in results if the missing mechanism is not taken into account. In other words, the missing mechanism is related to the unobserved responses.

Data are said to be non-ignorable missing if the probabilities of missingness depend on quantities that might not be included in the model. Classical pattern-mixture models for non-ignorable missing values are widely used for longitudinal data analysis because they do not require explicit specification of the missing mechanism, with the data stratified according to a variety of missing patterns and a model specified for each stratum. However, this usually results in under-identifiability, because of the need to estimate many stratum-specific parameters even though the eventual interest is usually on the marginal parameters. Pattern mixture models have the drawback that a large sample is usually required.

In this thesis, two studies are presented. The first study is motivated by an open problem from pattern mixture models. Simulation studies from this part show that information in the missing data indicators can be well summarized by a simple continuous latent structure, indicating that a large number of missing data patterns may be accounted by a simple latent factor. Simulation findings that are obtained in the first study lead to a novel model, a continuous latent factor model (CLFM).

The second study develops CLFM which is utilized for modeling the joint distribution of missing values and longitudinal outcomes. The proposed CLFM model is feasible even for small sample size applications. The detailed estimation theory, including estimating techniques from both frequentist and Bayesian perspectives is presented. Model performance and evaluation are studied through designed simulations and three applications. Simulation and application settings change from correctly-specified missing data mechanism to mis-specified mechanism and include different sample sizes from longitudinal studies. Among three applications, an AIDS study includes non-ignorable missing values; the Peabody Picture Vocabulary Test data have no indication on missing data mechanism and it will be applied to a sensitivity analysis; the Growth of Language and Early Literacy Skills in Preschoolers with Developmental Speech and Language Impairment study, however, has full complete data and will be used to conduct a robust analysis. The CLFM model is shown to provide more precise estimators, specifically on intercept and slope related parameters, compared with Roy's latent class model and the classic linear mixed model. This advantage will be more obvious when a small sample size is the case, where Roy's model experiences challenges on estimation convergence. The proposed CLFM model is also robust when missing data are ignorable as demonstrated through a study on Growth of Language and Early Literacy Skills in Preschoolers.

I would like to dedicate my thesis to Vivian, and my beloved family.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and guidance of my fantastic advisors, Dr. Reiser and Dr. Barber. They were selfless in their willingness to respond quickly to all of my questions and to meet with me whenever I needed their assistance. I am indebted to them for significant contribution of my graduate education, including encouragement to travel to JSM in San Diego, help with job hunting. They made a great team as co-advisors.

I am also very grateful to my other committee members - Dr. Kao, Dr. Wilson, and Dr. St Louis - for their valuable time and feedback, as well as their mentorship with coursework and professional development. I also would like to acknowledge all facuty members who supported me in my coursework. All of these faculty members from Arizona State have been extremely friendly and willing to share their knowledge. My heartful acknowledgement also goes to my graduate statistics club fellows. Their endless help made my graduate study more memorable. Finally, I need to thank the graduate coordinator, Debbie Olson, for her help in many administrative tasks.

Many thanks also go to the most important persons in my life who have helped me along the path to a Ph.D. I would like to thank my beloved parents and younger brother for their tremendous love and support. Lastly, and most importantly, I wish to thank my fiancee, Vivian Zhou, who has given me a great deal of help and encouragement as a colleague throughout the PhD program.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Missing values in multivariate studies pose many challenges. The primary research of interest focuses on accurate and efficient estimation of means and covariance structure in the population. The assumption and estimation of the covariance structure provide the foundation of many statistical models, for instance, structural equation modeling, principle component analysis, and so on. Literature on multivariate missing data methods was reviewed by Little and Rubin (2002) and Schafer (1997). For some frequentist statistical procedures, we may generally ignore the distribution of missingness only when the missing data are missing completely at random (MCAR), such as in the generalized estimation equations (GEE) estimation procedure. For likelihood or Bayes procedures, however, we may ignore the missing values when the missing data are missing at random (MAR), as in for example, the estimation procedure for linear mixed models. The Expectation Maximization (EM) algorithm for missing data (Dempster *et al.*, 1977) produces maximum likelihood (ML) estimation under an assumption of normality. A useful alternative to the ML approach is multiple imputation (MI) (Rubin, 1987; Schafer, 1997). In MI, each missing value is substituted with a set of plausible simulated values, which represent uncertainty about the missing data. The multiple imputed dataset can be analyzed by standard complete data methods, and model estimation and inference can be investigated from combined results. Apparently, the EM algorithm generates estimates of model-specific parameters. The EM algorithm and MI methods have also been extended to non-normal models, including log-linear models for multivariate categorical data, the general location model for mixed datasets containing both continuous and categorical data, and a multivariate linear mixed-effects model for multivariate panel data or clustered data (Schafer, 1997).

The above mentioned approaches to ML estimation are invariably applied under the assumption that the missing values in the dataset are MAR. The underlying meaning of MAR is that the probability of missing values may be related to observed data or covariates, but are conditionally independent of all missingness given the observed responses. However, this assumption is always challenged and can not be reasonable in some applications. The reason is that missing values are sometimes thought to depend on the values themselves. For instance, individuals may refuse to answer sensitive items (e.g. income or health history) on a questionnaire, and the missing value would be related to the underlying true values for those items; or in clinical trial, the dropout from a patient may be strongly related to outcome.

If missing at random in the data is questioned, and one suspects that the missing mechanism is NMAR, i.e. missingness may depend on missing values, then the joint modeling of the complete data and the missing indicators is required. The reason to follow this modeling method is that the resulting estimates of population parameters may be biased (Pirie and Leupker, 1988) unless these NMAR aspects of the data are taken into account in the analysis. Furthermore, the results of the study may not be feasible to generalize, because the observed respondents may not represent the target population.

From a practical aspect, investigators could not point out whether violations of the MAR assumption are severe enough to result in a conclusions that are not valid. It is also worthwhile to investigate how the results may change under different assumptions even if the primary analysis proceeds under an assumption of MAR. A standard ignorable analysis can be strengthened by a sensitivity analysis that includes non-ignorable alternatives. Results will be more convincing if estimates from different alternative models agree. If they do not agree, the differences afford a better sense of the true levels of uncertainty.

Models for NMAR data have been proposed for a few decades, including selection models (Diggle and Kenward, 1994a), pattern-mixture models (Diggle and Kenward, 1994b), as well as shared-parameter models (Diggle and Kenward, 1994b). The detailed review of these models will be given in Chapter 2. All of these forms lead to a rich class of models: latent class models are one of the prevalent members in longitudinal studies. However, the selection of number of latent classes, which is the key assumption for latent class modeling for missingness, is unstable due to many factors as shown by simulation studies presented in Chapter 3. This sensitivity hinders the direct application of the latent class modeling technique, and intensive simulation studies should be performed before applying it to application studies. The primary goal of this research is to develop a general method for non-ignorable modeling of incomplete multivariate data based on the idea of a continuous latent variable (Lord, 1952, 1953; Bock and Aitkin, 1981). We will summarize the distribution of the missingness indicators through a continuous latent factor model, and then relate to the model of interests by including an association of latent traits with subject-specific parameters from the population. A specific description of this new model will be given in Chapter 4.

Chapter 2

STATISTICAL THEORIES AND MODELS FOR NON-IGNORABLE MISSING DATA

When discussing missing data, it is useful to distinguish the missing-data pattern from the missing mechanism. Missing data pattern describes which values are observed in the data matrix and which values are missing; the missing mechanism concerns the relationship between missingness and the variable values in the data frame. Let $\mathbf{Y} = (y_{ij})$ be an $(n \times p)$ data frame, with ith row $\mathbf{Y_i} = (y_{i1}, \ldots, y_{ip})$ where y_{ij} is the value (response) of variable (item) j for subject i. Considering missing data, define the missingness indicator matrix $\mathbf{R} = (r_{ij})$, such that $r_{ij} = 1$ if y_{ij} is missing and $r_{ij} = 0$ if y_{ij} is observed. The matrix $\mathbf{R}$ then defines the pattern of missing data. Little and Rubin (2002) reviewed the theory of missing mechanism. If we keep the same notation for the complete data $\mathbf{Y}$ and missing-data indicator matrix $\mathbf{R}$ and borrow the notation for the unknown parameters θ from complete data model (in longitudinal studies the complete data model is commonly assumed to be a linear mixed model or generalized linear mixed model), as well as ψ from the indicator matrix model $\mathbf{R}$ from Little and Rubin, the missing mechanism is characterized by the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}$, i.e. $f(\mathbf{R}|\mathbf{Y}; \theta, \psi)$. If missingness does not depend on the values of the data frame $\mathbf{Y}$, (either missing or observed), i.e. if

$$f(\mathbf{R}|\mathbf{Y}; \theta, \psi) = f(\mathbf{R}|\psi) \quad for\ all\ \mathbf{Y},\ \theta,\ \psi \tag{2.1}$$

the data are called missing completely at random (MCAR). Further, we can decompose the data frame $\mathbf{Y}$ into the observed part $\mathbf{Y}^{obs}$ and the missing part $\mathbf{Y}^{mis}$, a less restrictive assumption than MCAR is that missingness depends only on the observed values of $\mathbf{Y}$, which is $\mathbf{Y}^{obs}$, and not on the components that are missing. That is,

$$f(\mathbf{R}|\mathbf{Y};\theta,\psi) = f(\mathbf{R}|\mathbf{Y}^{obs};\ \psi)\ \ for\ all\ \mathbf{Y}^{mis},\ \theta,\ \psi \tag{2.2}$$

This assumption will lead to missing at random (MAR). The missing mechanism is called not missing at random (NMAR) if the distribution of $\mathbf{R}$ depends on the missing values in the data matrix $\mathbf{Y}$.

Further, we assume for each individual i, $(i = 1, \ldots, n)$, there is a $(q \times 1)$ vector of covariates $\mathbf{X}_i = (x_{i1}, \ldots, x_{iq})$. To estimate unknown parameters θ and ψ, given fully observed covariate matrix $\mathbf{X} = (x_{ij})$, the general likelihood inferences are based on the observed-data likelihood, which is obtained by integrating the missing data $\mathbf{Y}_i^{mis}$ out of the density of $(\mathbf{Y}_i,\ \mathbf{R}_i)$:

$$L(\theta,\ \psi|\mathbf{R},\ \mathbf{Y}^{obs},\ \mathbf{X}) \propto \prod_{i=1}^{N} \int f(\mathbf{Y}_i,\ \mathbf{R}_i|\mathbf{X}_i;\ \theta,\ \psi) d\mathbf{Y}_i^{mis} \tag{2.3}$$

However, likelihood-based inferences are complicated due to missing data: in the above expression, a model for the joint distribution of $\mathbf{Y}$ and $\mathbf{R}$ is needed, rather than a model for the responses $\mathbf{Y}$. Hence, the data $mathbfY$ and missingness indicator $\mathbf{R}$ are related, and the parameter estimation tends to be sensitive to the assumptions for missing data.

Ignorable likelihood inference is favored by many researchers and many publications have flourished in recent decades (Horton and Fitzmaurice, 2002; Lee and Song, 2003). If we revisit parameters θ and ψ: θ describes model settings (e.g. parameters in a growth curve model) and ψ represents parameters for missingness, the missing-data mechanism is said to be ignorable if (a) the missing data are missing at random (MAR), and (b) the model parameters θ are distinct from missing mechanism parameters ψ, i.e. the joint parameter space of (θ, ψ) is the product of the parameter space of θ and the parameter space of ψ. The ignorable likelihood function can be written as

$$L_{ign}(\theta|\mathbf{Y}^{obs},\ \mathbf{X}) \propto \prod_{i=1}^{N} \int f(\mathbf{Y}_i|\mathbf{X}_i;\ \theta) d\mathbf{Y}_i^{mis} \propto \prod_{i=1}^{N} f(\mathbf{Y}_i^{obs}|\mathbf{X}_i;\ \theta) \tag{2.4}$$

By avoiding computing the integral in the full likelihood (2.3), the ignorable likelihood inference is generally easier to deal with. Furthermore, the ignorable likelihood does not need a model for missingness, which can be difficult to access. For these reasons, most likelihood approaches for incomplete longitudinal data with dropouts or intermittent missingness are based on (2.4) rather than the full likelihood (2.3).

When the ignorable assumptions are not met, one needs to consider the joint distribution of $\mathbf{Y}_i$ and $\mathbf{R}_i$. Depending on the factorization of the joint distribution, two models are widely investigated: selection model and pattern-mixture model. To demonstrate these two scenarios, we start from a fixed-effect model that does not include random effects for subjects.

Selection models factorize the joint distribution of $\mathbf{Y}_i$ and $\mathbf{R}_i$ as models for the marginal distribution of $\mathbf{Y}_i$ and the conditional distribution of $\mathbf{R}_i$ given $\mathbf{Y}_i$:

$$f(\mathbf{Y}_i,\ \mathbf{R}_i \mid \mathbf{X}_i;\ \theta,\ \psi) \ = \ f_Y(\mathbf{Y}_i \mid \mathbf{X}_i;\ \theta)\ f_{R|Y}(\mathbf{R}_i \mid \mathbf{X}_i,\ \mathbf{Y}_i;\ \psi) \tag{2.5}$$

where θ and ψ span the complete parameter space.

Pattern-mixture models specify the marginal distribution of $\mathbf{R}_i$ and the conditional distribution of $\mathbf{Y}_i$ given $\mathbf{R}_i$:

$$f(\mathbf{Y}_i,\ \mathbf{R}_i \mid \mathbf{X}_i;\ \theta,\ \psi) \ = \ f_R(\mathbf{R}_i \mid \mathbf{X}_i;\ \psi)\ f_{Y|R}(\mathbf{Y}_i \mid \mathbf{X}_i,\ \mathbf{R}_i;\ \theta). \tag{2.6}$$

where $(\theta,\ \psi)$ is the whole parameter space.

However, longitudinal studies require within-subject random effects $\mathbf{b}_i$ in most cases. Little (1995) points out that with NMAR data, the selection and pattern-mixture formulations can be expanded to allow the possibility that the missing-data mechanism depends on latent random effects. The mixed-effect selection models can be expressed as:

$$f(\mathbf{Y}_i,\ \mathbf{R}_i,\ \mathbf{b}_i \mid \mathbf{X}_i;\ \theta,\ \psi,\ \delta) \ = \ f(\mathbf{b}_i \mid \mathbf{X}_i;\ \delta)\, f(\mathbf{Y}_i \mid \mathbf{X}_i,\ \mathbf{b}_i;\ \theta)\, f(\mathbf{R}_i \mid \mathbf{X}_i,\ \mathbf{Y}_i,\ \mathbf{b}_i;\ \psi). \tag{2.7}$$

and mixed-effect pattern-mixture models have the form

$$f(\mathbf{Y}_i,\ \mathbf{R}_i,\ \mathbf{b}_i \mid \mathbf{X}_i;\ \theta,\ \psi,\ \delta) \ = \ f(\mathbf{R}_i \mid \mathbf{X}_i;\ \psi)\, f(\mathbf{b}_i \mid \mathbf{X}_i,\ \mathbf{R}_i;\ \delta)\, f(\mathbf{Y}_i \mid \mathbf{X}_i,\ \mathbf{b}_i,\ \mathbf{R}_i;\ \theta). \tag{2.8}$$

The independence assumption of $\mathbf{Y}_i$ and $\mathbf{R}_i$ given latent variables $\mathbf{b}_i$ yields the shared-parameter models: (Roy, 2003, 2007)

$$f(\mathbf{Y}_i,\ \mathbf{R}_i,\ \mathbf{b}_i \mid \mathbf{X}_i;\ \theta,\ \psi,\ \delta) \ = \ f(\mathbf{b}_i \mid \mathbf{X}_i;\ \delta)\ f(\mathbf{Y}_i \mid \mathbf{X}_i,\ \mathbf{b}_i;\ \theta)\ f(\mathbf{R}_i \mid \mathbf{X}_i,\ \mathbf{b}_i;\ \psi). \tag{2.9}$$

As the above factorizations leading to a rich class of models, latent class models (Goodman, 1978; Clogg, 1995), however, are one of the prevalent members and are discussed with missing-data patterns in longitudinal studies (Roy, 2007). A review of work for latent class models is given in the next chapter when a simulation study for these models is presented.

2.1 Latent Class Models and their Application to Missing-data Patterns in Longitudinal Studies

Roy (2003, 2007) considers using latent class models to describe intermittent missing-data patterns in longitudinal studies. Define latent class variable S as a categorical variable that can take values $\{1, \dots, M\}$. As an alternative, Roy et al. factorize the joint distribution of Y_i and R_i for subject i in expression 2.8 as:

$$f(\mathbf{Y}_i,\ \mathbf{R}_i \mid \mathbf{X}_i;\ \delta,\ \nu) \ = \sum_S f(\mathbf{R}_i \mid \mathbf{X}_i;\ \delta)\ f(S \mid \mathbf{R}_i;\ \nu_1)\ f(\mathbf{Y}_i \mid \mathbf{X}_i,\ S;\ \nu_2).$$

A series of surrogate measures (the time of the last observed value; the number of observed values; the number of transitions; etc.) for missing patterns are specified and used for modeling latent variable S in a logistic regression model, when assuming latent class variable S is distributed as multinomial and the observation process R is given. An ordinal logistic regression is applied to model observation process R_i given covariates X_i. With the assumption that conditional on latent class variable S the missing outcomes can be ignored, (i.e. given a set of subjects with similar observation patterns (given S) and similar covariates (given X_i), the observation process might no longer be informative about the missing outcomes.) The observed outcomes Y_i are modeled conditional on S and X_i. The following linear mixed-effects model is proposed in Roy's paper:

$$Y_{ij} \;=\; \{\sum_{s=1}^{M} X_{ij}^T \beta_s I(S_i = s)\} \;+\; Z_{ij}^T b_i \;+\; \epsilon_{ij} \tag{2.10}$$

In this model, $\beta_s,\ s = 1, \ldots, M$, is a class-specific regression coefficient vector, Z_{ij} is a vector of covariates, b_i are subject-specific random effects and ϵ_{ij} is the error term. The following assumptions were made: $b_i \sim N\{0,\ D(\theta)\}$, and independent of $\epsilon_{ij} \sim N(0,\ \sigma^2)$. The random effects covariance matrix D is parameterized by a vector of variance components θ.

As pointed out by Roy, while the regression coefficients for the sth class may be of substantive interest, generally marginal covariate effects will be of primary interest. The estimation of marginal covariate effects is carried out by averaging over the latent class distribution. Let ω be the vector of all model parameters, i.e. $\omega = (\theta, \psi)$, the likelihood function in Roy's model can be factored as

$$L(\omega;\ Y, S, R, X) = \prod_{i=1}^{n} [\ f(R_i|X_i;\omega)\ \sum_{s=1}^{M}\ f(Y_i|S_i = s, X_i;\omega) f(S_i = s|R_i;\omega)] \tag{2.11}$$

Estimation and inference can be based on this likelihood function.

In general, several areas in latent class modeling need to be carefully researched. The models rely on several assumptions that might be difficult to check. For example, one challenge is selecting the number of latent class, S. Some authors have suggested criterion such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) as a way of comparing models with different number of classes. Systematic Monte Carlo simulation studies on this model selection have been done through our work, and will be presented in next chapter. Roy suggests in his paper that one would like, ideally, to choose the smallest number of classes that allows the conditional independence assumption to hold. Having a model with too many classes can be a problem, as one or more of the class might be very small and hard to interpret. However, researchers have found that information in a set of missingness indicators is sometimes well summarized by a simple latent factor structure, indicating that a large number of missing patterns may be reduced to a few prototypes. Latent class modeling technique has been applied for non-ignorable modeling of incomplete multivariate data where factorization of a selection model is considered (Jung and Seo, 2011).

2.2 Latent Class Selection Model for Non-ignorable Missing Data

The selection model for non-ignorable missing data is often overlooked due to its instability and extreme sensitivity. Jung and Seo (2011) improved selection models by adopting a latent-class approach to modeling patterns of missingness. Rather than using an incomplete response vector to predict the probability of missingness for that item directly, they use this response vector to predict class membership, so that items and missingness are related only through latent classes. Here we keep the same notation and define a single column vector z_i which contains the y_{ij}'s, the x_{ij}'s and a constant term. The missing-data mechanism in Jung's model is

$$\begin{aligned} f(\mathbf{R}_i = \mathbf{r}_i \mid \mathbf{Z}_i = \mathbf{z}_i;\ \beta,\ \rho) &= \textstyle\sum_{l=1}^{C} f(\mathbf{R}_i = \mathbf{r}_i \mid L_i = l) P(L_i \mid Z_i = z_i) \\ &= \textstyle\sum_{l=1}^{C} \pi_l(z_i) \prod_{j=1}^{p} \rho_{j|l}^{1-r_{ij}} (1-\rho_{j|l})^{r_{ij}} \end{aligned} \tag{2.12}$$

where

$$\pi_l(z_i) = \frac{exp(z_i^T \beta_l)}{1 + \sum_{j=1}^{C-1} exp(z_i^T \beta_j)},$$

and $\rho_{j|l}$ is the conditional probability that an individual responds to item y_{ij} given that $L_i = l$.

The β and ρ parameters in the above model are nuisance since the questions of research interest usually pertain to the population of $\mathbf{Y}_i$. Let μ be parameters of the population distribution of $\mathbf{Y}_i$, and collect the missing data indicators for all subjects into a matrix $\mathbf{R}$, and the $\mathbf{Y}_i$'s and $\mathbf{X}_i$'s into another matrix $\mathbf{Z}$, the likelihood function for this model becomes

$$L(\psi \mid \mathbf{Z},\ \mathbf{R}) \propto \prod_{i=1}^{n} \left[f(z_i \mid \mu) \sum_{l=1}^{C} \pi_l(z_i) \prod_{j=1}^{p} \rho_{j|l}^{1-r_{ij}} (1-\rho_{j|l})^{r_{ij}} \right] \qquad (2.13)$$

where $\psi = (\mu,\ \beta,\ \rho)$ represents all parameters of the population model and the missingness mechanism. However, the above likelihood function cannot be used for inference because it depends on the missing items in $\mathbf{Y_i}$. By integrating out the missing items $\mathbf{Y}^{mis}$ in responses $\mathbf{Y}$, the following likelihood can be used in practice for observed responses $\mathbf{Y}^{obs}$:

$$L(\psi \mid \mathbf{Y}^{obs},\ \mathbf{X},\ \mathbf{R}) \propto \prod_{i=1}^{n} \left[\int f(z_i \mid \mu) \sum_{l=1}^{C} \pi_l(z_i) \prod_{j=1}^{p} \rho_{j|l}^{1-r_{ij}} (1-\rho_{j|l})^{r_{ij}} d\mathbf{Y}^{mis} \right] \qquad (2.14)$$

Similar to conventional latent class models, such as Roy's model, one of the important practical modeling issues in using the latent class selection models (LCSM) is to determine a proper number of latent classes. In the LCSM, too many classes may destabilize the posterior predictive distribution of $\mathbf{Y}^{mis}$, producing unstable inferences about the complete data population. Too few classes will produce a model that fails to adequately capture the relationships between the complete data and the missingness indicators. Latent class models and LCSM often result in under-identifiability due to many latent class-specific parameters even though the eventual interest is usually on the population-averaged parameters. Guo *et al.* (2004) extend the pattern mixture models to a random pattern mixture model, where the pattern-specific parameters are treated as nuisance parameters and modeled as random instead of fixed.

2.3 A Random Pattern-Mixture Model

Pattern mixture models (model (2.6), or (2.8)) are one of most popular models for longitudinal studies with non-ignorable missingness. This modeling technique stratifies the data according to missing patterns (e.g. dropout patterns), and forms a model for each stratum. The final estimate is a weighted average of the stratum-specific estimates. In pattern-mixture models, missing patterns could be well summarized by some surrogate measures, and conditional on the pattern, the missing mechanism is ignorable within a stratum. Hence information from the complete cases can be borrowed to predict the incomplete cases. However, a full pattern-mixture model usually has an over-parameterization issue. Guo et al. (2004) proposed a random pattern-mixture model by generalizing the definition of pattern and applied this model in a longitudinal study with dropouts. The pattern is defined based on a good surrogate for the dropout process which can be a baseline or time-varying covariate, or time to dropout. In this model, it is assumed that conditional on the latent pattern effects, the longitudinal outcome and the dropout process are independent.

The random pattern mixture model combines the features of the selection models (2.7) and fixed pattern-mixture models (2.8). Assuming that data can be stratified into m strata based on a surrogate for the dropout process, pattern effects are modeled as random effects and used to link responses $\mathbf{Y}$ and missing indicators $\mathbf{R}$. The random pattern mixture model implies the following factorization

$$\begin{aligned} & f(\mathbf{Y},\ \mathbf{R},\ \mathbf{b}|\mathbf{X}) \\ &\quad = \int f(\mathbf{Y}|\mathbf{X},\mathbf{b},\mathbf{u},\mathbf{R}) f(\mathbf{b}|\mathbf{X},\mathbf{u},\mathbf{R}) f(\mathbf{R}|\mathbf{X},\mathbf{u}) f(\mathbf{u}|\mathbf{X}) d\mathbf{u} \qquad (2.15) \\ &\quad = \int f(\mathbf{Y}|\mathbf{X},\mathbf{b},\mathbf{u}) f(\mathbf{b}|\mathbf{X},\mathbf{u}) f(\mathbf{R}|\mathbf{X},\mathbf{u}) f(\mathbf{u}|\mathbf{X}) d\mathbf{u} \end{aligned}$$

where $\mathbf{u}$ is the random pattern effects and $f(\mathbf{R}|\mathbf{X},\mathbf{u})$ gives the distribution of $\mathbf{R}$. The simplification in the above formula implies that $\mathbf{Y}$ and $\mathbf{b}$ depend on $\mathbf{R}$ through the random pattern effects $\mathbf{u}$. Further this model borrows the basic idea of stratification from the fixed pattern mixture model, i.e., conditional on the latent pattern effects, $\mathbf{u}$, the missing mechanism is ignorable within a stratum, and parameters of interest are the marginal estimates averaging over the latent pattern effects. This model has similar computational difficulty as the shared-parameter models because of the need to integrate over $\mathbf{u}$ and $\mathbf{b}$. To avoid the computational difficulty, the joint normal distribution is considered in Guo's paper (2004). They model the random effects as normally distributed and the outcome and dropout times as multivariate normal, as defined below.

Assume that data is stratified into m strata based on a selected surrogate, and consider the cases where subject j is nested within the ith stratum. Let $\mathbf{y}_{ij}$ be an n_{ij} vector of observed outcomes for the jth subject within the ith pattern, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Let r_{ij} be the corresponding dropout time for this subject (surrogate for the dropout process). Guo et al. (2004) modeled both the responses and the dropout times using mixed-effects models, with linking the two models by the random pattern effects. Then Guo's model could be expressed as follows

$$\mathbf{y}_{ij} = \mathbf{X}_{1ij}\alpha_1 + \mathbf{Z}_{ij}\mathbf{b}_{ij} + \mathbf{W}_{ij}\mathbf{u}_i + \mathbf{e}_{ij} \qquad (2.16)$$

and

$$r_{ij} = \mathbf{x}_{2ij}^T\alpha_2 + \beta^T u_i + \epsilon_{ij} \qquad (2.17)$$

where $\mathbf{X}_{1ij}$, $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ are the known design matrices for the fixed effects, subject level random effects, and pattern level random effects for $\mathbf{y}_{ij}$; α_1 is the vector of unknown fixed effects; $\mathbf{b}_{ij} \sim N(\mathbf{0}, \Sigma_{\mathbf{b}})$ is the unknown subject level random effects; $\mathbf{e}_{ij} \sim N(\mathbf{0}, \sigma^2_{\mathbf{e}}\mathbf{I}_{ij})$ is the residual term; $\mathbf{x}_{2ij}$ is the known design matrix linking the unknown parameter α_2 to r_{ij}; β is an unknown parameter vector linking u_i to r_{ij}; and $\epsilon_{ij} \sim N(0, s^2)$ is the residual for r_{ij}. The vector of parameters to be estimated in the model is $\theta = (\alpha_1, \alpha_2, \beta, \Sigma_{\mathbf{b}}, \Sigma_{\mathbf{u}}, \sigma^2_{\mathbf{e}}, s^2)$. By treating the pattern level random effects u_i and subject level random effects $\mathbf{b}_{ij}$ as missing data, an EM algorithm can be applied to calculate the maximum likelihood estimates of the above parameters. The complete data log-likelihood for the EM part of the algorithm is

$$l_c = \sum_{i=1}^{m} \sum_{j=1}^{n_i} log\ \phi(\mathbf{y}_{ij}|\mathbf{X}_{ij}, u_i, \hat{\theta})\phi(u_i|\hat{\theta})\phi(\mathbf{r}_{ij}|\mathbf{x}_{ij}, u_i, \hat{\theta})$$

where $\phi(\cdot)$ is the normal density function.

Instead of considering the joint inference of responses and missing patterns, Guo et al. (2004) redefined missing patterns (dropout) by surrogate variables, such as baseline or time varying covariates or time to dropout. Based on a good surrogate, pattern effects are defined and treated as a random variable. Data can be then divided into different strata according to the pattern effects. With the assumptions that the missing mechanism is ignorable within a stratum conditional on the latent pattern effects, and the responses $\mathbf{Y}$ are independent of missing process $\mathbf{R}$ given the random pattern effects, the joint distribution of responses and measures for the missing process (e.g. time to dropout) are modeled through a random pattern mixture model. For computational complexity reasons, In this paper, they only researched on the case that multivariate normal distribution is assumed for responses and missing process measures. In most real studies, however, it maybe impossible to find good measures for the missing mechanism. For instance, in a longitudinal study with many intermittent missing values, time to dropout is not necessarily a good measure, and it probably would not capture most features of missingness. Further, models other than the normal distribution will be required to describe the missing process. The violation of joint multivariate normality will lead to an increase of computation difficulties. We will extend the random pattern mixture models in Chapter 4 with general distributions.

Chapter 3

MONTE CARLO STUDY FOR LATENT CLASS MODEL

Latent class modeling now is wildly used and frequently appearing in medical and statistical journals. A potential application of latent class models (LCM) is for exploring missing data pattern (dropouts or intermittent missing) in longitudinal studies. In the intermittent missing cases, missing-data patterns could have many forms and the effects from missing patterns might be difficult to assess. For instance, in a series of depression studies described in Roy (2007), patients were randomly assigned to receive either drug plus psychotherapy or psychotherapy alone. Data were collected weekly during that period of 17 weeks including baseline. As mentioned in Roy's work, data at baseline were completely collected, but there was a large quantity of missing data afterwards. There were 379 unique missing-data patterns that were observed.

Latent class models with 3 latent classes were used by Roy to assess whether subjects from different missing-data patterns had different responses on the changes in depression over time. However, one of difficulties, also a key condition for using latent class models, is deciding the correct number of latent classes. Garrett and Zeger (2000) suggested using graphical methods for selecting the number of classes. Some researchers also proposed a Bayesian approach to select the number of latent classes by specifying a prior for the number of classes. One could select the model with the highest posterior probability for that number of classes. As the first contribution of this dissertation, we perform Monte Carlo simulation studies and investigate selection of the appropriate number of latent classes via conventional information criteria. Besides, in this particular study latent class models are used for modeling missing patterns and these patterns will highly related with a specific model structure. That is, different model structures will influence on the selection for latent class models. Hence, risk factors that related with model structures as well as missing values will be worthy to study, which includes mixture missing mechanisms in the study, number of covariates involved in the model, degree of correlation among repeated measure, as well as different magnitudes of missing probabilities. In Section 3.1, we make a short review of latent class models. In Section 3.2, we first give a brief description of simulation studies, then elaborate methods and models that are used in longitudinal simulation studies. We present and analyze simulation outputs in Section 3.3, with conclusion and discussion in Section 3.4.

3.1 Review of Latent Class Models

Lazarsfeld and P.F. (1950b,a) first proposed latent class models as a tool for building typologies based on observed dichotomous variables. The basic idea underlying LCM is some parameters of a postulated statistical model differ across unobserved subgroups. These subgroups form the categories of a categorical latent variable.

Let π_{ks} be the probability of a positive response on variable k for a person in category s $(k = 1, 2, \ldots, p;\; s = 0, 1, \ldots, M)$ and let η_s be the prior probability that a randomly chosen individual is in class s which satisfies $\sum_{s=0}^{M} \eta_j = 1$. For the case of M latent classes, the distribution of an individual responses becomes

$$f(\mathbf{x}) = \sum_{s=0}^{M} \eta_s \prod_{k=1}^{p} \pi_{ks}^{I(x_k=1)} (1 - \pi_{ks})^{1-I(x_k=1)} \tag{3.1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ is the response vector of an individual, $I(\cdot)$ is an indicator function and $x_k = 1$ represents a positive response on variable k. The posterior probability that an individual with response vector $\mathbf{x}$ belongs to category s is thus

$$h(s|\mathbf{x}) = \eta_s \prod_{k=1}^{p} \pi_{ks}^{I(x_k=1)} (1 - \pi_{ks})^{1-I(x_k=1)} / f(\mathbf{x}) \quad (s = 1, 2, \ldots, M) \tag{3.2}$$

We can use (3.2) to construct an allocation rule according to which an individual is placed in the class for which the posterior probability is greatest. The principle statistical task is to estimate parameters and testing goodness of fit. On the substantive side the main problem is to identify the latent classes, i.e. to interpret them in terms which make practical sense.

The parameter estimation could be found by maximum likelihood approaches. The log-likelihood function derived from (3.1) is complicated, but it can be maximized using standard optimization routines. McHugh (1956) showed the standard Newton-Raphson technique to solve this optimization problem. However, an easier method which enables larger problems to be tackled is offered by the EM algorithm. The fundamental reference for EM is Dempster *et al.* (1977) supplemented by Wu (1983), but the EM algorithm for latent class model was given by Goodman (1978). From (3.1) the log-likelihood with sample of size n is

$$l = \sum_{i=1}^{n} log\{\sum_{s=1}^{M} \eta_s \prod_{k=1}^{p} \pi_{ks}^{I(x_{ik}=1)} (1-\pi_{ks})^{1-I(x_{ik}=1)}\} \tag{3.3}$$

This log-likelihood function has to be maximized subject to $\sum \eta_s = 1$. Bartholomew (1987) found the parameter estimation in latent class model by taking partial derivatives:

$$\hat{\eta}_s = \sum_{i=1}^{n} h(s|\mathbf{x_i})/n \tag{3.4}$$

$$\hat{\pi}_{ks} = \sum_{i=1}^{n} x_{ik} h(s|\mathbf{x}_i)/n\hat{\eta}_s \tag{3.5}$$

where $k = 1, 2, \ldots, p;\ s = 1, 2, \ldots, M$.

By realizing that $h(s|\mathbf{x}_i)$ is a complicated function of $\{\eta_s\}$ and $\{\pi_{ks}\}$, which is given by

$$h(s|\mathbf{x}_i) = \eta_s \prod_{k=1}^{p} \pi_{ks}^{I(x_{ik}=1)} (1-\pi_{ks})^{1-I(x_{ik}=1)} / \sum_{s=1}^{M} \eta_s \prod_{k=1}^{p} \pi_{ks}^{I(x_{ik}=1)} (1-\pi_{ks})^{1-I(x_{ik}=1)} \tag{3.6}$$

However, if $h(s|\mathbf{x}_i)$ were known it would be easy to solve (3.4) and (3.5) for $\{\eta_s\}$ and $\{\pi_{ks}\}$. The EM algorithm could be applied on this fact by the following steps:

Step 1: choose an initial set of posterior probabilities $\{h(s|\mathbf{x}_i)\}$;

Step 2: update (3.4) and (3.5) to obtain a first approximation to $\{\eta_s\}$ and $\{\pi_{ks}\}$;

Step 3: substitute $\hat{\eta}_s$ and $\hat{\pi}_{ks}$ estimates into (3.6) to obtain improved estimates of $\{h(s|\mathbf{x}_i)\}$;

Step 4: return to step 2 to obtain second approximations to the parameters and continue the iteration until convergence is attained.

With the feasible and efficient estimating techniques, latent class models have been proposed in areas such as contingency table (S.E. Fienberg, 2007), longitudinal studies with dropout (Roy, 2003) and intermittent missing data (Lin *et al.*, 2004). Also, a number of recent papers have established fundamental connections between the statistical properties of latent class models and their algebraic and geometric features (Smith and Croft, 2003; Settimi and Smith, 2005; Rusakov and Geigerm, 2005). Though there are potentially benefits to implement latent class analysis in different discipline and fields, it is at the cost of making some strong assumptions. One of these assumptions is choosing the number of latent classes. As mentioned above, different methods are proposed to assess latent class models with different number of latent classes. However, no assessment has been investigated on latent class models for missing values. In the next section, we present the underlying methods and models of our simulation studies.

3.2 Methods and Models of Simulation Studies

Rubin (1976) proposed three different missing mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Data are said to be missing completely at random when the probability that responses are missing is unrelated to either the specific values that should have been obtained or the set of observed responses. For instance, in longitudinal studies, let T be the total discrete time points, Y_{ij} be an observation for subject i at time j, and $\mathbf{R_i}$ be an $T \times 1$ vector of response indicators for subject i: $\mathbf{R_i} = (r_{i1}, r_{i2}, \ldots, r_{iT})'$ with $r_{ij} = 0$ if the corresponding response Y_{ij} is observed and $r_{ij} = 1$ if Y_{ij} is missing. In addition, associated with $\mathbf{Y_i}$ is an $T \times p$ matrix of covariates, X_i. Given $\mathbf{R_i}$, the complete set of responses $\mathbf{Y_i}$ can be partitioned into two components ${\mathbf{Y_i}}^{obs}$ and ${\mathbf{Y_i}}^{mis}$, corresponding to those responses that are observed and missing, respectively. Longitudinal data are MCAR when $\mathbf{R_i}$ is independent of both ${\mathbf{Y_i}}^{obs}$ and ${\mathbf{Y_i}}^{mis}$,

$$Pr(\mathbf{R_i} | {\mathbf{Y_i}}^{obs}, {\mathbf{Y_i}}^{mis}, X_i) = Pr(\mathbf{R_i})$$

Data are said to be missing at random when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that should have been obtained. For instance, longitudinal data are MAR when U_i is conditionally independent of ${\mathbf{Y_i}}^{mis}$, given ${\mathbf{Y_i}}^{obs}$, i.e.

$$Pr(\mathbf{R_i} | {\mathbf{Y_i}}^{obs}, {\mathbf{Y_i}}^{mis}, X_i) = Pr(\mathbf{R_i} | {\mathbf{Y_i}}^{obs}, X_i)$$

The third type of missingness of data is referred to not missing at random. Missing data are said to be NMAR when the probability that responses are missing is related to the specific values that should have been obtained. That is, the conditional distribution of $\mathbf{R_i}$ is related to ${\mathbf{Y_i}}^{mis}$ given ${\mathbf{Y_i}}^{obs}$, and $Pr(\mathbf{R_i} | {\mathbf{Y_i}}^{obs}, {\mathbf{Y_i}}^{mis}, X_i)$ depends on at least some elements of ${\mathbf{Y_i}}^{mis}$. Our interests focus on two of three types of missingness (MCAR and NMAR) and corresponding mixture models. In the simulation studies that we have performed, datasets with different missing mechanisms are simulated and investigated by fitting latent class models. Three underlying assumptions of missingness in the datasets have been investigated : MCAR missing mechanism, NMAR missing mechanism and a mixture of both missing mechanisms, MCAR and NMAR. We considered a longitudinal study for 6 time points with mixed effects (or growth curve model):

$$y_{ij} = g_{0i} + g_{1i}t_j + \beta_2 x_{1ij} + \beta_3 x_{2ij} + \varepsilon_{ij} \tag{3.7}$$

where

$$g_{0i} = \beta_0 + b_{0i}$$

$$g_{1i} = \beta_1 + b_{1i}$$

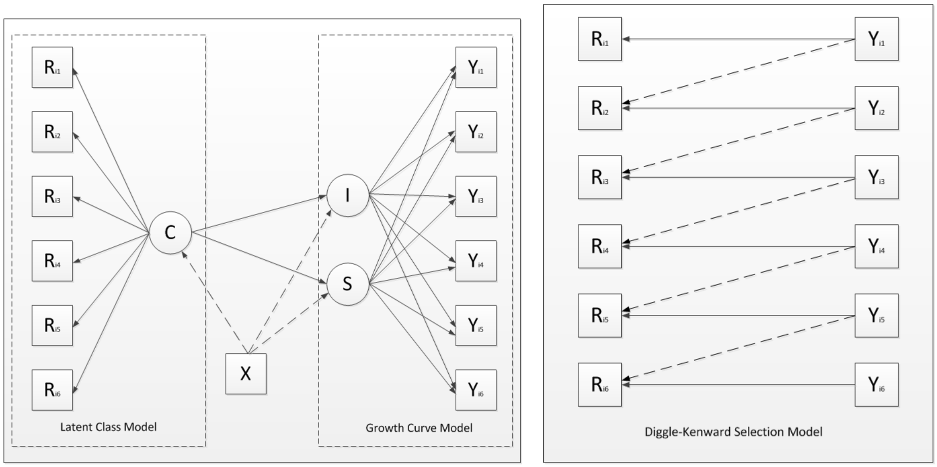
In this model, y_{ij} is the observation for subject i and time j, x_{1ij}, x_{2ij} are two covariates, b_{0i} is the random intercept for subject i with mean μ_{b_0} and variance $\sigma^2_{b_0}$, b_{1i} is the random slope for subject i with mean μ_{b_1} and variance $\sigma^2_{b_1}$. (In the simulated growth curve model, we assume the following parameters: random intercept b_{0i} and random slope b_{1i} are normally distributed with mean vector $[1, 2]$, and variance covariance structure $\begin{bmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$.) In this model, two time-invariant covariates x_1 and x_2 were also include for the analysis purpose. To represent missing values, we used the following Diggle-Kenward selection model to indicate missingness of a value at time j:

$$log[\frac{P(r_{ij} = 1|y_{ij}, y_{i,j-1})}{P(r_{ij} = 0|y_{ij}, y_{i,j-1})}] = \alpha_j + \xi_1 y_{ij} + \xi_2 y_{i,j-1} \tag{3.8}$$

where α_j is a const intercept in the above logit expression, ξ_1 and ξ_2 are the coefficients of the observations y_{ij} and $y_{i,j-1}$, respectively.

3.2.1 Simulation Model of MCAR Missing Mechanism

To illustrate the simulation methods, we started from a simple case: fitting latent class models in simulated data that contains one missing mechanism. To simulate datasets followed by the assumed model in equation (3.7), Monte Carlo technique is



(a) Diagram of simulated models

(b) Models for simulating missing data

Figure 3.1: Models studied in the simulations: latent class model and growth curve model (left); Diggle-Kenward selection model (right)

applied. In the simulation with MCAR missing mechanism, we set the coefficients of covariates in equation (3.8) to be zeros, that is: $\xi_1 = \xi_2 = 0$ and set the intercept in the logit expression $\alpha_j = 1$, which corresponds to a probability of 0.27 of having missing data on the dependent variables (observations), i.e.

$$P(r_{ij} = 1|y_{ij}, y_{i,j-1}) = \frac{1}{1 + exp(-1)} \tag{3.9}$$

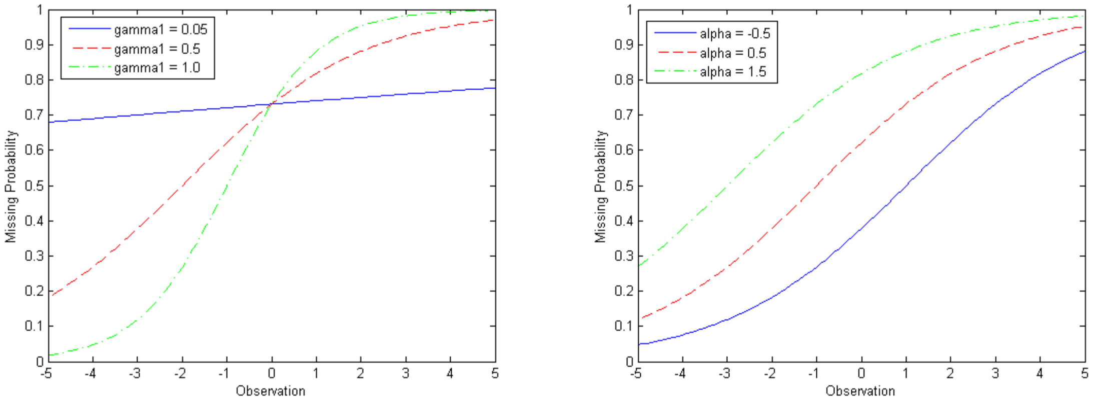
In this case, the missing probability is not related to either current or previous observations. This would reflect missing completely at random. A total of 1000 samples of MCAR were created using Monte Carlo method and each sample has 1000 observations. There are 64 different missing patterns in the simulated data, including the complete case. Latent class models with different number of classes have been applied for this data, in order to evaluate how the responses change through 6 time points from a grouping perspective. Covariates, as potential factors for explaining responses, were also investigated for whether they have effects on determing the number of latent classes.

3.2.2 Simulation Model of NMAR Missing Mechanism

Another type of simulation of interest was comparing latent class models for missing values under NMAR. In some cases, even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. This motivates us to fit latent class models for this type of missingness, and the conditional probability is defined as follows: considering the current observation of y_{ij} for subject i at time j, missingness of y_{ij} could partially or fully depends on the unobserved values of y_{ij}, the conditional probability has the same expression with equation (3.8), i.e.

$$P(r_{ij} = 1|y_{ij}, y_{i,j-1}) = \frac{1}{1 + exp\{-(\alpha_j + \xi_1 y_{ij} + \xi_2 y_{i,j-1})\}}$$

where coefficients $\alpha_j, \xi_2 \in R$ could take arbitrary values. In the above expression of conditional probability, changing the value of ξ_1 or ξ_2 will change the association between responses and missing values. For instance, we assume equation (3.8) only involves parameters α_j and ξ_1, which also means that the missingness for current observation is only related with current observation. Figure 2(a) shows that the parameter ξ_1 determines the steepness of the curve over the middle of the range. This means that a given change in the value of y_{ij} will produce a larger change in the missing probability of a positive response when this parameter is large than when it is small. Figure 2(b) demonstrates the missing probability curves by changing the values of α_j. With the increase of α_j, there is a larger chance for an observation to be missing, compared with a lower α_j. Therefore, changing parameter values in equation (3.8) should alter the association among the missing value indicators and might have an influence on deciding the number of latent classes. The related simulation studies and



(a) Missing probability curves for different values of ξ_1 and $\alpha_j = 1$, $\xi_2 = 0$

(b) Missing probability curves for different values of α_j and $\xi_1 = 0.5$, $\xi_2 = 0$

Figure 3.2: Missing probability curves

corresponding results will be given in the next section. For simulations in this part, each simulation generated 1000 replicates and each replicate had 1000 observations, followed by NMAR.

3.2.3 Simulation Mixture Model of MCAR and NMAR

In a longitudinal study, data are collected from baseline to the end of the study. The presence of a big amount of missing values is common, accompanying with complicate missing mechanisms. Though it's often difficult to distinguish what missing mechanisms are involved in the dataframe, ideally a combination with MCAR and NMAR is a possible case. This motivates us to investigate a mixture model of combining these two different types of missing mechanisms. For simulations in this part, we have generated 1000 samples and each sample is consisted of different proportions of MCAR and NMAR, either 500 observations for each of missing mechanism or 800 observations for MCAR and 200 observations of NMAR, depending on the research goals. We will announce this proportion in the simulation results. The conditional probabilities for MCAR and NMAR are defined in previous two formulas.

Besides exploring the method to choose the optimal number of latent classes, covarites in the growth curve model, different settings of missing probabilities, and the associations among the y's may be of interest and investigated on selection number of latent classes. To generate different associations among the observations, one could change the parameters of random slope in growth curve model (3.7). For instance, with a higher value of μ_{b_1}, samples with highly associated observations would be generated. All these factors of interests should be explored by fitting latent class models on samples with different settings.

3.3 Analysis of Simulation Results

To compare performance of latent class models with different number of latent classes, Clogg (1995) and Aitkin (1981, 1985) indicated that chi-squared likelihood ratio statistics were not theoretically correct for LCM selections. A $M-1$ class LCM is obtained by putting one parameter value at the boundary of a M-classes model. The likelihood ratio between the two LCMs may not follow a single χ^2 distribution if the constrained model ($M-1$ classes) is obtained from the full model (M classes) by placing parameters at their boundary values. Several alternative methods, including information criteria, parametric resampling, etc. were suggested to solve the problem. Information criteria are probably one of the most convenient methods than other methods such as parametric resampling. We apply as the efficient approaches and compare the performances of convectional information criteria to evaluate latent class models, including AIC, BIC, CAIC, DBIC, and other four information criteria.

3.3.1 Information Criteria

Yang. (2006) discussed many information criteria that can be used to compare LCMs. Akaike information criterion (AIC) was one of the earliest propositions of information criteria. AIC has the following form

$$AIC_g = -2logL(\theta_g) + 2p_g$$

where $log\ L(\theta_g)$ is log-likelihood from MLE, p_g is the total number of free parameters in model g. However, Woodruffe (1982) showed that AIC is not theoretically consistent; consequently, AIC will not select the correct model when sample size (N) is near infinity.

Schwarz (1978) proposed Bayesian information criterion (BIC) which has the following form

$$BIC_g = -2logL(\theta_g) + p_g\ log\ N$$

Haughton (1988) showed BIC is consistent when sample size goes large and hence can lead to a correct choice of model when N goes infinity.

Bozdogan (1987) derived a consistent version of AIC, called CAIC from the Kullback-Leibler information measure with the form

$$CAIC_g = -2logL(\theta_g) + p_g\ (log\ N + 1)$$

Since CAIC puts more severe penalty on over-parameterization than BIC or AIC, it tends to favor a model with fewer parameters.

Draper (1995) modified the penalty part of BIC, and DBIC is defined as follows

$$DBIC_g = -2logL(\theta_g) + p_g(log\ N - log\ 2\pi)$$

When sample size N goes infinity, the added term is asymptotically insignificant, but it has a notable effect on the log-likelihood for small to moderate sample sizes.

We also included HQ information criterion which was invented by Hannan (1979), HT-AIC information criterion discovered by Hurvich (1989), sample size adjusted BIC (BICa) and CAIC (CAICa) to compare the performance among latent class models with different latent classes. For each simulation that we investigated, 1000 samples were simulated on different missing probabilities and then fitted with latent class models with latent classes either from 1 to 5 or from 2 to 5, depending one which simulation is processed. When we performed simulations of MCAR or NMAR alone, LCMs with latent classes from 1 to 5 were compared. For simulations of mixture of MCAR or NMAR, we compared LCMs with latent classes from 2 to 5. One can check in the latter case, LCMs with one latent class won't be suggested by any of the information criteria among 1000 samples. For each information criterion, a smaller value indicates a better model fit on the simulated data. After fitting latent class models on 1000 samples, tallies were made for the numbers latent classes indicated by each criterion, with number of latent classes, ranging either from 1 to 5 latent classes, or from 2 to 5 latent classes. To illustrate directly, we summarize the tallies and corresponding proportions for each information criterion in tables and marked the favored LCM in red.

All simulations and implementation of latent class models with missing data are completed by Mplus version 5 (Muthen and Muthen, 2011), which is a statistical modeling program with specializing in fitting structural equation models, and provides extensive capabilities for Monte Carlo simulation studies. For each simulation study, we generate 1000 replicates that consist of 1000 individuals (observations) for each replicate. The repeated measures with covaraites in each replicate are generated via model (3.7), and missing data are obtained from model (3.8). Latent class model for fitting corresponding missing indicators are evaluated by Mplus program, the simulation results in different parameter settings are analyzed in next section.

3.3.2 Model Selection for LCMs

We first consider simulation results of the selection of LCMs with parameters given in equation (3.7) and (3.8), which are used to simulate the samples. Three different underlying missing types are investigated: MCAR, NMAR and a mixture of MCAR and NMAR. To simulate a growth curve model with MCAR missingness, we assume the random intercept is normally distributed with mean 1 and variance 1; the random slope is also normally distributed with mean 2 and variance 0.2; for the MCAR missingness, we choose the default intercept term $\alpha_j = 1$ in the logit expression (3.9). To simulate a growth curve model with NMAR type of missing, we use the same model parameters in (3.7) as former one and assume the missing status for current observation is only related with current observation, not previous one, i.e. $\alpha_j = 1$, $\xi_1 = 0.2$ and $\xi_2 = 0$. To simulate a growth curve model with a mixture of two types of missing mechanisms, 500 observations are generated from each missing mechanism using the same model parameters. The simulation results are shown in Tables 1-3.

Table 1 describes the simulation results of LCMs for MCAR missing mechanism. There are 10 replicates that failed to converge when fitting the models. All the information criteria support the LCM with one latent class, with spreading trends in both AIC and HT. Table 2 summarizes the results for NMAR missing mechanism, most information criteria suggest the model with one latent class, except AIC and HT. Both AIC and HT present significant spreading trends in the simulation results, and reverse the results to LCM with two latent classes. As discussed before, AIC tends to give an inaccurate suggestion due to its inconsistency when sample size gets large. HT information criteria is derived from AIC, and it inherits the inconsistency property as well. Simulation results demonstrate that a LCM with a homogeneous

Table 3.1: Number of latent class tallies on MCAR simulation

Information	Latent class model				
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	810 (0.82)	155 (0.16)	20 (0.02)	2 (0.002)	3 (0.003)
BIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
HT	823 (0.83)	149 (0.15)	14 (0.01)	1 (0.001)	3 (0.003)
BICa	988 (0.998)	2 (0.002)	0 (0.00)	0 (0.00)	0 (0.00)
CAICa	989 (0.999)	1 (0.001)	0 (0.00)	0 (0.00)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\xi_1 = 0$, $\xi_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

group is favored for single missing mechanism and fairly low probability of missing.

The simulation results for selection of LCMs for a mixture of two missing mechanisms are summarized in Table 3. All information criteria support LCM with two latent classes, while there are large dispersion of tallies over AIC and HT. By reviewing the way we simulate data for a mixture of two missing mechanism, two datasets with the single missing mechanism are merged. Simulation results indicate this mixing and suggest that LCM with two heterogeneous groups has a better model fit. Without loss of generality, we choose the results in Table 3 and the corresponding models as the reference results and models, to investigate the following factors of interests.

Table 3.2: Number of latent class tallies on NMAR simulatio

Information	Latent class model				
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	229 (0.23)	306 (0.31)	197 (0.20)	141 (0.14)	117 (0.12)
BIC	987 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	978 (0.99)	12 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	916 (0.925)	72 (0.073)	2 (0.002)	0 (0.00)	0 (0.00)
HT	253 (0.25)	343 (0.35)	197 (0.20)	120 (0.12)	77 (0.08)
BICa	899 (0.908)	88 (0.089)	3 (0.003)	0 (0.00)	0 (0.00)
CAICa	902 (0.911)	85 (0.086)	3 (0.003)	0 (0.00)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\gamma_1 = 0.2$, $\xi_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

3.3.3 Covariate Effects

In general, covariates potentially affects the relationship between the dependent variable and other independent variables of primary interest. Two covariates are included in our simulation studies, namely, X_1 and X_2, and both covariates are generated from standard normal distribution in Monte Carlo simulations. In equation (3.7), covariates provide extra information on observations y_{ij} and those observations potentially influence the missing indicators R_{ij}, as expressed in equation (3.8). The covariates effect on selection of LCMs may be of interest. To investigate this effect, we evaluate LCMs for the mixture of the two missing mechanisms, with or without incorporating covariates in LCMs. One could do the same study on LCMs for sin-

Table 3.3: Number of latent class tallies on mixture of MCAR and NMAR

Information	Latent class model				
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	387 (0.39)	282 (0.28)	201 (0.20)	125 (0.13)
BIC	0 (0.00)	992 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	992 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	987 (0.992)	8 (0.008)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	964 (0.969)	29 (0.029)	2 (0.002)	0 (0.00)
HT	0 (0.00)	438 (0.44)	286 (0.29)	177 (0.18)	94 (0.09)
BICa	0 (0.00)	952 (0.957)	41 (0.041)	2 (0.002)	0 (0.00)
CAICa	0 (0.00)	954 (0.959)	39 (0.039)	2 (0.002)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\xi_1 = 0(MCAR), = 0.2(NMAR)$, $\xi_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

gle missing mechanism. While fitting LCMs without covariates for 1000 replicates, 995 successfully converged; fitted converged 992 among 1000 samples for LCMs with covariates.

Table 3 describes the results of LCMs without incorporating covariates. All information criteria support a LCM with two latent classes, i.e. a LCM with two heterogeneous groups has a better model of fit. Table 4 lists the tallies of LCMs with covariates and most information criteria suggests the same number of latent classes as the case of without covariates, except AIC and HT. Due to the inconsistency of AIC and HT, they don't correctly identify a model, in particular, they select the model

Table 3.4: Number of latent class tallies on mixture of MCAR and NMAR with covariates

Information	Latent class model				
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	144 (0.14)	214 (0.22)	259 (0.26)	375 (0.38)	0 (0.00)
BIC	0 (0.00)	983 (0.991)	1 (0.001)	1 (0.001)	7 (0.007)
CAIC	0 (0.00)	792 (0.80)	157 (0.16)	30 (0.03)	13 (0.01)
DBIC	0 (0.00)	974 (0.982)	10 (0.010)	1 (0.001)	7 (0.007)
HQ	0 (0.00)	928 (0.936)	52 (0.052)	5 (0.005)	7 (0.007)
HT	0 (0.00)	286 (0.288)	290 (0.292)	222 (0.224)	194 (0.196)
BICa	0 (0.00)	775 (0.78)	168 (0.17)	35 (0.04)	14 (0.01)

*With covariates,low missing probabilities, high association among responses.

with more latent classes than it actually had. Simulations have shown the covariates do not alter the choice of number of latent classes of LCMs when the models are applied for data with two missing mechanisms, MCAR and NMAR. However, the auxiliary information provided by covariates ”un-stabilizes” the selection of LCMs by information criteria. For instance, one of the best performing information craiteria, BIC supports a two latent class model in most cases (with probability $p \approx 0.997$) when there is no covariate considered; it loses this performance when covariates are incorporated (with probability $p \approx 0.991$). Other information criteria have more significant loss on this performance when incorporating covariates into models. AIC and HT are severely sensitive to the covariates effects. AIC drops this probability from 0.39 to 0.14 for supporting a LCM with two latent classes.

Table 3.5: Number of latent class tallies on mixture of MCAR and NMAR with low associations among responses(without covariates)

Information	Latent class model				
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	701 (0.706)	232 (0.234)	41 (0.041)	19 (0.019)
BIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	986 (0.993)	7 (0.007)	0 (0.00)	0 (0.00)
HT	0 (0.00)	737 (0.742)	215 (0.217)	29 (0.029)	12 (0.012)
BICa	0 (0.00)	984 (0.991)	9 (0.009)	0 (0.00)	0 (0.00)
CAICa	0 (0.00)	985 (0.992)	8 (0.008)	0 (0.00)	0 (0.00)

*$\alpha_j = 1$, $\xi_1 = 0(MCAR), = 0.2(NMAR)$, $\xi_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 1$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

3.3.4 Association Effect among Responses

As we discussed before, model parameters in (3.7) and (3.8) are initialized at the beginning of data simulations. In this part we consider the changes on parameters in equation (3.7), more specifically, we simulate growth curve models with missingness by altering the parameters in the random slope term to different values, i.e. the mean and variance of b_{1i}. To avoid the redundant tables, we provide one of the simulations with two different initialized mean values of b_{1i}: while $\mu_{b_{1i}} = 1$ represents a lower association among observations, $\mu_{b_{1i}} = 2$ indicates a higher association.

Table 5 displays the results for the lower association. All information criteria agree a LCM with two heterogeneous groups will fit the missing values better. By comparison, the results for the higher association case are shown in Table 3. It is indicated that with increasing the degree of associations among responses, the choice of number of latent classes won't change. However, the problem of changes in the "selection certainty" draws our attention again. One of the worst behaviored information criteria AIC losses its choice certainty from 0.706 to 0.39.

3.3.5 Missing Probability Effect

To investigate the selection of LCMs for missing values, Diggle-Kenward selection model are intensively used in our simulation studies, as described in equation (3.8). In this expression, the missing probability for the current observation y_{ij} is determined by the value of previous observation $y_{i,j-1}$, current observation y_{ij} and initialized parameter values α_j, ξ_1, and ξ_2. Changing any one of these values will lead to a change in missing probabilities and potentially affect the structure of LCMs. For instance, increasing the coefficient ξ_1 will lead to a higher missing probability for the current observation y_{ij}, while holding other parameters fixed. Table 3 and 6 present the model selection results for a paired values of ξ_1 (0.2, 0.4) which are set to simulate the missingness. $\xi_1 = 0.2$ corresponds to a lower missing probability, when $\xi_1 = 0.4$ corresponds to a higher missing probability. $\alpha_j = 1$ and $\xi_2 = 0$ are fixed in this comparison.

Table 3 illustrates the simulation results for the lower missing probability: a LCM with two latent classes is suggested by all information criteria. Clearly it is suggested that LCM is changed in the higher missing probability case, based on the cell values in Table 6. While both BIC and CAIC support a LCM with three heterogeneous groups, all the other information criteria tend to indicate for four latent classes. This change shows evidence of the influence of missing probability on the LCM selection,

Table 3.6: Number of latent class tallies on mixture of MCAR and NMAR with high missing probability(without covariates)

Information			Latent class model		
Criteria	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	0 (0.00)	5 (0.005)	546 (0.553)	437 (0.442)
BIC	0 (0.00)	15 (0.015)	849 (0.859)	124 (0.126)	0 (0.00)
CAIC	0 (0.00)	15 (0.015)	849 (0.859)	124 (0.126)	0 (0.00)
DBIC	0 (0.00)	0 (0.00)	470 (0.476)	511 (0.517)	7 (0.007)
HQ	0 (0.00)	0 (0.00)	186 (0.19)	766 (0.77)	36 (0.04)
HT	0 (0.00)	0 (0.00)	8 (0.008)	609 (0.616)	371 (0.376)
BICa	0 (0.00)	0 (0.00)	167 (0.169)	777 (0.786)	44 (0.045)
CAICa	0 (0.00)	0 (0.00)	169 (0.169)	776 (0.786)	43 (0.045)

*$\alpha_j = 1$, $\xi_1 = 0(MCAR), = 0.6(NMAR)$, $\xi_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

i.e. with a higher missing probability, LCMs with more heterogeneous groups are preferred.

To investigate the selection of LCMs, we have checked the missing mechanisms and related factors that derived from changing parameters in either model equation (3.7) or missing values generating mechanism (3.8), and through simulation studies we conclude their influences on deciding the number of latent classes. To fit the datasets which consist of two assumed missing mechanisms groups, the cases where a LCM with three heterogeneous groups is suggested are worthy to be researched further. However, the assumed missing mechanisms usually cannot be identified in practice.

In particular, there is no statistical method or test on NMAR and the mixture of MCAR and NMAR. By contrast, missing patterns could be directly observed and it may provide another perspective to understand LCMs. In the last part of this section, we focus on exploring the behavior of missing patterns on LCMs with three latent classes.

3.3.6 Missing Patterns in LCMs

In the above simulations, longitudinal studies with 6 time points are considered. We define R_{ij} as the missing indicator for subject i at time j. The possible missing patterns are $2^6 = 64$. For a large sample size, many of the missing patterns will be repeated. In our simulations, each sample has 1000 observations and a list of the observed missing patterns together with their associated frequencies is given in the Appendix. The posterior probability $h(s|\mathbf{x})$ of an individual with missing pattern $\mathbf{x}$ belonging to sth group could be obtained when the corresponding LCM is fitted, based on the definition in equation (3.6). In our case, three posterior probabilities for each latent class would be calculated for each missing pattern and these results are given in the Appendix as well. A missing pattern $\mathbf{x}$ is allocated in the class for which the posterior probability is greatest.

Let $C_{s|\mathbf{x}}$ be the posterior count for sth latent class given missing pattern $\mathbf{x}$, which can be calculated as the product of observed frequency f and posterior probability $h(s|\mathbf{x})$. Based on the posterior counts we could explore the missing patterns in deciding allocation of latent classes. For instance, LCMs with three heterogeneous groups in our simulation studies are of interest to investigate further. For instance, Table 7 lists the posterior probabilities and counts for the first 10 missing patterns in one of our simulation studies. '0' in missing pattern represents observed response, and '1' means missing response. Two numbers in the parenthesis for frequency item are frequencies counted from MCAR and NMAR, respectively. The total frequency for the 8th missing pattern $\mathbf{x}_8$ is 223, where responses are only observed at the first time point. 220 out of 223 come from NMAR mechanism, only 3 come from MCAR mechanism. Among on the posterior counts $C_{s|\mathbf{x}_8}$ $(s = 1, 2, 3)$ for this pattern, the second latent class has $C_{2|\mathbf{x}} = 208.282$. Therefore, this pattern is associated with the second latent class. In fact, the second latent class essentially consists of three missing patterns: $\mathbf{x}_6$, $\mathbf{x}_7$ and $\mathbf{x}_8$. $\mathbf{x}_8$ are the majority in this group, i.e. observations with this type of missing pattern will be allocated in the second latent class.

From the inspection on all missing patterns in each simulation, one could find that the first two latent classes mainly consist of missing patterns from NMAR mechanism, and missing patterns from MCAR forms the third class. Compared with cases where LCMs with two classes are preferred, we find that there is a seperation in the NMAR mechanism, which lead to an additional class. Further, we could observe that in LCMs, latent classes are represented by homogeneous responses, i.e. homogeneous missing patterns fall into a single class.

3.4 Discussion

This chapter described simulation studies on selecting the number of latent classes for missing values and the comparison of results based on eight information criteria. The Bayesian information criteria, consistency version of AIC (CAIC) and sample adjusted BICa are noteworthy information criteria to choose correct latent classes. AIC presents its inconsistency property in the simulation studies. HT has less consistent performance as well. These inconsistent information criteria are not recommended for real case studies.

Covariates and degree of association among responses do not influence deciding how many latent classes are best for fitting the data with different missing mech-

Table 3.7: Posterior probability for LCMs with three classes (first 10 frequent missing patterns)

Missing Pattern	Frequency f	$h(1\|\mathbf{x})$	$h(2\|\mathbf{x})$	$h(3\|\mathbf{x})$	$C_{1\|\mathbf{x}}$	$C_{2\|\mathbf{x}}$	$C_{3\|\mathbf{x}}$
011111	223 (3,220)	0.056	0.934	0.009	12.488	208.282	2.007
001111	96 (2,94)	0.948	0	0.052	91.008	0	4.992
101111	54 (1,53)	0.955	0	0.045	51.57	0	2.43
000011	50 (35,15)	0.115	0	0.885	5.75	0	44.25
000111	44 (16,28)	0.707	0	0.293	31.108	0	12.892
000001	39 (36,3)	0	0	1	0	0	39
000000	33 (32,1)	0	0	1	0	0	33
001011	33 (21,12)	0.494	0	0.506	16.302	0	16.698
000010	30 (29,1)	0	0	1	0	0	30
010111	27 (5,22)	0.167	0.625	0.207	4.509	16.875	5.589

*Missing data are simulated using Diggle-Kenward model ($\alpha_j = 1,\ \xi_1 = 0.4,\ \xi_2 = 0.4$). 0 is observed response, 1 is missing response.

anisms. However, changing these parameters will influence ”selection certainty” of all inforamtion criteria. Increasing the degree of associations among responses or incorporating covariates in the simulation model will lead to the loss of ”selection certainty”. We also find that the selection by AIC and HT are more sensitive to these changes. Compared with those less-influential factors, missing probabilities directly have effects on deciding number of latent classes. A higher missing probability tends to make the number of latent classes larger. Bayesian Information Criterion (BIC) and consistent version of AIC (CAIC) suggest conservative LCMs with three classes, while other information criteria indicate that four classes are preferred. One would like to choose the smallest number of classes that allows the assumption of conditional independence to hold. A latent class model with too many classes can produce problems. One of these problems is difficulty to interpret these classes due to the small size of classes.

Missing patterns are also investigated for the chosen latent classes. Posterior counts for each pattern are calculated and compared. The allocation for each pattern is based on the largest posterior probability, i.e. assgin a pattern to the class where the posterior probaility is the greatest. Studies indicate that latent classes in LCMs are represented by homogeneous missing patterns. And the underlying missing mechanism could account for the classes. For the two class LCMs, one class mainly comes from missing patterns generated by MCAR, when the other is consisted of missing patterns from NMAR. LCMs with three classes in the simulations could be illustrated as a separation of missing patterns in NMAR.

If one wants to apply LCMs to capture the group characteristics for missing values, a simulation on deciding the number of latent classes is recommended before fitting the model. Further research on latent variables for missing indicators may be of interests. As shown in Figure 1, the assumed latent class C is related with latent variables i, s which are used as random intercept and slope in the growth curve model. If the observations $\mathbf{Y}$ are continuous, both random terms could be continuous and the linked latent variables for missing indicator could be continuous as well.

Chapter 4

A CONTINUOUS LATENT FACTOR MODEL FOR NON-IGNORABLE MISSING DATA IN LONGITUDINAL STUDIES

In this chapter we will review linear mixed models that incorporate unobserved responses and present a novel parametric approach to modeling longitudinal data when non-ignorable missing values are involved. Mixed effects modeling is one of the prevalent method for the analysis of correlated data where correlation can arise from repeated measurements, longitudinal data or clustering. Since the foundation paper of Laird and Ware (1982), a vast amount of literature has developed that extends a range of model fitting techniques and applications. (Diggle and Zeger, 1994; McCulloch and Searle, 2001; Fitzmaurice and Ware, 2004) These together provide a comprehensive description of methods for estimation and prediction of linear, generalized linear and nonlinear mixed-effects modeling. Many longitudinal studies suffer from missing data due to subjects dropping into or out of a study or not being available at some measurement times, which can cause bias in the analysis if the missingness are informative. For likelihood procedures of estimating linear mixed models, we may generally ignore the distribution of missing indicators when the missing data are MAR (or ignorable likelihood estimation), that is missingness depends only on observed information. However, when the missing data mechanism is related to the unobservable missing values, the missing data are non-ignorable and the distribution of missingness has to be considered. To account for informative missingness, a number of model based approaches have been proposed to jointly model the longitudinal outcome and the non-ignorable missing mechanism. Little and Rubin (2002) described three major formulations of joint modeling approaches: selection model, pattern-mixture model and shared-parameter model (the detailed formulations were given in Chapter 2), while Verbeke and Molenberghs (2000) provided applications for these models in their book. Other researchers have extended this field in the last decade. Some authors have incorporated latent class structure into pattern-mixture models to jointly describe the pattern of missingness and the outcome of interest (Lin *et al.*, 2004; Muthn *et al.*, 2003; Roy, 2003). Lin et al. (2004) proposed a latent pattern-mixture model where the mixture patterns are formed from latent classes that link a longitudinal response with a missingness process. Roy (2003) investigated latent classes to model dropouts in longitudinal studies to effectively reduce the number of missing-data patterns. Muthen et al. (2003) also discussed how latent classes could be applied to non-ignorable missingness. Jung et al. (2011) extended traditional latent class models, where the classes are defined by the missingness indicators alone.

All the above extensions are from the family of pattern-mixture models, and these models stratify the data according to time to dropout or missing indicators alone and formulate a model for each stratum. This usually results in under-identifiability, since we need to estimate many pattern-specific parameters even though the eventual interest is usually on the marginal parameters. Further, there is a controversial and also important practical modeling issue in using latent class models which is determining a suitable number of latent classes. Some authors suggested criterion approach as a way of comparing models with different number of classes. In the simulation studies in Chapter 3, we investigated eight different information criteria systematically on their performances when different missing data settings were handled. In our work, we found that the selection of latent classes is sensitive to many factors that relate to missing data, and a simulation study on selection latent classes is strongly recommended if one wants to apply latent class modeling for missing data. Moreover, the uncertainty of model selection makes latent class models inefficient in estimating population parameters. Instead of modeling missing indicators with latent categorical classes, one possible alternative approach is to model missingness as continuous latent variables.

As the alternative, Guo *et al.* (2004) extended pattern-mixture to a random pattern-mixture model for longitudinal data with dropouts. The review work for Guo's paper is given in Chapter 2. The extended model works effectively on the case where a good surrogate for the dropout can be representative for the dropout process. In most real studies, however, it maybe impossible to find good measures for the missing mechanism. For instance, in a longitudinal study with many intermittent missing values, time to dropout is not necessarily a good measure, and it probably wouldn't capture most features of missingness. That is, this measurement can not represent for subjects who have drop-in responses. Instead, modeling for missing indicators is necessary in this case. Further, models other than the normal distribution will be required to describe the missingness process. The violation of joint multivariate normality will lead to an increase of computation difficulties. In the proposed new model, missing indicators are directly modeled with a continuous latent variable, and this latent factor is treated as a predictor for latent subject-level random effects in the primary model of interests. Some informative variables related with missingness (e.g. time to first missing, number of switches between observed and missing responses) will be served as covariates in the modeling of missing indicators. The detailed description of the new model will be given in next section.

4.1 Background of Continuous Latent Factor Model for Binary Outcome

For analyzing multivariate categorical data, continuous latent factor modeling which is often referred to as categorical variable factor analysis (Muthen, 1978) and item response modeling (Lord, 1980; Embretson and Reise, 2000) probably is the most widely used method. In the terminology of educational testing, the involved binary variables are called items and the observed values are referred to as binary or dichotomous responses. In this paper, we will extend this model to describe missing data procedure.

Let $r_{i1}, \ldots, r_{iJ}$ be the J binary responses (missing indicators) on J given time points for a given individual i out of a sample of n individuals, $i = 1, \ldots, n$ and $j = 1, \ldots, J$. In concrete cases 1 and 0 may correspond to a observed or unobserved outcome in a longitudinal study. In the continuous latent factor model there are two sets of parameters. The probability of r_{ij} being 1 or 0 can depend on an individual parameter u_i, specific and characteristic for the individual in study. This parameter is also referred to as a latent parameter. In addition, the probability may depend on a parameter for different time points (items) τ_j, characteristic for the particular time point.

We use the following notation to define the probability of a missing outcome as a function of the latent individual factor:

$$\pi_{ij}(\tau_j) = Pr(r_{ij} = 1 | u_i).$$

It is usually assumed that $\pi_{ij}(\tau_j)$ is monotonously increasing from 0 to 1 as u_i runs from $-\infty$ to ∞, and that ξ_j is the 50%-point, i.e. $\pi_{ij}(\xi_j) = 0.5$. A typical latent trait plot is shown in Figure 4.1.

In the literature two main models for a latent trait have been suggested. The normal ogive model or probit model is given by

$$\pi_{ij}(u_i) = \Phi(u_i - \tau_j)$$

where $\Phi(x)$ is the cumulative normal distribution function. Alternatively we may use

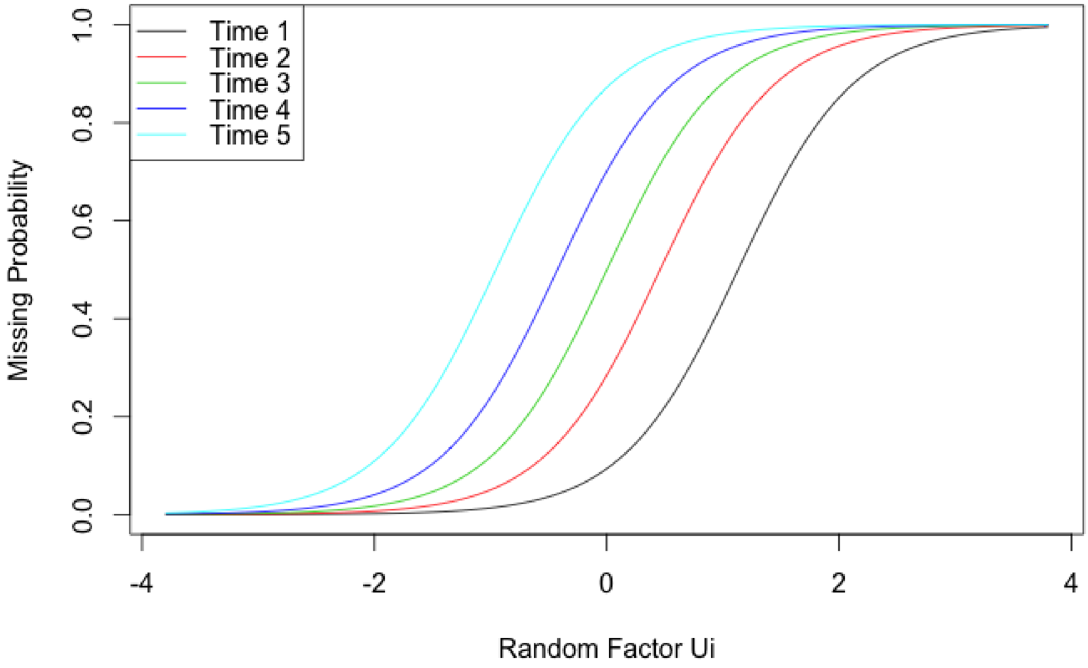


Figure 4.1: A typical latent trait plot

the logistic model or logit model,

$$\pi_{ij}(u_i) = \Psi(u_i - \tau_j)$$

where $\Psi(x) = e^x/(1+e^x)$ $(-\infty < x < \infty)$ is the cumulative distribution function of standard logistic random variable.

There is a series of continuous latent variable models for different kinds of categorical data. Here, we present the 2-parameter (2PL) item response model for binary data, which could be reduced to the model discussed above. The 2PL model is used to estimate the probability (π_{ij}) of a missing response for subject i and time point j while considering the item (time)-varying parameters, τ_{2j} for item (time) location parameters and τ_{1j} for item (time) slope parameters, which allow for different weights for different times, and the person-varying latent trait variables u_i. The 2PL model is expressed as

$$logit(\pi_{ij}) = \tau_{1j}(u_i - \tau_{2j}).$$

As τ_{1j} increases, the item (time) has a stronger association with the underlying missingness. When τ_{1j} is fixed to be 1, the 2PL model is reduced to be a Rasch model (Rasch, 1960) or a 1PL model. As τ_{2j} increases, the response is more likely to be observed. This 2PL model has been shown to be mathematically equivalent to be confirmatory factor analysis model for binary data (Takane, 1987). The IRT models can be expressed as generalized mixed or multilevel models (Adams, 1997; Rijmen, 2003). Considering a mixed logistic regression model for binary data:

$$p(r_{ij} = 1|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \beta, \mathbf{u}_i) = \frac{exp(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{u}_i)}{1 + exp(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{u}_i)}$$

where r_{ij} is the binary response variable for subject i at time j, $i = 1, \ldots, n$; $j = 1, \ldots, J$; $\mathbf{x}_{ij}$ is a known P-dimensional covariate vector for the P fixed effects; $\mathbf{z}_{ij}$ is a known Q dimensional design vector for the Q random effects; β is the P-dimensional parameter vector of fixed effects; and $\mathbf{u}_i$ is the Q-dimensional parameter vector of random effects for subject i. In this model, the binary responses are assumed to be independent Bernoulli conditional on the covariates, the fixed effects, as well as the random effects. This conditional independence assumption is often referred to in the latent variable model literature as the assumption of local independence. The described model comes from the family of the generalized linear mixed model in which the observations are relations from a Bernoulli distribution (belonging to the exponential family), mean $\mu_{ij} = p(r_{ij} = 1|\mathbf{x}_{ij})$, and the canonical link function is the logit function. The IRT model is formally equivalent to a nonlinear mixed model, where the latent variable $\mathbf{u}_i$ is the random effect; time covariate τ_{2j} and slope parameter τ_{1j} are treated as fixed effects. Raudenbush (2003) also reexpressed the

Rasch model and the 2PL model as a two-level logistic model by including dummy variables indicating item numbers (time locations).

4.2 Proposed Model

In this section we present a continuous latent factor model (CLFM) in the longitudinal data with non-ignorable missingness. For a J-time period study which may have as many as 2^J possible missing patterns; modeling the relationship among the missing indicators and their relationships to the observed data is a challenge. The underlying logic of our new model comes from the assumption that a continuous latent variable exists and allows flexibly for modeling missing indicators. Suppose we have a data set with n independent individuals. For individual i ($i = 1, \cdots, n$), let $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iJ})'$ be a J-dimensional observed vector with continuous elements used to measure a q-dimensional continuous latent variable $\mathbf{b}_i$. Let $\mathbf{R}_i = (r_{i1}, \cdots, r_{iJ})'$ be a J-dimensional observed missing vector with binary elements and u_i be a continuous latent variable, which is used to measure $\mathbf{R}_i$. The primary model of interest will be the joint distribution of $\mathbf{Y}_i$ and $\mathbf{R}_i$, given u_i and possibly additional observed covariates $\mathbf{X}_i$, where $\mathbf{X}_i$ represents p-dimensional fully observed covariates. Figure 4.2 (model D) provides a diagram representing the proposed model for all the observed and latent variables. As indicated in Figure 4.2, $\mathbf{X}_{1i}$, containing both time-variant and time-invariant attributes for subject i, is the p_1 dimensional covariates and used in model B; $\mathbf{X}_{2i}$ is the p_2 dimensional covariates used in model A; a p_3 dimensional time-invariant covariate vector $\mathbf{X}_{3i}$ is used in modeling link function between $\mathbf{b}_i$ and u_i. These three covariate-vector form the covariate for model D, i.e. $p = p_1 + p_2 + p_3$.

One of the fundamental assumptions of this new model is that $\mathbf{Y}_i$ is conditionally independent of $\mathbf{R}_i$ given the latent variables u_i and $\mathbf{b}_i$. This is a natural assumption when modeling relationships between variables measured with error, i.e., we want

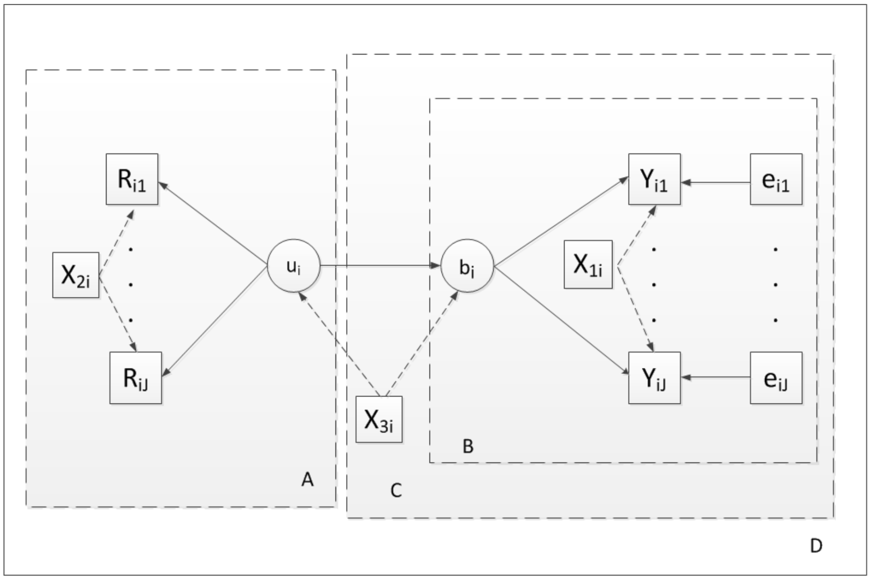


Figure 4.2: Proposed model diagram: observed quantities are described in squared boxes, latent quantities are in circled boxes

to model the relationship between the underlying variables, not the ones with error. Finally, we assume that $\mathbf{Y}_i$ is conditionally independent of u_i given $\mathbf{b}_i$, and likewise, $\mathbf{R}_i$ is conditionally independent of $\mathbf{b}_i$ given u_i. Hence, we introduce the following model for the joint distribution of the responses $\mathbf{Y}_i$ and missing indicators $\mathbf{R}_i$,

$$f(\mathbf{Y}_i,\ \mathbf{R}_i|\mathbf{X}_i) = \iint f(\mathbf{Y}_i|\mathbf{b}_i,\ \mathbf{X}_{1i}) f(\mathbf{R}_i|u_i,\ \mathbf{X}_{2i}) f(\mathbf{b}_i|u_i, \mathbf{X}_{3i}) f(u_i) du_i d\mathbf{b}_i \tag{4.1}$$

with specific parametric models specified as follows: ($N_p(\mathbf{a}, B)$ denotes the p-variate normal distribution with mean $\mathbf{a}$ and covariance matrix B)

$$(\mathbf{Y}_i|\mathbf{b}_i,\ \mathbf{X}_{1i}) \sim_{ind} N_J(\mathbf{X}_{1i}\beta + \mathbf{Z}_{1i}\mathbf{b}_i,\ \Sigma_\epsilon) \tag{4.2}$$

$$(\mathbf{b}_i|u_i, \mathbf{X}_{3i}) \sim_{ind} N_q(\mathbf{X}_{3i}^{'}\gamma,\ \zeta_i) \tag{4.3}$$

$$u_i \sim_{ind} N_1(0,\ \sigma_u^2) \tag{4.4}$$

$$f(\mathbf{R}_i|u_i,\ \mathbf{X}_{2i}) = \prod_{j=1}^{J} \pi_{ij}^{r_{ij}}(1-\pi_{ij})^{1-r_{ij}} \tag{4.5}$$

A linear mixed model (growth curve) is used for the relationship between $\mathbf{Y}_i$ and $\mathbf{b}_i$ (model B in Figure 4.2), where $\mathbf{X}_{1i}$ is a known $(J \times p_1)$ design matrix containing fixed within-subject and between-subject covariates (including both time-invariate and time-varying covariates), with associated unknown $(p_1 \times 1)$ parameter vector β, $\mathbf{Z}_{1i}$ is a known $(J \times q)$ matrix for modeling random effects, and $\mathbf{b}_i$ is an unknown $(q \times 1)$ random coefficient vector. We specify $\mathbf{Y}_i = \mathbf{X}_{1i}\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$, where the random error term ϵ_i is a J-dimensional vector with $E(\epsilon_i) = \mathbf{0}$, $Var(\epsilon_i) = \Sigma_\epsilon$, and ϵ_i is assumed independent of $\mathbf{b}_i$. Furthermore, the $J \times J$ covariance matrix Σ_ϵ is assumed to be diagonal, that any correlations found in the observation vector $\mathbf{Y}_i$ are due to their relationship with common $\mathbf{b}_i$ and not due to some spurious correlation between ϵ_i. A continuous latent variable model is assumed for the relationship between $\mathbf{R}_i$ and u_i (model A in Figure 4) with $\pi_{ij} = Pr(r_{ij} = 1)$ representing the probability that the response for subject i at time point j is missing. We apply the logit link for the probability of the missingness, i.e., $log(\frac{\pi_{ij}(u_i,\ \mathbf{X}_{2i})}{1-\pi_{ij}(u_i,\ \mathbf{X}_{2i})}) = u_i - \tau_j \equiv X_{2i}\alpha + Z_{2i}u_i$, where τ_j are unknown parameters for determining an observation at time point j is missing. As discussed earlier, this relationship is equivalent to a random logistic regression, with appropriate design matrices $\mathbf{X}_{2i}$ and Z_{2i}. A latent variable regression, $\mathbf{b}_i = \mathbf{X}_{3i}^{'}\gamma + \zeta_i$, is used to establish the relationship between latent variable $\mathbf{b}_i$ and u_i, where $\mathbf{X}_{3i}^{'} = [\mathbf{X}_{3i}\ \ u_i]$ is a $p_3 + 1$ dimensional vector combining $\mathbf{X}_{3i}$ and u_i, γ is the $(p_3+1) \times q$ unknown regression coefficients for $\mathbf{X}_{3i}^{'}$ and the $q \times q$ matrix Ψ determines variance-covariance structure for error term ζ_i. Finally the latent continuous variable u_i is assumed to be normally distributed with mean 0 and variance σ_u^2.

Note that the maximum likelihood (ML) estimation of the model (4.2) - (4.4) requires the maximization of the observed likelihood, after integrating out missing data $\mathbf{Y}^{mis}$ and latent variables $\mathbf{b}$ and $\mathbf{u}$ from complete-data likelihood function. Detail of the ML estimation technique will be given in next section.

4.3 Maximum Likelihood Estimation

The main objective of this section is to obtain the ML estimate of parameters in the model and standard errors on the basis of the observed data $\mathbf{Y}^{\mathbf{obs}}$ and $\mathbf{R}$. The ML approach is an important statistical procedure which has many optimal properties such as consistency, efficiency, etc. Furthermore, it is also the foundation of many important statistical methods, for instance, the likelihood ratio test, statistical diagnostics such as Cook's distance and local influence analysis, among others. To perform ML estimation, the computational difficulty arises because of the need to integrate over continuous latent factor $\mathbf{u}$, random subject-level effects $\mathbf{b}$, as well as missing responses $\mathbf{Y}^{mis}$. The classic Expectation-Maximization (EM) algorithm provides a tool for obtaining maximum likelihood estimates under models that yield intractable likelihood equations. The EM algorithm is an iterative routine requiring two steps in each iteration: computation of a particular conditional expectation of the log-likelihood (E-step) and maximization of this expectation over the parameters of interest (M-step). In our situations, in addition to the real missing data $\mathbf{Y}^{mis}$, we will treat the latent variables $\mathbf{b}$ and $\mathbf{u}$ as missing data. However, due to the complexities associated with the missing data structure and the nonlinearity part of the model (model A in Figure 4.2), the E-step of the algorithm, which involves the computations of high-dimensional complicated integrals induced by the conditional expectations, is intractable. To solve this difficulty, we propose to approximate the conditional expectations by sample means of the observations simulated from the appropriate conditional distributions, which is known as Monte Carlo Expectation Maximization algorithm. We will develop a hybrid algorithm that combines two advanced computational tools in statistics, namely the Gibbs sampler (Geman and Geman, 1984) and the Metropolis Hastings (MH) algorithm (Hastings, 1970) for simulating the observations. The M-step does not require intensive computations due to the distinctness of parameters in the proposed model. Hense, the proposed algorithm is a Monte Carlo EM (MCEM) type algorithm (Wei and Tanner, 1990). The description of the observed likelihood function is given in the following.

Given the parametric model (4.2) - (4.4) and the i.i.d. $J\times 1$ variables $\mathbf{Y}_i$ and $\mathbf{R}_i$, for $i = 1, \ldots, n$, estimation of the model parameters can proceed via the maximum likelihood method. Let $\mathbf{W}_i = (\mathbf{Y}_i^{obs}, \mathbf{R}_i)$ be the observed quantities, $\mathbf{d}_i = (\mathbf{Y}_i^{mis}, \mathbf{b}_i, u_i)$ be the missing quantities, and $\theta = (\alpha, \beta, \tau_j, \gamma, \Psi, \sigma_u^2, \Sigma_\epsilon)$ be the vector of parameters relating $\mathbf{W}_i$ with $\mathbf{d}_i$ and covariates $\mathbf{X}_i$. With Birch's regularity conditions for parameter vector θ (see Appendix C), the observed likelihood function for the model (4.2) - (4.4) can be written as

$$L_o(\theta|\mathbf{Y^{obs}}, \mathbf{R}) = \prod_{i=1}^{n} f(\mathbf{W}_i|\mathbf{X};\theta) = \prod_{i=1}^{n} \int f(\mathbf{W}_i, \mathbf{d}_i|\mathbf{X}_i;\theta) d\mathbf{d}_i \tag{4.6}$$

where the notation for the integral over $\mathbf{d}_i$ is taken generally to include the multiple continuous integral for u_i and $\mathbf{b}_i$, as well as missing observations $\mathbf{Y}_i^{mis}$. In detail, the above function can be rewritten as following:

$$\begin{aligned}
&L_o(\theta|\mathbf{Y^{obs}}, \mathbf{R}) = \prod_{i=1}^{n} \\
&\iiint \frac{1}{\sqrt{2\pi}}|\Sigma_\epsilon|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i)^T \Sigma_\epsilon^{-1} (\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i)\right\} \\
&\frac{1}{\sqrt{2\pi}}|\Sigma_b|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma)^T \Sigma_b^{-1} (\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma)\right\} \frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left\{-\frac{u_i^2}{2\sigma_u^2}\right\} \\
&\left\{\prod_{j=1}^{J}\left(\frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)}\right)^{r_{ij}} \left(1 - \frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)}\right)^{1-r_{ij}}\right\} du_i d\mathbf{b}_i d\mathbf{Y}_i^{mis}
\end{aligned} \tag{4.7}$$

where $\mathbf{Y}_i^{com} = (\mathbf{Y}_i^{obs}, \mathbf{Y}_i^{mis})$, $\Sigma_b = \sigma_u^2\gamma\gamma^T + \Psi$. As discussed above, the E-step involves complicated, intractable and high dimension integrations. Hence, the Monte Carlo

EM algorithm is applied to obtain ML estimates. Detail of the technique for MCEM will be given in the following section.

4.3.1 Monte Carlo EM

Inspired by the key idea of the EM algorithm, we will treat $\mathbf{d}_i$ as missing data and implement the expectation and maximization (EM) algorithm for maximizing (4.7). Since it is difficult to maximize the observed data likelihood L_o directly, we construct the complete-data likelihood and apply the EM algorithm on the augmented log-likelihood $ln\ L_c(\mathbf{W}, \mathbf{d}|\theta)$ to obtain the MLE of θ over the observed likelihood function $L_o(\mathbf{Y^{obs}}, \mathbf{R}|\theta)$ where it is assumed that $L_o(\mathbf{Y^{obs}}, \mathbf{R}|\theta) = \int L_c(\mathbf{W}, \mathbf{d}|\theta)d\mathbf{d}$. ($\mathbf{W}$ and $\mathbf{d}$ are ensemble matrices for vectors $\mathbf{W}_i$ and $\mathbf{d}_i$ defined in (4.6)). In detail, the EM algorithm iterates between a computation of the expected complete-data likelihood

$$Q(\theta|\hat{\theta}^{(r)}) = E_{\hat{\theta}^{(r)}}\{ln\ L_c(\mathbf{W}, \mathbf{d}|\theta)|\mathbf{Y}^{obs}, \mathbf{R}\} \tag{4.8}$$

and the maximization of $Q(\theta|\hat{\theta}^{(r)})$ over θ, where the maximum value of θ at the $(r+1)$th iteration is denoted by $\hat{\theta}^{(r+1)}$ and $\hat{\theta}^{(r)}$ denotes the maximum value of θ evaluated at the rth iteration. Specifically, r represents the EM iteration. Under regularity conditions the sequence of values $\{\hat{\theta}^{(r)}\}$ converges to the MLE $\hat{\theta}$. (See Wu (1983))

As discussed above, the E-step in our case is analytically intractable, so we may estimate the quantity (4.8) from Monte Carlo simulations. One could notice that the expectation in (4.8) is over the latent variables $\mathbf{d}$. In particular,

$$E_{\hat{\theta}^{(r)}}\{ln\ L_c(\mathbf{W}, \mathbf{d}|\theta)|\mathbf{Y}^{obs}, \mathbf{R}\} = \int ln\ L_c(\mathbf{W}, \mathbf{d}|\theta)g(\mathbf{d}|\mathbf{Y}^{obs}, \mathbf{R}; \hat{\theta}^{(r)})d\mathbf{d}$$

where $g(\mathbf{d}|\mathbf{Y}^{obs}, \mathbf{R}; \hat{\theta}^{(r)})$ is the joint conditional distribution of the latent variables given the observed data and θ. A hybrid algorithm that combines the Gibbs sampler and the MH algorithm is developed to obtain Monte Carlo samples from above conditional distribution. Once we draw a sample $\mathbf{d}_1^{(r)}, \ldots, \mathbf{d}_T^{(r)}$ from the distribution $g(\mathbf{d}|\mathbf{Y}^{obs}, \mathbf{R}; \hat{\theta}^{(r)})$, this expectation can be estimated by the Monte Carlo average

$$Q_T(\theta|\hat{\theta}^{(r)}) = \frac{1}{T}\sum_{t=1}^{T} ln\ L_c(\mathbf{W}, \mathbf{d}_t^{(r)}|\theta) \tag{4.9}$$

where T is the MC sample size and also denotes the dependence of current estimator on the MC sample size. By the law of large numbers, the estimator given in (4.9) converges to the theoretical expectation in (4.8). Thus the classic EM algorithm can be modified into an MCEM where the E-step is replaced by the estimated quantity from (4.9). The M-step maximizes (4.9) over θ.

4.3.2 Execution of the E-step via the Hybrid Algorithm

Let $h(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u})$ be a general function of $\mathbf{Y}^{mis}$, $\mathbf{b}$ and $\mathbf{u}$ that involved in $Q(\theta|\hat{\theta}^{(r)})$, the corresponding conditional expectation given $\mathbf{Y}^{mis}$, $\mathbf{b}$ and $\mathbf{u}$ is approximated by

$$\hat{E}\{h(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u})|\mathbf{Y}^{obs}, \mathbf{R}; \theta\} = \frac{1}{T}\sum_{t=1}^{T} h(\mathbf{Y}^{mis(t)}, \mathbf{b}^{(t)}, \mathbf{u}^{(t)}) \tag{4.10}$$

where $\{(\mathbf{Y}^{mis(t)}, \mathbf{b}^{(t)}, \mathbf{u}^{(t)})\};\ t = 1, \ldots, T$ is a sufficiently large sample simulated from the joint conditional distribution $g(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}; \theta)$. We apply the following three-stage Gibbs sampler to sample these observations. At the tth iteration with current values $\mathbf{Y}^{mis(t)}$,$\mathbf{b}^{(t)}$ and $\mathbf{u}^{(t)}$, (t represents Gibbs sampling iteration)

Step I: Generate $\mathbf{Y}^{mis(t+1)}$ from $f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{R}, \mathbf{b}^{(t)}, \mathbf{u}^{(t)}; \theta)$,

Step II: Generate $\mathbf{b}^{(t+1)}$ from $f(\mathbf{b}|\mathbf{Y}^{obs}, \mathbf{R}, \mathbf{Y}^{mis(t+1)}, \mathbf{u}^{(t)}; \theta)$,

Step III: Generate $\mathbf{u}^{(t+1)}$ from $f(\mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}, \mathbf{Y}^{mis(t+1)}, \mathbf{b}^{(t+1)}; \theta)$.

where function $f(\cdot|\cdot)$ specifies full conditionals that are applied for each step of Gibbs sampler. The full conditional for $\mathbf{Y}^{mis}$ is easily specified due to the conditional independence assumptions between $\mathbf{Y}$ and $\mathbf{R}$, $\mathbf{u}$, given $\mathbf{b}$ as showed in Figure 4. Hence, the full conditional for $\mathbf{Y}^{mis}$ can be simplified as $f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{b}; \theta)$ which is again another normal distribution from the property of conditional distribution of multivariate normal. This conditional can be further simplified in our case due to the assumption of variance-covariance matrix Σ_ϵ in model (4.2) is diagonal. In detail, for subject $i = 1, \ldots, n$, since $\mathbf{Y}_i$ are mutually independent given $\mathbf{b}_i$, $\mathbf{Y}_i^{mis}$ are also mutually independent given $\mathbf{b}_i$. Since Σ_ϵ is diagonal, $\mathbf{Y}_i^{mis}$ is conditionally independent with $\mathbf{Y}_i^{obs}$ given $\mathbf{b}_i$. Hence, it follows from model (4.2) that:

$$f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{b}; \theta) = \prod_{i=1}^{n} f(\mathbf{Y}_i^{mis}|\mathbf{b}_i; \theta)$$

and

$$(\mathbf{Y}_i^{mis}|\mathbf{b}_i; \theta) \sim MVN(\mathbf{X}_{1i}^{mis}\beta + \mathbf{Z}_{1i}^{mis}\mathbf{b}_i,\ \Sigma_{\epsilon,i}^{mis})$$

where $\mathbf{X}_{1i}^{mis}$ and $\mathbf{Z}_i^{mis}$ are submatrices of $\mathbf{X}_{1i}$ and $\mathbf{Z}_i$ with rows corresponding to observed components deleted, and Σ_ϵ^{mis} is a submatrix of Σ_ϵ with the appropriate rows and columns deleted. In fact, the structure of $\mathbf{Y}^{mis}$ may be very complicated with a large number of missing patterns, however, the corresponding conditional distribution only involves a product of relatively simple normal distributions. Hence, the computational cost for simulating $\mathbf{Y}^{mis}$ is low. Due to the hierarchical structure for the model (4.2) - (4.4), the joint distribution that is required in full conditionals for $\mathbf{b}$ and $\mathbf{u}$ can be obtained by multiplying the corresponding densities together, and on the basis of the definition of the model and its assumptions, the following set of full conditionals for $\mathbf{b}$ and $\mathbf{u}$ can be derived: (see Chapter 7, Robert and Casella

(2010))

$$
\begin{aligned}
\mathbf{b}_i|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta \propto exp & \left\{ -\frac{1}{2}(\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i)^T \Sigma_\epsilon^{-1} (\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i) \right. \\
& \left. -\frac{1}{2}(\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma)^T \Psi^{-1} (\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma) \right\} \\
u_i|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta \propto exp & \left\{ -\frac{u_i^2}{2\sigma_u^2} - \frac{1}{2}(\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma)^T \Psi^{-1} (\mathbf{b}_i - \mathbf{X}_{3i}^{'}\gamma) \right\} \\
& \prod_{j=1}^{J} \left(\frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)} \right)^{r_{ij}} \left(1 - \frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)} \right)^{1-r_{ij}}
\end{aligned}
\tag{4.11}
$$

Based on expressions (4.11), it is shown that the associated full conditional distributions for **b** and **u** are not standard and are relatively complex. Hence we choose to apply the M-H algorithm for simulating observations efficiently. The M-H algorithm is one of the classic MCMC methods that has been widely used for obtaining random samples from a target density via the help of a proposed distribution when direct sampling is difficult. Here $p_1(\mathbf{b}_i|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta)$ and $p_2(u_i|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta)$ are treated as the target densities. Based on the discussion given in Robert and Casella (2010), it is convenient and natural to choose $N(\cdot, \sigma^2\Omega)$ as the proposed distributions, where σ^2 is a chosen value to control the acceptance rate of the M-H algorithm, and $\Omega_1^{-1} = \Sigma_b^{-1} + \mathbf{Z}_i^T \Sigma_\epsilon^{-1} \mathbf{Z}_i$ for $\mathbf{b}_i$ and $\Omega_2^{-1} = (\sigma_u^2)^{-1} + \Sigma_b^{-1}$ for u_i. The implementation of M-H algorithm is as follows: at the tth iteration with current value $\mathbf{b}_i^{(t)}$ and $u_i^{(t)}$, new candidates $\mathbf{b}_i^*$ and u_i^* are generated from $N(\mathbf{b}_i^{(t)}, \sigma^2\Omega_1)$ and $N(u_i^{(t)}, \sigma^2\Omega_2)$, respectively. The acceptance of new candidates is decided by the following probabilities:

$$
min\left\{1, \frac{p_1(\mathbf{b}_i^*|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta)}{p_1(\mathbf{b}_i^{(t)}|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta)}\right\}, \quad min\left\{1, \frac{p_2(u_i^*|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta)}{p_2(u_i^{(t)}|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta)}\right\}
$$

where $p_1(\cdot)$ and $p_2(\cdot)$ are calculated from equation (4.11). The quantity σ^2 can be chosen such that the average acceptance rate is approximately 1/4, as suggested by Robert and Casella (2010).

Instead of allowing the candidate distributions for **b** and **u** to depend on the present state of the chain, an attractive alternative is choosing proposed distributions to be independent of this present state, then we get a special case which is named Independent Metropolis-Hastings. To implement this method, we generate candidate for $\mathbf{b}_i$ at step t, $\mathbf{b}_i^*$, from a multivariate normal distribution with mean vector $\mathbf{0}$ and variance covariance Σ_b (denote as the function $h_1(\cdot)$); generate candidate for $\mathbf{u}_i$ at step t, $\mathbf{u}_i^*$, from a univariate normal distribution with mean 0 and variance σ_u^2 (denote as the function $h_2(\cdot)$). The acceptance probability for proposed distributions of $\mathbf{b}_i^{(t+1)}$ and $\mathbf{u}_i^{(t+1)}$ $(i = 1, 2, \ldots, n)$ can be obtained by

$$min\left\{1, \frac{p_1(\mathbf{b}_i^*|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta)\ h_1(\mathbf{b}_i^{(t)})}{p_1(\mathbf{b}_i^{(t)}|\mathbf{Y}_i^{com}, \mathbf{R}_i, u_i; \theta)\ h_1(\mathbf{b}_i^*)}\right\},\quad min\left\{1, \frac{p_2(u_i^*|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta)\ h_2(u_i^{(t)})}{p_2(u_i^{(t)}|\mathbf{Y}_i^{com}, \mathbf{R}_i, \mathbf{b}_i; \theta)\ h_2(u_i^*)}\right\}$$

Let $(\mathbf{Y}_i^{mis(t)}, \mathbf{b}_i^{(t)}, u_i^{(t)});\ t = 1, \ldots, T;\ i = 1, \ldots, n$ be the random samples generated by the proposed hybrid algorithm from the joint conditionals $(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}; \theta)$. Conditional expectations of the complete data sufficient statistics required to evaluate the E-step can be approximated via these random samples as follows: let $\mathbf{Y}_i = (\mathbf{Y}_i^{obs}, \mathbf{Y}_i^{mis})$, and define $Y_i^{(t)} = (Y_i^{obs(t)}, Y_i^{mis(t)})$, where $Y_i^{obs(t)}$ is sampled with replacement from Y_i^{obs},

$$\begin{aligned}
&E[\mathbf{Y}_i - \mathbf{Z}_{1i}\mathbf{b}_i|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}(\mathbf{Y}_i^{(t)} - \mathbf{Z}_{1i}\mathbf{b}_i^{(t)})\\
&E[\epsilon_i\epsilon_i'|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}(\mathbf{Y}_i^{(t)} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i^{(t)})(\mathbf{Y}_i^{(t)} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i^{(t)})'\\
&E[\mathbf{b}_i|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}\mathbf{b}_i^{(t)}\\
&E[\psi_i\psi_i'|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}(\mathbf{b}_i^{(t)} - \mathbf{X}_{3i}'^{(t)}\gamma)(\mathbf{b}_i^{(t)} - \mathbf{X}_{3i}'^{(t)}\gamma)'\\
&E[u_i|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}u_i^{(t)},\quad E[u_iu_i'|\mathbf{Y}_i^{obs}, \mathbf{R}_i; \theta] = T^{-1}\sum_{t=1}^{T}u_i^{(t)}u_i'^{(t)}
\end{aligned} \tag{4.12}$$

where $\mathbf{X}_{3i}'^{(t)} = [\mathbf{X}_{3i}\ \ u_i^{(t)}]$.

4.3.3 Maximization Step

At the M-step we need to maximize $Q(\theta|\theta^{(\mathbf{r})})$ with respect to θ. In other words, the following systems are needed to be solved:

$$\frac{\partial Q(\theta|\theta^{(\mathbf{r})})}{\partial\theta} = E\{\frac{\partial}{\partial\theta} lnL_c(\mathbf{W},\mathbf{d}|\theta)|\mathbf{Y}^{obs},\mathbf{R};\theta^{(r)}\} = 0 \tag{4.13}$$

It can be shown that

$$\begin{aligned}
\frac{\partial lnL_c(\mathbf{W},\mathbf{d}|\theta)}{\partial\beta} &= \sum_{i=1}^{n}\mathbf{X}_i^T\Sigma_\epsilon^{-1}(\mathbf{Y}_i-\mathbf{Z}_{1i}\mathbf{b}_i-\mathbf{X}_{1i}\beta)\\
\frac{\partial lnL_c(\mathbf{W},\mathbf{d}|\theta)}{\partial\Sigma_\epsilon} &= \frac{1}{2}\Sigma_\epsilon^{-1}\sum_{i=1}^{n}\left[(\mathbf{Y}_i-\mathbf{X}_{1i}\beta-\mathbf{Z}_{1i}\mathbf{b}_i)(\mathbf{Y}_i-\mathbf{X}_{1i}\beta-\mathbf{Z}_{1i}\mathbf{b}_i)^T-\Sigma_\epsilon\right]\Sigma_\epsilon^{-1}\\
\frac{\partial lnL_c(\mathbf{W},\mathbf{d}|\theta)}{\partial\gamma} &= \sum_{i=1}^{n}u_i\Psi^{-1}(\mathbf{b}_i-\mathbf{X}_{3i}^{'}\gamma)\\
\frac{\partial lnL_c(\mathbf{W},\mathbf{d}|\theta)}{\partial\Psi} &= \frac{1}{2}\Psi^{-1}\sum_{i=1}^{n}\left[(\mathbf{b}_i-\mathbf{X}_{3i}^{'}\gamma)(\mathbf{b}_i-\mathbf{X}_{3i}^{'}\gamma)^T-\Psi\right]\Psi^{-1}\\
\frac{\partial lnL_c(\mathbf{W},\mathbf{d}|\theta)}{\partial\alpha} &= \sum_{i=1}^{n}\sum_{j=1}^{J}\left\{r_{ij}X_{2ij}-\frac{exp(X_{2ij}\alpha+Z_{2ij}u_i)}{1+exp(X_{2ij}\alpha+Z_{2ij}u_i)}\cdot X_{2ij}\right\}
\end{aligned} \tag{4.14}$$

Due to distinctness of parameters in the model, the ML estimates can be obtained separately: for β and Σ_ϵ in the linear mixed model, as well as γ and Ψ in latent variable regression model, the corresponding ML estimates can be obtained from sufficient statistics in the E-step, which is given in (4.12); to estimate α, we will implement a quasi-Newton method because of no closed expression; the estimates of Σ_b and σ_u can be obtained from simulated random samples by applying law of total variance.

With the assumption that the missing mechanism is ignorable given latent factors **u**, and **b**, the computation of proposed MCEM algorithm can be further reduced. That is, the ML estimates can be obtained from observed components in **Y**, given information of **u**, and **b**. Specifically, the dimension of integration in E-step will reduced to two, instead of three.

4.3.4 Monitor Convergence of MCEM via Bridge Sampling

In order to obtain valid ML estimates, one needs to investigate the convergence of the EM algorithm. However, in our case, determining the convergence of the MCEM algorithm is not straightforward. Meng and Schilling (1996) pointed out that the log-likelihood function can 'zigzag' along the iterates even without implementation or numerical errors, due to the variability introduced by simulation at the E-step. Further to evaluate the observed-data log-likelihood function, some numerical method has to be used because of a closed forms is lacking. In the absence of accurate evaluation of the observed-data log-likelihood function, we could not judge whether any large fluctuation is due to the implementation errors, to the numerical errors in computing the log-likelihood values, or to non-convergence of the MCEM algorithm. We will implement bridge sampling to solve this problem, as suggested by Meng and Schilling (1996).

In the determination of the convergence of a likelihood function, only the evaluation changes in likelihood are of interest, and these changes can be expressed by the logarithm of the ratio of two consecutive likelihood values. In our case, the ratio is given by

$$K(\theta^{(r+1)}, \theta^{(r)}) = log \frac{L_o(\mathbf{Y}^{obs}, \mathbf{R}|\theta^{(r+1)})}{L_o(\mathbf{Y}^{obs}, \mathbf{R}|\theta^{(r)})}$$

Due to the complexity of the observed likelihood function, the accurate value of $K(\theta^{(r+1)}, \theta^{(r)})$ is difficult to obtain. However, as pointed out by Meng and Schilling (1996), it can be approximated by

$$\begin{aligned} \hat{K}(\theta^{(r+1)}, \theta^{(r)}) = log & \left\{ \sum_{t=1}^{T} \left[\frac{L_c(\mathbf{W}, \mathbf{d}^{r,(t)}|\theta^{(r+1)})}{L_c(\mathbf{W}, \mathbf{d}^{r,(t)}|\theta^{(r)})} \right]^{\frac{1}{2}} \right\} \\ -log & \left\{ \sum_{t=1}^{T} \left[\frac{L_c(\mathbf{W}, \mathbf{d}^{r+1,(t)}|\theta^{(r)})}{L_c(\mathbf{W}, \mathbf{d}^{r+1,(t)}|\theta^{(r+1)})} \right]^{\frac{1}{2}} \right\} \end{aligned} \tag{4.15}$$

where $\mathbf{d}^{r,(t)}$, $t = 1, \ldots, T$ are random samples generated from $g(\mathbf{d}|\mathbf{W}, \theta^{(r)})$ by the hybrid algorithm. In determining the convergence of the MCEM algorithm, we plot $\hat{K}(\theta^{(r+1)}, \theta^{(r)})$ against iteration index r. Approximate convergence is claimed to be achieved if the plot shows a curve converging to zero.

4.3.5 Standard Error Estimates

Standard error estimates of the ML estimates can be obtained by inverting the Hessian matrix or the information matrix of the log-likelihood function based on observed data $\mathbf{Y}^{obs}$ and missing pattern matrix $\mathbf{R}$. Unfortunately, these matrices don't have closed forms. Thus, we apply the formula by Louis (1982) formula and random samples generated from $g(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}, \theta)$ via the hybrid algorithm to obtain standard error estimates. From Louis (1982) we have

$$\begin{aligned}-\frac{\partial^2 L_o(\mathbf{Y}^{obs}, \mathbf{R}|\theta)}{\partial\theta\partial\theta^T} &= E\left\{-\frac{\partial^2 L_c(\mathbf{Y}^{obs}, \mathbf{R}, \mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\theta)}{\partial\theta\partial\theta^T}\right\} \\ &\quad - Var\left\{\frac{\partial L_c(\mathbf{Y}^{obs}, \mathbf{R}, \mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\theta)}{\partial\theta}\right\}\end{aligned} \tag{4.16}$$

The above expectation involved calculations of expectation and variance with respect to the conditional distribution of $(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u})$ given $\mathbf{Y}^{obs}$, $\mathbf{R}$ and θ, and the whole expression is evaluated at $\hat{\theta}$. Again, it is difficult to evaluate the above expression in closed forms; however, they can be approximately by the sample mean and sample variance-covariance matrix of the distinct random sample $\{(\mathbf{Y}^{mis(t)}, \mathbf{b}^{(t)}, \mathbf{u}^{(t)});\ t = 1, \ldots, T_1\}$ generated separately from $g(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}, \hat{\theta})$ using the hybrid algorithm. Let $\mathbf{W} = (\mathbf{Y}^{obs}, \mathbf{R})$ and $\mathbf{d} = (\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u})$, we have

$$\begin{aligned}&-\frac{\partial^2 L_o(\mathbf{Y}^{obs}, \mathbf{R}|\theta)}{\partial\theta\partial\theta^T} = T_1^{-2}\left(\sum_{t=1}^{T_1}\frac{\partial L_c(\mathbf{W}, \mathbf{d}^{(t)}|\theta)}{\partial\theta}\right)\left(\sum_{t=1}^{T_1}\frac{\partial L_c(\mathbf{W}, \mathbf{d}^{(t)}|\theta)}{\partial\theta}\right)^T\Bigg|_{\theta=\hat{\theta}} \\ &+ T_1^{-1}\sum_{t=1}^{T_1}\left\{-\frac{\partial^2 L_c(\mathbf{W}, \mathbf{d}^{(t)}|\theta)}{\partial\theta\partial\theta^T} - \left(\frac{\partial L_c(\mathbf{W}, \mathbf{d}^{(t)}|\theta)}{\partial\theta}\right)\left(\frac{\partial L_c(\mathbf{W}, \mathbf{d}^{(t)}|\theta)}{\partial\theta}\right)^T\right\}\Bigg|_{\theta=\hat{\theta}}\end{aligned} \tag{4.17}$$

Finally, the standard errors are obtained from the diagonal elements of inverse Hessian matrix $-\partial^2 L_o(\mathbf{Y}^{obs}, \mathbf{R}|\theta)/\partial\theta\partial\theta^T$, evaluated at $\hat{\theta}$.

4.4 An Empirical Simulation Study for Obtaining MLEs

To study the performance of the proposed model and sensitivity of the model assumptions, we simulated data using different assumptions and fit different models to investigate how much the results from these models change accordingly. We conducted a simulation study to evaluate the performance of the proposed model (4.2) - (4.4). In this simulation we generated missing indicators for 500 individuals from model (4.5), with the known fixed effects and random effects, so that approximately 52% of the subjects had missing values, and 48 different missing patterns in a 6 time points study. We removed 8 individuals that didn't have any observed values, and kept the remaining 492 individuals in the study. Given the fixed effects, random effects, error variance-covariance as well as link parameters, we generated the growth-curve data and removed observations for each subject to be missing based on the observed missing indicators. Once the simulation data was generated using the true known parameters associated with the underlying model, we fitted the proposed model (4.2) - (4.4) to the data.

In this simulation, the true underlying model was

$$\begin{aligned}
&Y_{ij} = 1.00 + 2.00t_{ij} + 1.00X_1 + 0.5X_2 + b_i + \epsilon_{ij}\\
&b_i = 0.6u_i + \zeta_i\\
&logit(\pi_{ij}) = u_i - (3.5,\ 3,\ 2.5,\ 2,\ 1.5,\ 1)\ I_{ij}
\end{aligned}$$

where π_{ij} is the missing probability for subject i at time point j, i.e. $\pi_{ij} = P(r_{ij} = 1)$; t_{ij} is the jth visiting time for subject i; $\tau = (3.5,\ 3,\ 2.5,\ 2,\ 1.5,\ 1)^T$ is true values for time location parameters, that is we assume an individual has a higher missing probability at the later stage of the study, I_{ij} is a 1×6 vector with the jth element 1, 0 elsewhere. In this simulation, we also allow the missing mechanism to depend on a subject-level latent random effect u_i, with normal distribution of mean 0 and variance 2. This unobserved random effect further influences the growth-curve model via the specified link model, which in this simulation we consider influences on subject-level random intercept in the growth-curve model. Parameters in the link model and growth-curve model are given as follows: $\epsilon_{ij} \sim N(0, 0.5)$, $u_i \sim\ N(0, 2)$, $\zeta_i \sim\ N(0, 0.28)$. It can be shown that the subject-level random effects b_i has variance 1, based on the link model (4.3).

The total number of unknown parameters in this simulation study was 19. ML estimates were obtained by fitting proposed model (4.2) - (4.4). The proposed MCEM algorithm was used to produce the ML estimates and standard errors estimates in 100 replications. In the MH algorithm of the E-step, we set proposed distribution to be independent of chain state. The number of observations generated from the conditional distribution $g(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}; \theta)$ via the hybrid algorithm for completing the E-step at the rth iteration of the MCEM algorithm was $50 + 10r$. This number was increased with the EM iteration and was larger near convergence where parameters values in the conditional distribution were closer to the ML estimates. Starting values for variance elements were all set to 1.0 and starting values for the remaining unknown parameters were 0.0. The convergence of model fitting procedure was assessed by plotting log-likelihood ratio versus EM iteration, see Figure 4.3 for a summary of convergence in a randomly selected replication. We observed that the log-likelihood ratio K of the bridge sampling is sufficiently small after 100 iterations for all replications. To be conservative, we took the parameters values at the 150th iteration as the ML estimates in all the replications of the simulation study. Finally the standard error estimates were calculated from Equation 4.17 on $3,000$ observations simulated from $g(\mathbf{Y}^{mis}, \mathbf{b}, \mathbf{u}|\mathbf{Y}^{obs}, \mathbf{R}; \hat{\theta})$ by the hybrid algorithm with 100

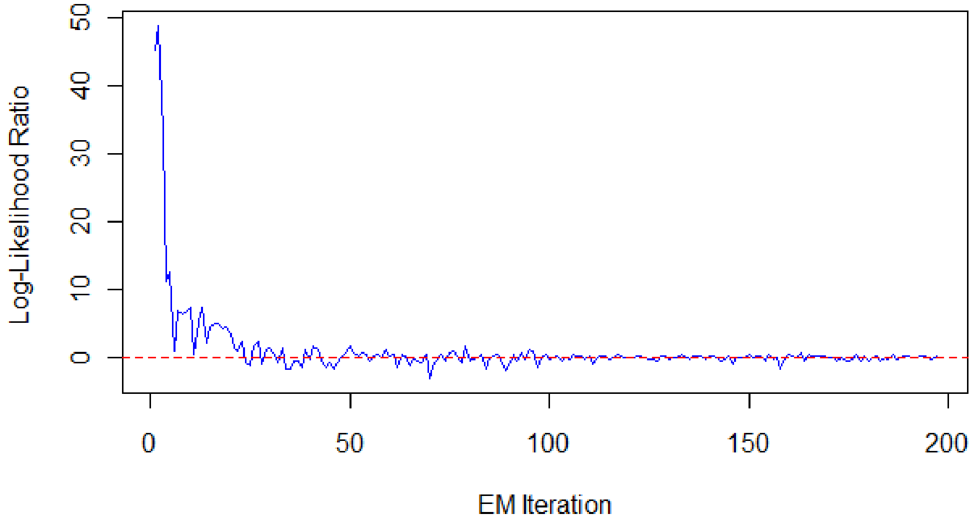


Figure 4.3: Log-likelihood ratio versus EM iteration from the third iteration

burn-in iterations. Based on 100 replications, the mean of the estimates and the mean of the standard errors were computed and given in Table 4.1. We observed that the mean estimates are quite close to the true values, although the true parameters in the missing model are slightly different from default values that used to generate missing indicators, due to 8 individuals excluded from the simulation study. The convergence trace plot for fixed effects in growth-curve model was given in Figure (4.4).

4.5 Bayesian Approach for Model Estimation

In the previous section, we presented a maximum likelihood approach to obtain estimates for model (4.2) - (4.4). However, for small sample sizes, likelihood-based inference can be unreliable with variance components being particularly difficult to estimate. Meanwhile, the properties of ML estimators can be only guaranteed on a large sample size. Even worse, the computation of MCEM could be tedious, be-

Table 4.1: ML estimates of the parameters in the simulation study

	Parameters	True Value	Proposed Model	Standard Error
Growth-curve Model	I	1.00	0.992	0.012
	S	2.00	2.001	0.010
	X_1	1.00	1.005	0.015
	X_2	0.50	0.512	0.020
	$\sigma^2_{b_0}$	1.00	0.989	0.070
	$\sigma^2_{\epsilon_1}$	0.50	0.506	0.042
	$\sigma^2_{\epsilon_2}$	0.50	0.549	0.044
	$\sigma^2_{\epsilon_3}$	0.50	0.44	0.037
	$\sigma^2_{\epsilon_4}$	0.50	0.561	0.045
	$\sigma^2_{\epsilon_5}$	0.50	0.412	0.039
	$\sigma^2_{\epsilon_6}$	0.50	0.474	0.047
Linked Model	γ	0.60	0.625	0.304
	ψ	0.28	0.264	0.091
Missing Model	τ_1	3.32	3.295	0.221
	τ_2	3.15	3.148	0.214
	τ_3	2.65	2.582	0.195
	τ_4	2.24	2.198	0.182
	τ_5	1.71	1.699	0.168
	τ_6	1.18	1.158	0.157
	σ^2_u	1.88	1.856	0.732

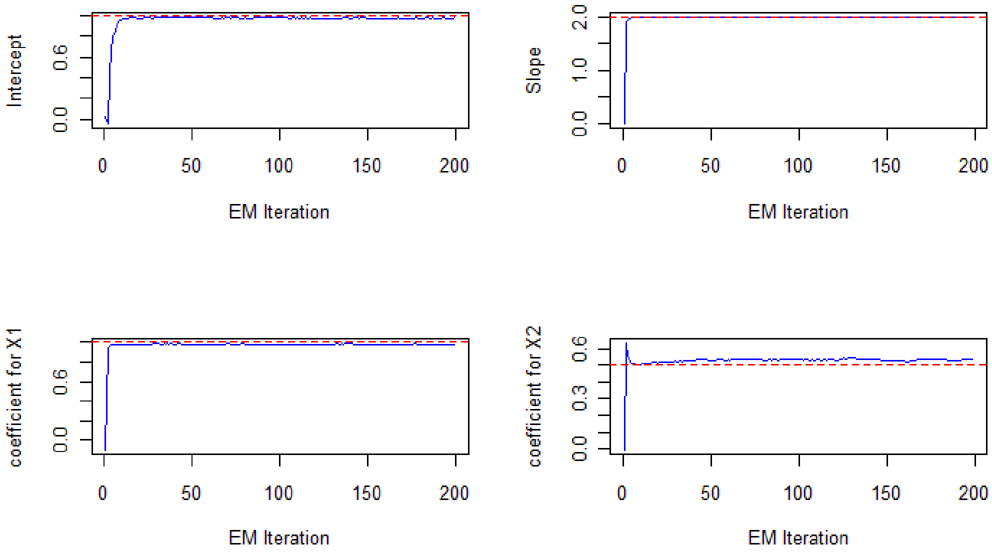


Figure 4.4: Convergence plot for fixed effects in growth-curve model. True values were plotted as dot line.

cause in each iteration, new variation will be introduced by the Monte Carlo scheme. The convergence of MCEM typically cannot achieve the expected difference between two consecutive iterations. Instead, one needs to monitor the convergence trace of MCEM and terminate the implementation if a stable fluctuation along a fixed value is present. For example, one wants to determine the convergence of MCEM via monitoring value changes of the log-likelihood function. The MCEM could be terminated if a convergence plot shows a stable fluctuation around 0, but this waiting time will be long, depending on model complexity. In the previous empirical simulation study, the average computation time for one replication is more than 2 hours. (The program was implemented on a Macintosh machine with Processor 2.8GHz, Intel Core i7.) One approach to improve the computation efficiency is to choose appropriate starting values. The estimates which are obtained from ignorable likelihood approach will be an ideal option for initial values for MCEM algorithm. As an alternative, a Bayesian approach is appealing and worth to be further explored. In this section, we will present the basic idea of Bayesian methods and a Bayesian approach based on Markov Chain Monte Carlo (MCMC) method for model (4.2) - (4.4).

4.5.1 Basic Ideas of Bayesian Inference

Bayes Theorem

Bayesian analysis is based on assumptions that the concept of probability can be applied to the degree to which a person believes a hypothesis or proposition. The degree of belief in proposition H can be represent as $Pr(H)$. Here we adopt the same notation from a published work by Zhang and Hamagami (2007). $Pr(H)$ is also known as the prior degree of belief in H. A conventional Bayes theorem states,

$$Pr(H|E) = \frac{Pr(E \cap H)}{Pr(E)} = \frac{Pr(E|H)Pr(H)}{Pr(E)},$$

which indicates that the degree of belief in H given the observed evidence E is equal to the ratio between joint probability of H and E and the probability of E. $Pr(H|E)$ is known as posterior degree of belief in H, in the sense of being the updated belief after observing the evidence.

In most cases, one will have more than one hypothesis in research. For instance, if we have N different hypotheses, $H_1,\ H_2, \ldots, H_N$ to account for a phenomenon, then Bayes theorem is given as

$$Pr(H_i|E) = \frac{Pr(E|H_i)Pr(H_i)}{\sum_{i=1}^{N} Pr(E|H_i)Pr(H_i)}$$

The above expression explains that the posterior belief on H_i not only depends on the observed evidence E but also depends on our prior beliefs regarding each hypothesis.

Bayes theorem is useful because it provides a tool to calculate the probability of a hypothesis based on the evidence or data. After obtaining the evidence, the calculation of $Pr(E|H_i)$ is straightforward. However, when we observe some evidence or collect some data, we are interested in the probability of the hypotheses conditional on the evidence, $Pr(H_i|E)$. Bayes theorem provides a way to calculate this probability by noticing that this calculation also depends on the prior probabilities $Pr(H_i)$. Hence, Bayes theorem provides a natural way to update prior belief $Pr(H_i)$ concerning the hypothesis to posterior belief $Pr(H_i|E)$ based on the evidence E that we collected.

In parallel, the hypotheses can be represented by one or more continuous parameters from a model denoted by θ for a continuous probability setting. Assume the evidence, also known as data, is denoted by $\mathbf{Y}$. Bayes theorem can be rewritten as,

$$p(\theta|\mathbf{Y}) = \frac{p(\theta)p(\mathbf{Y}|\theta)}{p(\mathbf{Y})} = \frac{p(\theta)p(\mathbf{Y}|\theta)}{\int_\theta p(\theta)p(\mathbf{Y}|\theta)d\theta}$$

in which $p(\theta)$ is a prior probability distribution of θ, $p(\theta|\mathbf{Y})$ is the posterior probability distribution of θ, and $p(\mathbf{Y}|\theta)$ is the probability of the data which is also known as the likelihood $L(\theta;\mathbf{Y})$ in maximum likelihood estimations (MLE). In Bayesian framework, $\int_\theta p(\theta)p(\mathbf{Y}|\theta)d\theta$ is a normalized constant, hence in most situations, we will express the relationship between the posterior and prior distributions as follows:

$$p(\theta|\mathbf{Y}) \propto p(\theta)p(\mathbf{Y}|\theta) = p(\theta)L(\theta;\mathbf{Y}),$$

which states that a posterior is proportional to the prior times the likelihood.

Choice of Priors

Bayes theorem shows that the prior belief is required for Bayesian analysis. A prior is the available information or knowledge about the hypothesis and unknown parameters before the data are collected and should be specified in advance. The prior is classified as either an informative prior or a non-informative prior.

When no reliable prior information or knowledge concerning the hypotheses or parameters exists, or an inference based only on the data at hand is desired, non-informative priors can be used. A non-informative prior does not favor any hypothesis or value of a parameter. For example, for a discrete distribution, the prior $Pr(H_i) = 1/N,\ i = 1, \ldots, N$ is a non-informative prior because it assigns equal probability to each hypothesis H_i. Similarly, for the continuous case, one could assign a non-informative prior as $\pi(\theta) = c,\ any\ c > 0$. This prior is usually called an improper prior because its integration is infinity. Further, priors with little information about the unknown parameters are also called non-informative priors. For example, researchers sometimes give a wide variance range for a normal prior. In this case, a large variance will provide vague information. In the Bayesian framework, the use of non-informative priors typically yields similar results to MLE.

In another perspective, informative priors make Bayesian analysis more subjective because different priors can result in different conclusions, which is a situation that has been criticized by frequentists for a long time. An informative prior may be constructed from previous studies. For example, if one want to predict tomorrow's temperature, it is reasonable to use a normal distribution prior with the mean and variance equal to the mean and variance of the temperature on the same day over the past 20 years (An example from Zhang and Hamagami (2007)). Intuitively, the use of priors provides a method to utilize current knowledge to a future study. For instance, before any experiment is carried out, we may know nothing about a parameter and thus specify a non-informative prior $p(\theta)$. After an experiment in which we obtain the data $\mathbf{Y}_1$, we update our knowledge about the parameter to $p(\theta|\mathbf{Y}_1)$. With an additional experiment, we obtain the data $\mathbf{Y}_2$, and can use the posterior $p(\theta|\mathbf{Y}_1)$ from the first experiment as the prior to update the knowledge about that parameter again.

Regardless of whether informative priors are adopted, many investigators prefer to using conjugate priors when they are appropriate to simplify computation. A conjugate prior is a prior from the family of probability density functions from which the derived posterior density functions have similar function forms to the priors. For instance, a normal prior will to lead a normal posterior based on the Bayes theorem, then this prior is a conjugate prior. The use of conjugate priors can reduce the computation complexity of the posterior distribution largely. The exponential family, which includes the normal distribution, gamma distribution, beta distribution, and so on, is the most often used family of distributions and has conjugate priors.

Statistical Inference on Posteriors

Once the posterior distribution of the parameters is obtained, statistical inference can be performed. Since the posterior distribution of the unknown parameters are steadily revealed by Bayesian analysis, we can demonstrate their densities in plots. However, such plots carry so much information that they become difficult to comprehend. Several statistics can be used to summarize the information of the posterior and are analogous to parameter estimates and standard errors from MLE. In particular, we consider point estimation and credible intervals.

Of the many point estimations, the mean is the most widely used statistic. Given the posterior, the mean is calculated by

$$\bar{\theta} = \int \theta p(\theta|\mathbf{Y}) d\theta \tag{4.18}$$

which is the classical definition of the mean. Similarly, the associated variance can be obtained with

$$Var(\theta) = \int (\theta - \bar{\theta})(\theta - \bar{\theta})' p(\theta|\mathbf{Y}) d\theta \tag{4.19}$$

These two terms are also referred to as posterior mean and posterior variance, respectively.

In Bayesian statistics, credible intervals are used for purposes similar tho those of confidence intervals in frequentist statistics. Formally, a $100\times\ (1-\alpha)\%$ credible interval for θ is obtained by

$$1-\alpha \leq \int_{L}^{U} p(\theta|\mathbf{Y})d\theta \tag{4.20}$$

where L and U are lower and upper bounds, respectively.

One has to pay attention to the interpretation of credible intervals. Because the parameter θ is considered a random variable, we can interpret the credible interval as "The probability that θ lies in the interval (L,U) given the observed data is at least $100\times\ (1-\alpha)\%$." In frequentist statistics, the confidence interval means that "If the experiment is repeated many times and the confidence interval is calculated each time, then overall $100\times\ (1-\alpha)\%$ of them contain the true parameter θ." Thus, the credible interval has a more intuitively appealing interpretation.

Markov Chain Monte Carlo methods

Statistical inference presented above can be done when the integration in equations $(4.18)-(4.20)$ can be solved analytically. However, this is usually impossible in practice especially when multiple unknown parameters are present. In practice, Markov Chain Monte Carlo (MCMC) methods are generally used to circumvent the difficulty of multiple dimension integration. Different versions of MCMC methods have been proposed, such as Metropolis-Hastings (M-H) sampling, Gibbs sampling, and slice sampling. For model estimation within Bayesian framework, we focus on Gibbs sampling scheme.

Gibbs sampling is an numerical implementation to generate a data point from the conditional distribution of each parameter, conditional on the current values of the other parameters. Here is a procedure in detail: let $\theta = (\theta_1, \theta_2, \dots, \theta_K)$ be K unknown parameters in the model of interest. The full conditional distribution (or referred as conditional density function) $\pi(\theta_k|\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_K; \mathbf{Y})$ for θ_k can be obtained directly from standard manipulations on probability density/mass functions. Then we can use following scheme to sample the data points from the conditional distributions: at the $(t+1)$th iteration with current value $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_K^{(t)})$, update $\theta^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_K^{(t+1)})$ by means of sequentially generating

$$\begin{gathered}
\theta_1^{(t+1)} \text{ from } \pi(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_K^{(t)}; \mathbf{Y}) \\
\theta_2^{(t+1)} \text{ from } \pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_K^{(t)}; \mathbf{Y}) \\
\vdots \\
\theta_K^{(t+1)} \text{ from } \pi(\theta_K|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{K-1}^{(t+1)}; \mathbf{Y})
\end{gathered}$$

From this updating scheme, the first parameter is updated on values of parameters from the previous iteration. The second parameter is updated based on the just-updated first parameter estimate and the not-yet-updated third to Kth parameters. This process of updating parameters is performed up to the Kth parameter to finish one complete iteration. The iteration process above can be repeated T times. Geman and Geman (1984) showed that for sufficiently large T, $\theta^{(T)}$ can be viewed as a simulated observation from the posterior distribution $\pi(\theta|\mathbf{Y})$. The simulated observations after T iterations are recorded, and for convenience, we denote as $\theta^t, \ t = 1, \dots, T$. Sometimes, there are highly positive autocorrelation between consecutive iterations. To reduce autocorrelation and computing memory space, one could pick points with a fixed interval (or thinning process) a indexed $1, 1+a, 1+2a, 1+3a, \dots$ to perform further analysis. The point estimation is calculated by

$$\bar{\theta} = \frac{1}{T} \sum_{t=0}^{T} \theta^{1+ta},$$

with variance expression

$$Var(\bar{\theta}) = \frac{1}{T-1}\sum_{t=0}^{T-1}(\theta^{1+ta} - \bar{\theta})(\theta^{1+ta} - \bar{\theta})^T$$

To construct the credible interval, one could use the percentiles of the generated sequences. For instance, the lower bound of the $100 \times (1-\alpha)\%$ credible interval is equal to the $\alpha/2$ percentile of the sequence and the upper bound is equal to the $1-\alpha/2$ percentile. To determine the convergence of the generated Markov chain, or equivalently determine T, the typical approach that we use is "eyeball" method, i.e., monitoring the convergence by visually inspecting the history plots of the generated sequences.

4.5.2 Specification of Priors

With a brief overview of Bayesian analysis, we will in the following demonstrate a full Bayesian scheme to achieve parameter estimations for models (4.2) - (4.4). In section 4.3 we derived the observed likelihood function and completed tasks to find conditional distribution for parameters of interests. Formulas (4.7), (4.11), as well as Gibbs sampling scheme, had been utilized to find MLEs in a previous section, and further can be adopted here to find corresponding Bayesian estimates. In addition, one needs to specify a prior for each parameter in models (4.2) - (4.4), in order to invoke a full Bayesian approach. In the following, we will focus on specifying priors.

As we discussed earlier, conjugate priors are substantially able to reduce computation burdens since they provide the same distribution family with posteriors. The conditional independence assumptions of models (4.2) - (4.4) further break down the complexity of those models and make posterior calculation feasible and more easier. Hence, we will adopt conjugate priors for each parameter in each simulation study. In particular, for all regression coefficients, we assign them normal priors with means 0 and large variances 10^3. For variance components, we assign inverse-gamma priors for single variance components, and inverse-Wishart priors for variance-covariance structures. In the simulation studies, all priors are given in the sense of providing vague knowledge on parameters of interests, which guarantees the comparability among models from MLE approaches and proposed models (4.2) - (4.4) from full Bayesian approaches. In real case applications, we will adopt different priors, including diffuse priors and informative priors, in order to demonstrate whether the influence from prior knowledge dominate the conclusions. All Bayesian approaches and investigations are performed with a combination of a widely-used free software WinBUGS (Lunn and Spiegelhalter, 2000), and a R cran package 'R2WinBUGS' (Sturtz and Gelman, 2005).

Chapter 5

APPLICATIONS

In this chapter, we will present results of simulation studies and illustrate several applications.

5.1 Simulation Studies

To study the effectiveness of the continuous latent factor model (CLFM), we simulated data that includes non-ignorable missingness from Diggle-Kenward selection model and fitted different models to investigate how much the results changed accordingly. Firstly, three simulation studies were generated with 500 replicates in each simulation, as follows. Given the known fixed effects, random effects, and link parameter values, plus the random error covariances, we generated missing values for each subject in the study. For sample size, we included two different sizes, a moderate sample size 300, as well as a small sample size 80 in the first three simulations. That is, we simulated data from baseline and at follow-up times that were observed. The total length of time in the study was six time points. Once each replicate was generated using the true known parameter values associated with the underlying model, three models were fitted and compared, including classic model where missing data are excluded from estimation, Roy's model, and CLFM model. Since Roy's model usually requires a larger sample size to obtain estimation convergence, the fourth simulation study was conducted with 1000 subjects in each replicate, and total number of replicates was 200. The simulation model was the same as the first two simulation studies.

For the first two simulation studies, the true underlying parameters for repeated measures were as follows,

$$Y_{ij} = 3.00 + 1.00(Visit\,time) + 2.00(Age) + 1.00(Treatment) + b_{0ij} + b_{1ij}(Visit\,time) + \epsilon_{ij}$$

where age is a standardized continuous variable with mean 0 and variance 1 at baseline, subjects are randomly assigned into two treatment groups with a $1:1$ ratio. $b_{0ij} \sim N(0,1)$, $b_{1ij} \sim N(0,0.2)$, $cov(b_{0ij}, b_{1ij}) = -0.3$ and $\epsilon_{ij} \sim N(0,0.5)$. For each simulation study, we fit three different models, including a classic linear mixed model by ignoring missingness, Roy's model and continuous latent factor model (CLFM) via full Bayesian approach. One needs to know that we only include an empirical study for CLFM via MCEM approach in the last chapter due to its long computation time. In the first simulation study, repeated measures are missing with lower missing probabilities. Figure 5.1 describes the missing proportion and missing patterns from two scenarios, which corresponds to two sample sizes. In this plot, there are two colors: blue represents for observed measures or responses, and red means a response is missing. The observation time is listed from bottom to top, and each column in the figure is the record of a subject. The one on the left is the missing pattern plot of the first replicate from a simulation with 80 subjects in the study; by comparison, the plot on the right is the first one with 300 subjects. As shown in this plot, lots of subjects do not have any missing values. In both simulation studies, around half individuals follow the study at all times. For the case with 80 subjects, there are 40 complete cases; and for the 300 individual study, a total 146 do not have any missing values. On average, the proportions of missing values for both studies are around 13 percent.

Figures 5.2 and 5.3 summarize the results for the fixed effects on 80 subjects' study and 300 subjects' study, respectively. Since there is only a small proportion of missingness on repeated-measures, and few subjects drop out from the study, all three

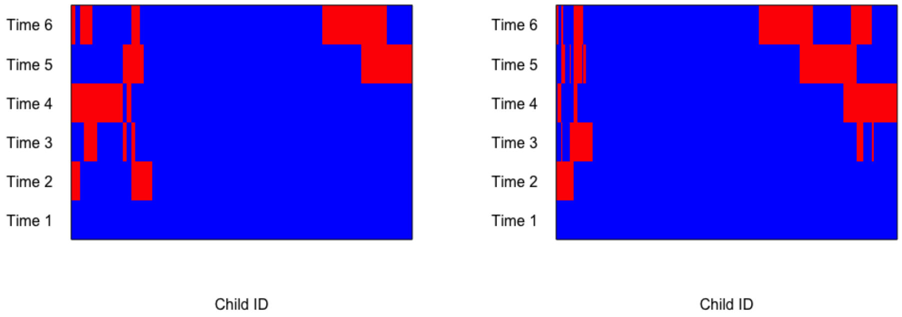


Figure 5.1: Missing proportion and missing patterns of simulation studies on repeated-measure model with lower missing probability. Blue color represents observed measures and red color means missing measures. Missing observations are non-ignorable which are generated from Diggle-Kenward selection model. Two different sample sizes were simulated: 80 individuals (left) and 300 individuals (right)

models provide reasonable point estimates for fixed effects, as well as corresponding standard errors for both scenarios. However, one can determine, with a closer investigation, CLFM produces more accurate estimates when comparing with other two models, in terms of mean square error (MSE) or root mean of square error (RMSE). For instance, if we consider the effect of time in this study, that is, estimating the regression coefficient of time in the model, the RMSE from CLFM is 0.089, which is much smaller than the RMSE from the ignorable model, 1.981. (These two numbers are from the study with 80 subjects. For 300 subjects, we have two similar RMSEs: the one from CLFM is 0.069, and the other one from MAR is 1.982.) Further, Roy's model underestimates the size of the age effect, as well as average change rate on response (Time). In estimating procedures, Roy's model does not converge on many replications with a smaller sample size, which contributes to smaller standard errors in the study of 80 subjects. These shortened confidence intervals also exist in es-

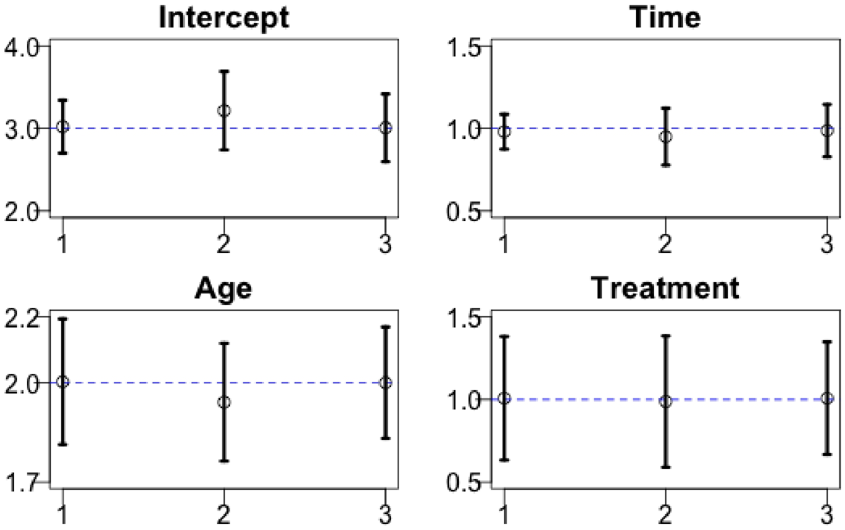


Figure 5.2: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with lower missing probability. The study sample size is 80. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

timating random effects, which include random intercept, random slope, as well as covariance between random intercept and slope.

Figure 5.4 describes estimates and corresponding 95 percent confidence interval (credible interval for Bayesian estimates from CLFM) for random effects in the simulated model, when a repeated-measure model with 80 subjects is of interest. In this result, the missing proportion is close to 17 percent. For this case, one can see intervals from Roy's model cannot cover true variances of random intercept and slope, while ignorable likelihood approach and CLFM could generate more accurate point estimates and corresponding variabilities. Specifically, the point estimate on variance for the random intercept from the ignorable likelihood approach is 0.940; from Roy's

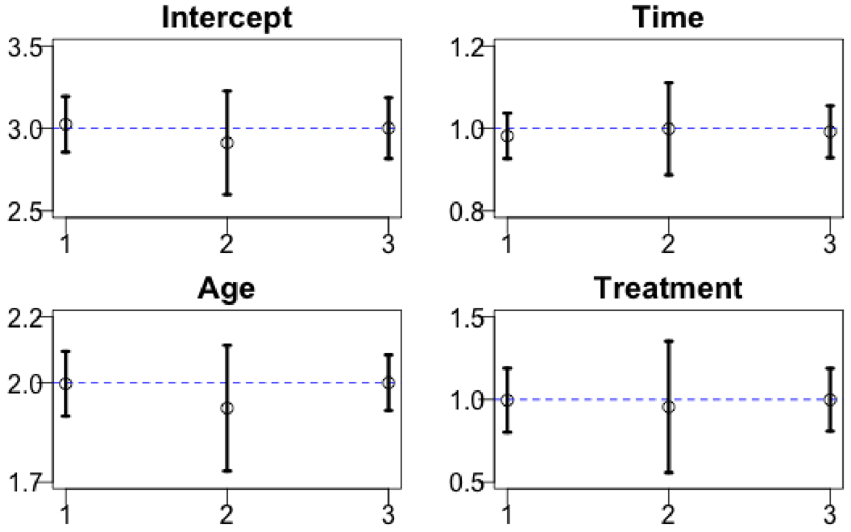


Figure 5.3: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with lower missing probability. The study sample size is 300. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

model with two latent classes is 0.632, and from CLFM is 1.008, given true variance of random intercept term is 1. Similarly, for estimating the variance of the random slope term, (underlying true value in the simulation is 0.2) point estimates from three models are 0.191, 0.124, 0.196, respectively. In estimation of covariance, CLFM also gives the best performance. Similar results can be obtained for estimating random effects with a larger sample size. However, comparing performance among the presented models via point estimates alone is not valid. In all simulation studies, we apply the mean square error (MSE) to evaluate model performance as we described above. It is well known that MSE of an estimator measures both variability of this estimator and its bias from the true value. That is, the MSE is equal to the sum of the

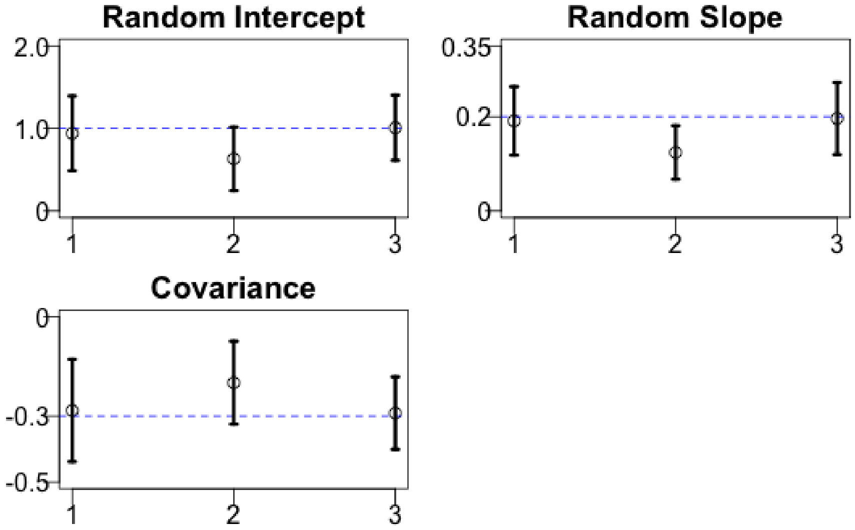


Figure 5.4: Point estimates and confidence interval (credible interval for Bayesian estimates) for random effects from simulated repeated-measure model with lower missing probability. The study sample size is 80. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

variance and squared bias of the estimator. Hence, the MSE assesses the quality of an estimator. This measurement is calculated and compared for each parameter across all the models of interest, and one can conclude from the comparison that estimators from CLFM have the best qualities for estimating on both fixed and random effects.

In the second simulation study we compare CLFM with two other models when data contains a large proportion of missingness. Missing values in this study are generated by the Diggle-Kenward selection model, and the mechanism of missingness is non-ignorable. Two sample sizes are considered in this case, 80 and 300. For each scenario, we simulated 500 replicates, as we specified at the beginning of the simulation

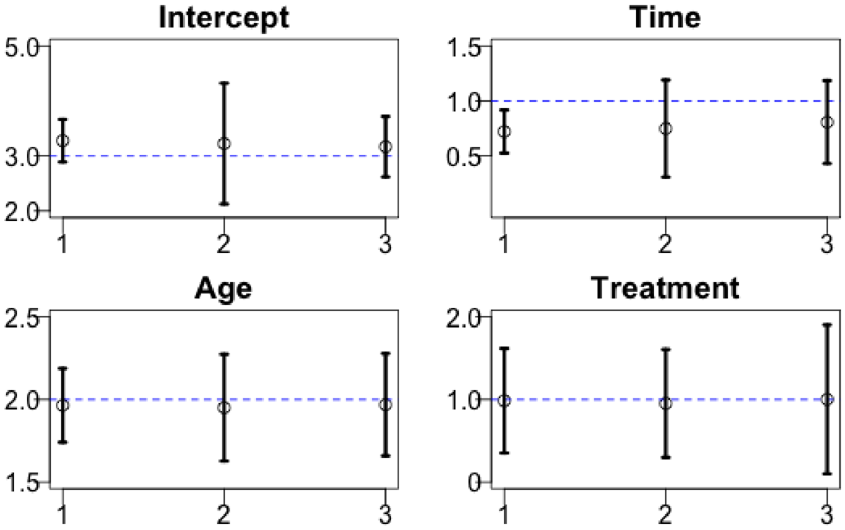


Figure 5.5: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with higher missing probability. The study sample size is 80. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

study. On average, the missing proportion exceeds 50 percent in the data. For most replicates, this proportion achieves 70 percent. Figures 5.5 and 5.6 summarize point estimates and standard errors for both fixed and random effects. As expected, CLFM produced the best results for both cases, in terms of MSE. More specifically, one can observe that the ignorable likelihood approach tends to underestimate fixed intercept and slope in the model; furthermore 95 percent confidence intervals obtained from this approach do not cover the true values.

In Figure 5.6, true values of variance components in random effects, including $\sigma^2_{b_0}$, $\sigma^2_{b_1}$ and $\sigma_{b_0 b_1}$, are labeled as blue dotted lines, and red lines represent the non-significant level. Based on these plots, the variance components are indicated to be

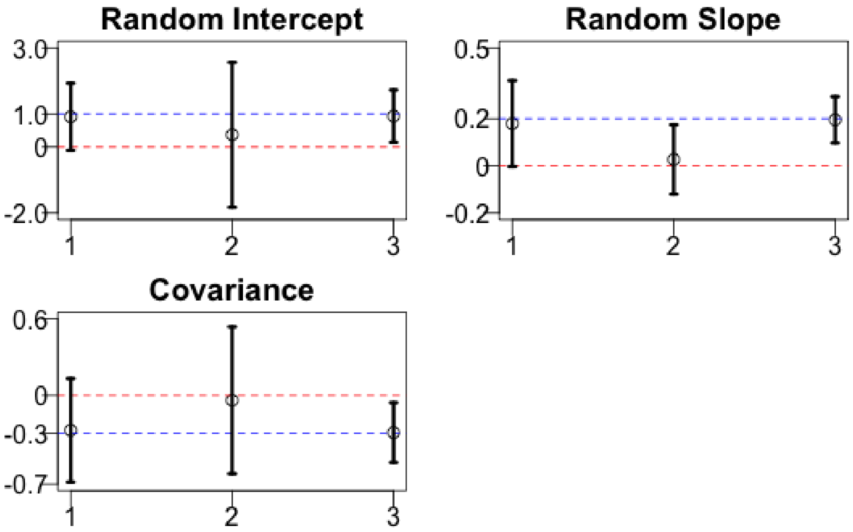


Figure 5.6: Point estimates and confidence interval (credible interval for Bayesian estimates) for random effects from simulated repeated-measure model with higher missing probability. The study sample size is 80. The true values are indicated by the blue dashed lines, the non-significant level is indicated by the red dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

non-significant from the ignorable likelihood approach and Roy's model, but CLFM shows the correct result. In summary, CLFM can correct bias and generate efficient estimators when missing values are not ignorable in a study that contains lots of missingness.

The third simulation is motivated from the clinical studies where patients in different treatment groups tend to have different missing trends. For example, patients who are randomized to control group tend to miss their evaluations with a higher probability, compared with those who are in treatment group. This is due to the fact that patients can not receive treatment benefit in this group. Furthermore, patients are more likely to have non-ignorable missing data in the control group since due to lack of efficacy. In this simulation study, we focus on different missing probabilities across treatment groups and assume missingness in both groups are due to non-ignorable missing data mechanism. The simulation model is given as follows:

$$Y_{ij} = 3.00 - 1.00(Time) + 2.00(Age) - 0.50(Treatment \times Time) + b_{0ij} + b_{1ij}(Time) + \epsilon_{ij}$$

where age is a standardized continuous variable with mean 0 and variance 1 at baseline, subjects are randomly assigned into two treatment groups with a $1:1$ ratio. $b_{0ij} \sim N(0,1)$, $b_{1ij} \sim N(0,0.2)$, $cov(b_{0ij}, b_{1ij}) = -0.3$ and $\epsilon_{ij} \sim N(0,0.5)$. Because patients are randomized to treatment and control, the mean response at baseline is assumed to be the same in the two groups. That is, we don't include a single treatment term in the model. On average, the missing proportion from the control group is 40 percent, and 20 percent for treatment group.

The regression parameters (fixed effects) and variances of subject random effects are given in Figures 5.7 and 5.8. The corresponding asymptotic 95 percent confidence intervals are also presented in the table. From Figures 5.7, we can observe that the ignorable model that excludes missing data from analysis tends to underestimate the average change rate in response. Roy's model performs even worse than the conventional approach on estimating parameters in the model. Specifically, Roy's model underestimates intercept and slope parameters, and overestimates other parameters in the fixed effects, including patients' age, as well as interaction between treatment group and time. Roy's model also generates large standard errors on point estimates, compared with the ignorable model and proposed CLFM. This is due to singularity of the Hessian matrix. Figures 5.8 corresponds to estimation of variance components for random effects. Based on these figures, we find that Roy's model has a tendency to underestimate variances for both random intercept and random slope terms, and

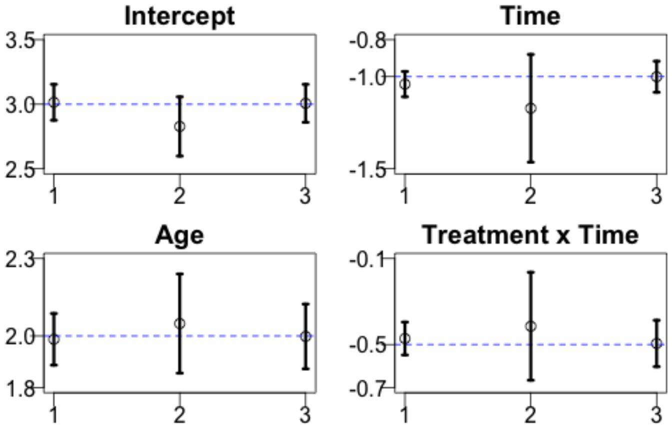


Figure 5.7: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with various missing probabilities in different groups. The study sample size is 300. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

overestimate the covariance. In addition, Roy's model presents larger standard errors for each point estimate as well.

In the above three simulation studies, the estimation difficulties for Roy's model were observed. This was due to small sample size in the simulations. Typically, a large sample size is required to estimate parameters in Roy's model. The last simulation study was designed by generating large enough sample size such that Roy's model can be fitted for all replicates. In this study, Roy's model was able to be fitted with sample size 1000. Each replicate was simulated from the linear mixed model which was the same as in the first two studies. In this simulation, only 20 percent or less of subjects had complete observations across all 6 time points, and the average missing proportion was more than 70 percent. Model evaluations and

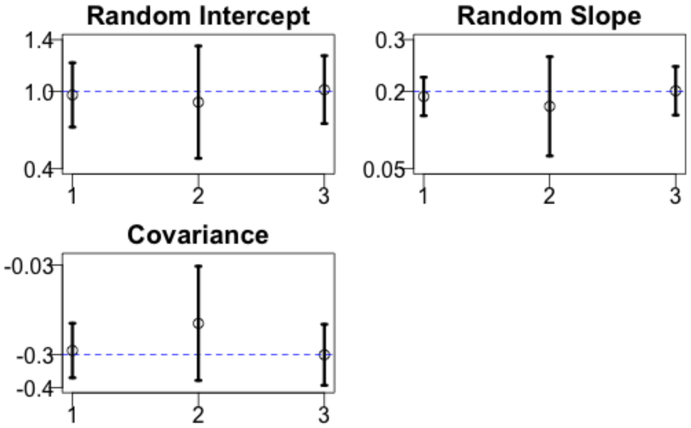


Figure 5.8: Point estimates and confidence interval (credible interval for Bayesian estimates) for random effects from simulated repeated-measure model with various missing probabilities in different groups. The study sample size is 300. The true values are indicated by the blue dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

comparisons followed the same procedures as before. Figure 5.9 and 5.10 describe point estimates for regression parameters, variance components in the linear mixed model, with corresponding 95 percent confidence intervals.

Based on the information from Figure 5.9 and 5.10, the proposed CLFM model produced the most accurate estimates on regression parameters, variance components, especially in estimating intercept and slope effects. Roy's model in this simulation can be fitted on all replicates and had a better performance, compared with the conventional analysis (ignorable model) that excludes missing values. Results from ignorable model suggested that ignoring missing data from analysis led to biased estimates on regression parameters. In specific, ignorable model overestimated intercept term in the linear mixed model, the change slope term, and covariates terms were tented to

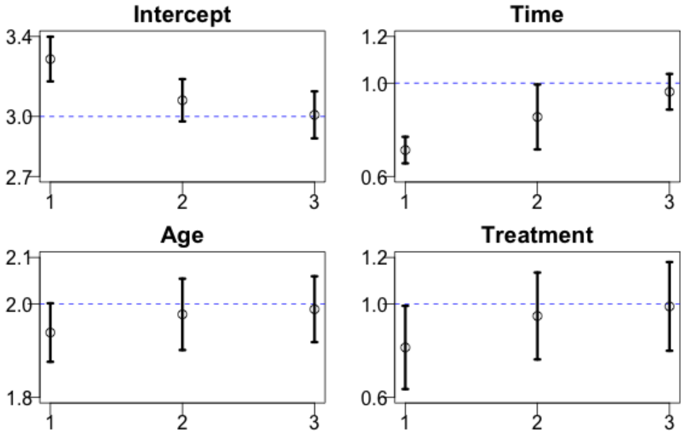


Figure 5.9: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with various missing probabilities in different groups. The study sample size is 1000. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

be underestimated. Subject variability in terms of repeated measures changes across study period was overestimated.

5.2 An Application on Peabody Picture Vocabulary Test Data

5.2.1 Description of Data

In this section, we present an application on an observational study that is from the National Longitudinal Survey of Youth (NLSY79 Child Survey). The NLSY79 Child and Young Adult cohort is a longitudinal project that follows the biological children of the women in the NLSY79. As of 2010, more than 10,000 children have been interviewed in at least one survey round. In 1986, a separate survey of all children born to NLSY79 female respondents began, greatly expanding the breadth

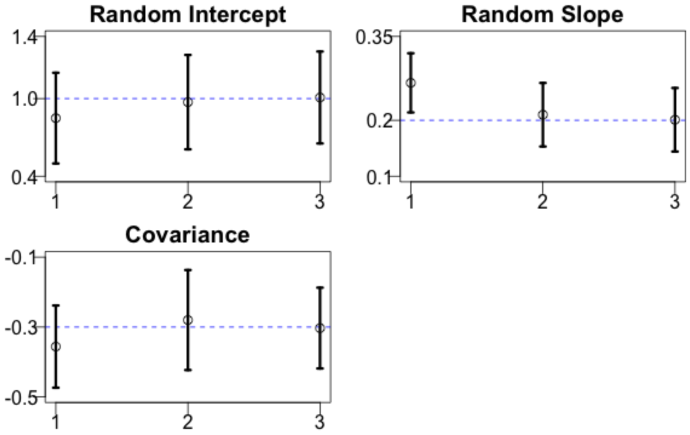
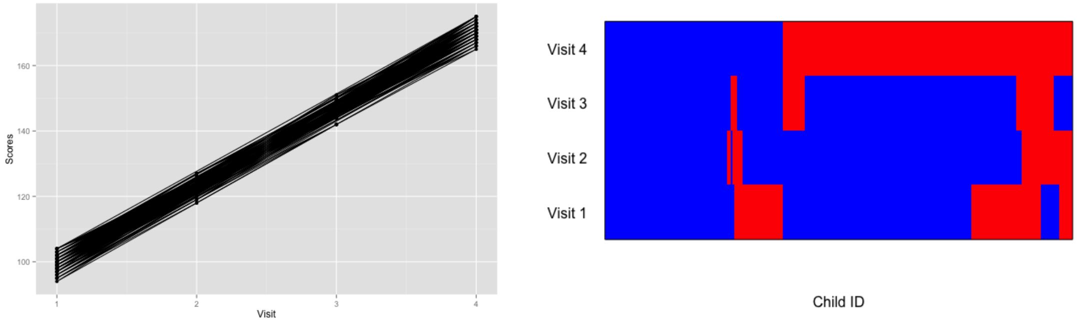


Figure 5.10: Point estimates and confidence interval (credible interval for Bayesian estimates) for random effects from simulated repeated-measure model with various missing probabilities in different groups. The study sample size is 1000. The true values are indicated by the blue dashed lines. 1, ignorable model; 2, Roy's model; 3, CLFM model from Bayesian approach.

of child-specific information collected. The children of NLSY79 female respondents are assessed and interviewed every two years. These assessments measure cognitive ability, temperament, motor and social development, behavior problems, and self-competence of the children, as well as the quality of their home environment. One of important assessing tools is referred to as the Peabody Picture Vocabulary Test (PPVT) and its revised version PPVT-R (Weber, 2007). PPTV measures an individual's receptive vocabulary for Standard American English, as well as verbal ability or scholastic aptitude. In this example, we focus on children with age ranging from 72 months to 83 months. In this test, a child listens to a word uttered by interviewer and then selects one of four pictures that best describes the word's meaning. The PPVT-R consists of 175 stimulus words and 175 corresponding image plates. Each image plate contains 4 black-and-white drawings, one of which best represents the meaning of the corresponding stimulus word. There are also 5 training words and image plates. For those five training items, they are administrated at the beginning of the PPVT assessment in order to familiarize children with the task. The first item, or starting point, is determined based on the child's PPVT age. Starting at an age-specific level of difficulty is intended to reduce the number of items that are too easy or too difficult, in order to minimize boredom or frustration for children.

PPVT begins with the starting point and proceeds forward until the child makes an incorrect response. If the child has made 8 or more correct responses before the first error, a "basal" is established. The basal is defined as the last item in the highest series of 8 consecutive correct answers. Once the basal is established, testing proceeds forwards, until the child makes six errors in eight consecutive items. If, however, the child gives an incorrect response before 8 consecutive correct answers have been made, testing proceeds backwards, beginning at the item just before the starting point, until 8 consecutive correct responses have been made. If a child does not make eight consecutive responses even after administering all of the items, he or she is given a basal of one. If a child has more than one series of 8 consecutive correct answers, the highest basal is used to compute the raw score. A "ceiling" is also established when a child incorrectly identifies six of eight consecutive items. The ceiling is defined as the last item in the lowest series of eight consecutive items with six incorrect responses. If more than one ceiling is identified, the lowest ceiling is used to compute the raw score. The assessment is complete once both a basal and a ceiling have been established.

A child's raw score is the number of correct answers below the ceiling. One needs to note that all answers below the highest basal are counted as correct, even if the child answered some of these items incorrectly. The raw score can be calculated by



(a) Growth Curve Plot for Complete Cases

(b) Missing Pattern Plots

Figure 5.11: PIAT studies on children: trajectory plot for complete cases (left; scores are in original scale); missing pattern plots (right; red presents a missing value, blue correspondes to an observed value)

subtracting the number of errors between the highest basal and lowest ceiling from the item number of the lowest ceiling. The total score for this test ranged in value from 0 to 84, but in our study, this score will be rescaled by dividing by 10. (A complete description about this test can be found on the NLSY webpage.) In total, this example includes 323 children. At the first measurement in 1992, the children were about 6-7 years of age. The same children were then repeatedly measured at 2-year intervals for three additional measurement occasions (1994, 1996 and 1998). One empirical research question of interest was stated as ”Is there systematic change in verbal ability and scholastic aptitude, and individual difference in this change over time (8 years)?” The equivalent question to be answered in statistical words, ”Is there evidence to suggest that the slope term across time and random effects at subject level are statistically significant?”.

Figure 5.11 provides some basics of this study; Figure 5.11a presents trajectory changes of children’s PPVT score who have completed all four assessments. In this plot, we set the first assessment as a baseline value, i.e. treat the year of 1992 as the first visit time point. The following three assessments are taken at visit 2, 3 and 4, respectively. Note that testing scores here are in the original scale. From this plot, one can observe that testing scores showed a gradual increase. Furthermore, missing data in this study draws intensive attention from researchers and investigators. The data set is missing many observations from different subjects, and Figure 5.11b plots missing patterns in this study. In this plot, we use a red color to denote missing values and blue to represent observed. Only 84 children had PPVT scores recorded for all 4 assessment points; there were 13 children who missed all the four assessments, and they were excluded from the analysis; the missing proportion in this study approaches 30 percent. It is possible that the visiting process itself was not ignorable. For instance, children might skip a test when they feel this test is too difficult or too easy for them. Our approach is to implement models (4.2-4.4) to account for the missing visit process, and to compare estimation results with several other models which make different assumptions of the missing mechanism.

5.2.2 Model Results for Ignorable Model, Roy's model, and CLFM

Denote by $\mathbf{Y}_i$ the observed PPVT score vector for subject i. Let $\mathbf{R}_i$ be the 4 vector of observed-data indicators for subject i (including the baseline week). In general, the jth element of $\mathbf{R}_i$ is equal to 1 if PPVT score is missing for subject i in the jth visit, and equal to 0 if it is observed. The primary covariate of interest is visit time points, as we presented earlier. From Figure 5.11a one can see that a linear trend is enough to capture the changes of PPTV scores. Hence the first order term of time will be of interest and included in the model. Moreover, from this plot one can also observe that children at the baseline assessment had variate starting PPTV scores, and as tests went on, they presented different change rates on test scores. This evidence suggested a linear mixed-effect model will be a good fit. We first fitted a linear mixed-effects model ignoring the missing process. We assumed $\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon$, where the design matrix $\mathbf{X}_i$ included an intercept, and assessment time. The random effects design matrix included an intercept and time (i.e. random intercept and slope model). The variance-covariance matrix of the random effects included three parameters: $\sigma^2_{b_0}$ (variance of random intercept), $\sigma_{b_0b_1}$ (covariance) and $\sigma^2_{b_1}$ (variance of random slope). The error term was assumed be normally distributed with mean 0 and variance σ^2.

The above analysis was based on the assumption of an ignorable visit process. We then assumed the missing process cannot be ignored and fitted both Roy's model and CLFM described in an earlier section. In Roy's model, we tried two latent classes and three latent classes. Based on the information criteria and discussion on model selection for latent class models in the previous chapter, a two latent-class model is preferred. As we discussed earlier, Roy's model attempts to categorizing missing patterns into two clusters, and a linear mixed model was fitted for each cluster. The marginal inference was primarily of interest, by averaging estimates across both clusters. There were 20 parameters in total to be estimated. The most significant issue from Roy's model is identifiability, as we reviewed in Chapter 3. In order to avoid this issue and also singularity of the information matrix caused by this issue, one multinomial logit parameter was fixed.

In the following, we applied a CLFM to the PPVT study. We modeled missing probabilities using model 4.5. Two factors were included in this model to describe key features of the missing process: the time location of each observation process, i.e. there are a set of 'location' parameters τ_j, $(j = 1, 2, 3, 4)$ to describe the probability of missing at each time location (or we referred as assessment points); and a random variation term u_i for child i, $(i = 1, 2, \ldots)$ to represent individual tendency in this missing process. For example, a child i with a higher value in u_i may have a higher probability of missing an assessment. This random variation, in another perspective, might contribute to model 4.2 through another random effect term $\mathbf{b}_i$. For instance, it was possible that a child who was more likely to miss the second test tended to have a lower baseline testing score. Or a child may be likely to miss a test in the middle due to his/her slow improvement or fast improvement on the performance at previous tests. By these examples, one can see this random variation term u_i tended to have influences on subject-level random effects in the primary model of interest (i.e. model 4.2). Hence, we adopted a link model 4.3 to describe this relationship. Finally, the observed response data were modeled, conditional on the continuous latent factor $\mathbf{u}$ and subject-level random effects $\mathbf{b}$, using model 4.2. The design matrices X_i and Z_i were the same as used in the mixed-effects model described above that ignored the observation process.

Parameters from models (4.2-4.4) were obtained via both approaches: maximum likelihood estimates from Monte-Carlo Expectation and Maximization (MCEM) method and full Bayesian approach. The MCEM approach was implemented in R, and full Bayesian estimates were solved in WinBUGS. MCEM tended to take longer time to achieve good stable estimates due to extra variation that is introduced in each iteration. The computation time was more than two hours to get a reasonable convergence on likelihood values. This method was implemented on a Macintosh machine with Processor 2.8GHz, Intel core i7 and RAM 4G.

By comparison, we also implemented the full Bayesian approach. In this approach, we started with 2 initialized chains for MCMC method, specifying a prior for each parameter. With a multivariate normal distribution for the complete PPVT data, it is customary to apply improper non-informative or mild informative priors to the mean μ and variance-covariance structure Σ. In this example, we employed a mild informative ridge prior distribution. We suppose that, given covariance structure Σ, mean μ is conditionally multivariate normal, and the variance Σ is inverted-Wishart,

$$\mu|\Sigma \sim N(\mu_0, \lambda^{-1}\Sigma),$$

and

$$\Sigma \sim W^{-1}(m, \Lambda),$$

where $\lambda > 0$, $\Lambda > 0$ and m are user-specified hyperparameters. That is, we specified normal priors with large variances for all regression coefficients in models 4.2-4.4; for a high dimensional variance-covariance structure, an inverted-Wishart prior was employed; an inverted-gamma prior was adopted for a one dimensional precision parameter. To obtain estimates, we let the program have $10,000$ burn-in iterations and used another $20,000$ iterations to conclude the posterior mean for each parameter. The thinning size was set to be 10, in order to reduce correlation between two consecutive sampling points. Figure 5.12 describes trace plots for the fixed effects estimates in linear mixed model 4.2 over 20,000 iterations. One can observe that under the specified priors, two chains were mixed well and both estimates approach stationarity within the allowed iterations.

5.2.3 Results

Under different settings that we described above, we fit three different models that include four scenarios. We employed two different approaches to fit a CLFM. The results from the assumption of ignorable observation process are given in Table 5.2 under 'MAR' column. PPVT score tends to be larger, on average, for children with higher baseline testing score. There is a significant linear time effect, i.e. this suggested that PPVT score will increase as children grew up. Results also indicated that children showed baseline variability in testing scores that is significant.

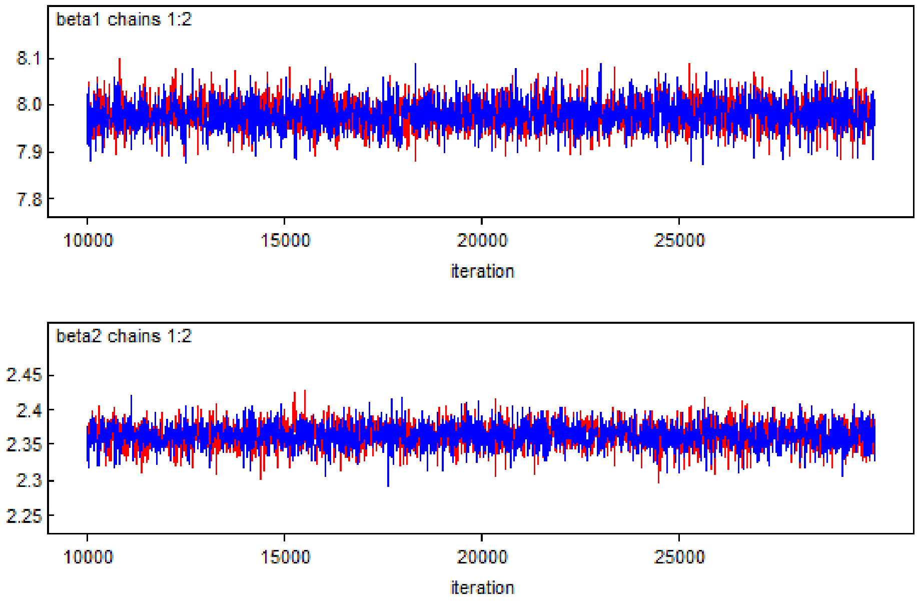


Figure 5.12: Trace plots of children's baseline average test score and average increase on the PPVT score

The point estimates and corresponding standard errors from Roy's model, which incorporates non-ignorable missing values, are given in Table 5.2 under the 'Roy' column. Roy's model gave similar results as the ignorable model (MAR case), except in estimation for variance $\sigma^2_{b_1}$ of random slope and covariance $\sigma_{b_0 b_1}$ between random intercept and random slope. We will discuss those together with estimation from CLFM.

The estimated regression coefficients, variance components and their standard errors from two approaches to fitting a CLFM are displayed in the 'MCEM' and 'Bayesian' columns of Table 5.2. The estimated coefficients (intercept and time effects) and standard errors are very similar to those from the models either ignoring the observation process or Roy's approach. The estimated variance components for random error terms are also close. This suggest that, if one were specifically interested in inference about the marginal covariate effects, then incorporating for the missing process might not have been necessary. Further, all approaches indicated that children had baseline differences in testing scores, and this variability cannot be ignored. The improvement in testing score was also supported by all methods. MAR suggested the variance of random slope term should be significant with a p-value 0.002 of a student-t test. However, there was a disagreement between MAR and the approaches accounting for non-ignorable missing process and this disagreement provided additional information. All approaches that account for non-ignorable missing process suggested that children had substantial differences improvement rates in PPVT evaluation system, depending on their baseline abilities in testing. By comparison, the MAR approach gave a p-value 0.104 of student-t test on testing statistical significance of covariance between random intercept and random slope terms. That means, when the missing process was included in the analysis, we found that change of a child's performance on testing scores depended on his/her baseline status on the test. Specifically, if a child has lower initial test score in the study, he/she had a larger improvement, and his/her improvement on test scores was significantly larger, comparing with a child who had a high score in the first assessment. This finding can be seen in Figure 5.11a. In this plot, many overlaps among growth curves can be observed across the study; for a child with a lower initial test score, his/her growth curve tended to end with a higher score. For a child with a higher baseline score, he/she showed a stable improvement across the study.

Table 5.1 displays the parameter estimates from the continuous latent factor model (τ_1-τ_4, σ_u^2) which describes the missing process (equivalent to observation process), and from the link model (γ_1, γ_2) that gives the relationship between random variation in missing process and children's variability in the growth curve model. The first two rows are the estimated coefficients from the link model. Both approaches (MECM and Bayesian) suggested that children's variability in the missing process was related to their baseline score. That means a child who was far from average performance at baseline tended to have a large chance to be missing in the study, either because they felt tests were too easy or too difficult for them. However, there is no evidence to support that a child's variability in missing process was related with his/her improvements on tests. The left rows, from 3-7 of Table 5.1 display parameter estimates from CLFM. τ_1-τ_4 are location parameters which describe the missing probability at each assessment. A smaller value of τ_j means a larger chance to be missing. In Table 5.1 one can see τ_4 has the smallest value, which means children were more likely to be missing at the last assessment. Further the first assessment also has a higher probability to be skipped.

In this example, data provide no empirical information to favor one model over the other. The fact that the non-ignorable assumption implies that a child's improvement rate on PPVT depends on his/her baseline test score from the average level, however, is very telling. This result, in our opinion, reflects a real fact that children who are below the average, in terms of test score, have a large improvement and tend to show a faster improvement rate on the tests.

Table 5.1: Parameter estimates and estimated standard errors for the missing process model and link model

Variables	MCEM		Bayesian	
	Estimate	SE	Estimate	SE
γ_1	0.147	0.030	0.149	0.031
γ_2	-0.009	0.008	-0.010	0.014
σ_u^2	1.414	0.150	1.413	0.152
τ_1	1.276	0.172	1.275	0.174
τ_2	2.433	0.218	2.424	0.219
τ_3	2.421	0.215	2.420	0.217
τ_4	-0.667	0.155	-0.668	0.158

Table 5.2: Parameter estimates and estimated standard errors for the regression coefficients and variance components from MAR model, Roy's model with two classes, and from CLFM

Variables	MAR		Roy		MCEM		Bayesian	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	7.958	0.024	7.613	0.028	7.980	0.025	7.979	0.032
Time	2.365	0.005	2.366	0.010	2.363	0.015	2.362	0.018
$\sigma^2_{b_0}$	0.146	0.015	0.048	0.009	0.213	0.020	0.215	0.021
$\sigma_{b_0 b_1}$	-0.004	0.002	-0.033	0.003	-0.035	0.008	-0.037	0.009
$\sigma^2_{b_1}$	0.002	0.001	0.062	0.002	0.080	0.007	0.081	0.007
σ^2_{ϵ}	0.013	0.001	0.015	0.003	0.011	0.004	0.011	0.004

5.3 Randomized Study of Dual or Triple Combinations of HIV-1 Reverse Transcriptase Inhibitors

In this section, we will illustrate another application of CLFM by using data from a randomized, double-blind, study of AIDS patients with advanced immune suppression, which is measured as CD4 counts ≤ 50 cells/ mm^3. (Henry and Erice, 1998)

5.3.1 Description of Study

Patients in an AIDS Clinical Trial Group (ACTG) Study 193 A were randomized to dual or triple combinations of HIV-1 reverse transcriptase inhibitors. Specifically, HIV patients were randomized to one of four daily regimens containing 600 mg of zidovudine: zidovudine plus 2.25 mg of zalcitabine; zidovudine plus 400 mg of didanosine; zidovudine alternating monthly with 400 mg didanosine; or zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine (triple therapy). In this study, we focus on the comparison of the first three treatment regimens (dual therapy) with the forth (triple therapy)as described in Fitzmaurice's work. (Fitzmaurice and Laird, 2004)

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to unequal measurements and also CD4 counts have missing data that were caused by skipped visits and dropout. Table 5.3 presents four randomly selected subjects. The number of measurements of CD4 counts during the first 40 weeks of follow-up varied from 1 to 9, with a median of 4, based on the available data. The goal in this study is to compare the dual and triple therapy groups in terms of short-term changes in CD4 counts from baseline to week 40. The responses of interest are based

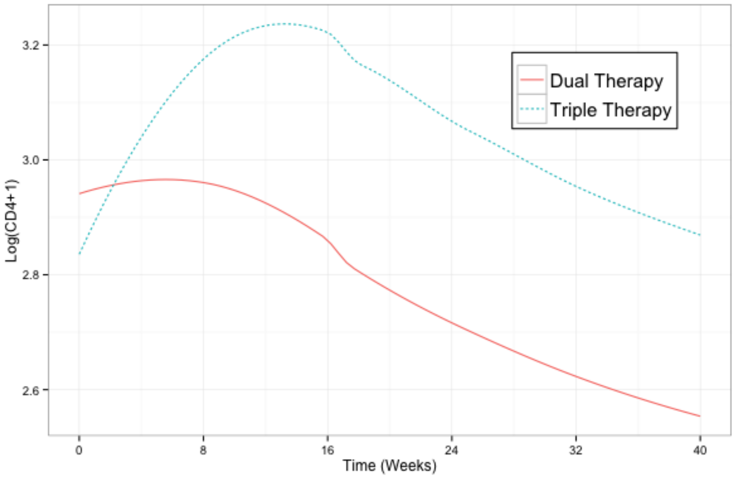


Figure 5.13: Lowess smoothed curves of log(CD4 + 1) against time (in weeks), for subject in the dual and triple therapy groups in ACTG study 193A

on log transformation CD4 counts, log(CD4 counts + 1), available on 1309 patients.

Figure 5.13 describes the trend in the mean response in the dual and triple therapy groups via lowess smoothed curves on observed data. The curves reveal a modest decline in the mean response during the first 16 weeks for the dual therapy group, followed by a steeper decline from week 16 to week 40. By comparison, the mean response increases during the first 16 weeks and declines after for the triple therapy group. The rate of decline from week 16 to week 40 appears to be similar for the two groups. However, one has to notice that there is a substantial amount of missing data in the study, therefore the plot of the mean response over time can be potentially misleading, unless the data are missing completely at random (MCAR). Moreover, based on a small random sample of individuals, we observed that those with drop-out tend to have large CD4 counts. In other words, there is a trend that a patient in the study tended to skip a visit due to a large magnitude of current CD4 count. That is, a patient tends to skip a visit because of no treatment benefits or side effects. When data are missing due to this reason, a plot of the mean response over time can be deceptive. Figure 5.14 describes observed responses at different visit points in each group. Almost all patients from both groups are treated at baseline and their CD4 count data are collected. There are two sharp decrease in response rate, one is from week 0 to week 8 and the other is from week 32 to week 40. Approaching to the end of the study, most patients are dropping out from study, and response rates at week 40 are close to 20 percent for both treatments. The missing information can substantially influence the analysis and even bias our findings. In the example, we will implement CLFM which assumes missing data are not ignorable, and compare with the conventional model that ignores missingness.

In the following we describe a model for the mean response that enables the rates of change before and after week 16 to differ within and between groups, and this model was also been adopted by Fitzmaurice and Laird (2004) in their work. Specifically, one could assume that each patient has a piecewise linear spline with a knot at week 16. That is, the response trajectory of each patient can be described with an intercept and two slopes–one slope for the changes in response before week 16, another slope for the changes in response after week 16. Further, we assume the average slopes for changes in response before and after week 16 are allowed to vary by group. Because this is a randomized study, the mean response at baseline is assumed to be the same in the two groups, as supported by Figure 5.13. Hence instead of the conventional growth curve model, we applied a special growth curve model to capture changing trends of responses on CD4 counts.

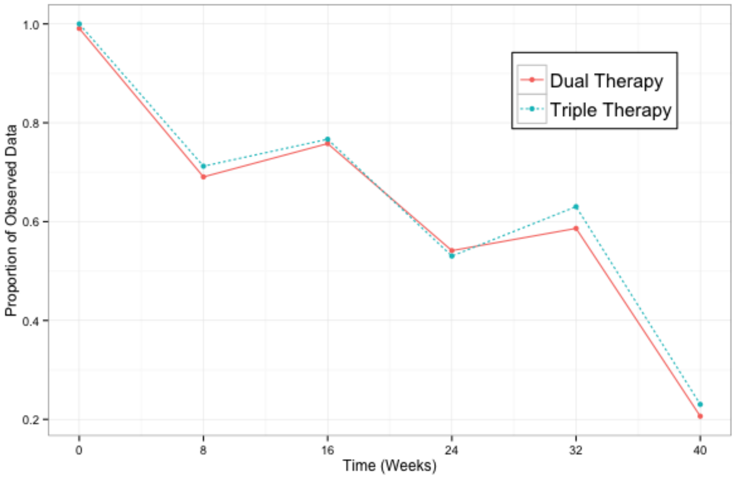


Figure 5.14: Proportions of observed responses in the dual and triple therapy groups in ACTG study 193A

5.3.2 Model Specification

Let t_{ij} denote the time since baseline for the jth measurement on the i-th subject with $t_{ij} = 0$ at baseline, we consider the following linear mixed effects model:

$$
\begin{aligned}
E(Y_{ij}|b_i) =& \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 Group_i \times t_{ij} + \beta_5 Group_i \times (t_{ij} - 16)_+ \\
& + b_{1i} + b_{2i} t_{ij} + b_{3i}(t_{ij} - 16)_+
\end{aligned}
$$

where $Group_i = 1$ if the ith subject is randomized to triple therapy, and $Group_i = 0$ otherwise; $(t_{ij} - 16)_+ = t_{ij} - 16$ if $t_{ij} > 16$ and $(t_{ij} - 16)_+ = 0$ if $t_{ij} \le 16$; b_{1i}, b_{2i} and b_{3i} are random effects in this splined growth curve model. In this model, $(\beta_1 + b_{1i})$ is the intercept for the ith subject and has an interpretation as the true log CD4 count as baseline, i.e. when $t_{ij} = 0$. Similarly, $\beta_2 + b_{2i}$ is the ith subject's slope, or rate of change in log CD4 counts from baseline to week 16, if this patient is randomized to dual therapy; $(\beta_2 + \beta_4 + b_{2i})$ is the ith subject's slope if randomized to triple therapy. Finally, the ith subject's slope from week 16 to week 40 is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$ if randomized to dual therapy and $\{(\beta_2 + \beta_3 + \beta_4 + \beta_5) + (b_{2i} + b_{3i})\}$ if randomized to triple therapy. The model described above will be fitted without incorporating missing data. In order to fit CLFM, one has to specify the model for the missing part. Assume that $\mathbf{R}$ is a missing indicator matrix where its (i, j)th element $r_{ij} = 1$ if Y_{ij} is missing and $r_{ij} = 0$ if it is observed. Within a framework of CLFM, we incorporate information on missing values through modeling the missing information matrix $\mathbf{R}$ with time location parameters, and a continuous latent factor $\mathbf{u}$. Further, there are strong indications which support a application of this model. Based on Figure 5.14 one can see that the response variable tends to be missing over time. In other words, time locations are good indicators for explaining missing data. From Figure 5.14 one might also notice that the two therapies have identical missing proportions which suggests a group effect for therapies is not necessary in modeling $\mathbf{R}$. The continuous latent factor $\mathbf{u}$ is used to describe individuals' variability in missingness, and two regression parameters γ_1 and γ_2 are specified to provide information on random intercept $\mathbf{b}_0$ and slope $\mathbf{b}_1$, in order to correct estimation bias. A third regression parameter was also explored which links $\mathbf{u}$ with $\mathbf{b}_3$, but analysis results showed that this parameter is not significant. Hence we exclude this parameter in the final results. To estimate CLFM, we adopt both approaches: MECM to obtain ML estimates and full Bayesian estimates with specified conjugate priors. Point estimates and corresponding standard errors from a Bayesian perspective are summarized by posterior mean and standard deviation. Roy's model is also implemented by summarized missing patterns from $\mathbf{R}$ into three latent classes. (The number of latent classes for Roy's model is determined by information criteria)

5.3.3 Summary of Analyses under MAR and MNAR

In this study, one research question of interest is treatment effects in the changes in log CD4 counts. The null hypothesis of no treatment group differences can be expressed as $H_0: \beta_4 = \beta_5 = 0$. The ML estimates on fixed effects from three models are given in Table 5.4, including the conventional model with a MAR assumption, Roy's model that handles non-ignorable missing data from pattern-mixture modeling and CLFM. The Bayesian estimates for CLFM are also displayed in Table 5.4. For the likelihood approach with MAR assumptions, a test of $H_0: \beta_4 = \beta_5 = 0$ yields a Wald statistic, $W^2 = 59.12$, with 2 degrees of freedom, and corresponding p-value is less than 0.0001. For the full Bayesian approach, we compute Deviance information criterion (DIC) to compare two models: one assumes no treatment effects by excluding interaction terms between treatment groups and study time; the other assumes treatment effects are significant. DIC for a model with embracing treatment effects is 15792.7, which is less than the one from the model with no groups effects, 18076.5. Based on the criteria, 'the smaller the better', there is evidence to support the fact that treatment group differences in changes in log CD4 counts are significant. The tests from Roy's model and MCEM approach on CLFM also support this group variety, with p-values for both less than 0.0001. Based on the magnitude of the estimate of β_4, and its standard error from all approaches, there is a significant group difference in the rates of change from baseline to week 16. The estimated response curve for two groups are displayed in Figure 5.4. In this figure, dashed lines represent the response curve from CLFM, dotted lines correspond to results from Roy's model, while solid lines are results from the MAR approach; blue color describes dual therapy group, and red one corresponds to triple therapy. In the dual therapy group, there is a significant decrease in the mean of the log CD4 counts from baseline to week 16, based on the ignorable likelihood approach. The estimated change during the first 16 weeks is -0.12, which can be obtained from 16×-0.0073. On the untransformed scale, this corresponds to an approximate 10% decrease in CD4 counts. However, CLFM which assumes missing data are not ignorable suggests that this decrease is not significant, since the 95 percent credible interval for β_2 covers zero ($[-0.01638, 0.006517]$). Further, Roy's model also confirms this finding with the 95 percent confidence interval $[-0.016076, 0.005876]$. By observing missingness from baseline to week 16, subjects with higher log CD4 counts tend to be missing. CLFM involves non-ignorable missing data in the analysis, and the average of log CD4 counts tend to recover to a higher value. Hence, the decrease in the mean of the log CD4 counts from baseline to week 16 is not significant, when non-ignorable missing data are considered. By comparison, in the triple therapy group, there is a significant increase in the mean response. Based on the ignorable approach, the estimated change during the first 16 weeks in the triple therapy group is 0.31, $(16 \times (-0.0073 + 0.0269))$; the estimated slope for the triple therapy group is 0.0196 with a standard error 0.0033. In terms of the untransformed scale, it corresponds to an approximate 35 percent increase in CD4 counts. In CLFM, a similar estimate is obtained: the corresponding estimated change is 0.36. $(16 \times (-0.0047 + 0.0273))$; the estimated slope for the triple therapy group is 0.0226, and it corresponds to an approximate 40 percent increase in CD4 counts.

The loess curves in Figure 5.13 suggest that the rate of decline from week 16 to week 40 is similar for the two groups. The null hypothesis of no treatment group difference in the rates of change in log CD4 counts from week 16 to week 40 can be expressed as $H_0: \beta_4 + \beta_5 = 0$. The estimates of β_4 and β_5 from all approaches appear to support the null hypothesis since they are of similar magnitude but with opposite signs. In the work of Fitzmaurice and Laird (2004), a test of the null hypothesis,

$H_0: \ \beta_4 + \beta_5 = 0$, is given and a Wald statistic is yielded with $W^2 = 0.07$, with 1 degree of freedom. The corresponding p value is greater than 0.75 based on the ignorable likelihood approach. DIC comparison for the Bayesian version of CLFM also suggests that two groups have similar rate of decline from week 16 to week 40. The Wald tests for Roy's model and MCEM version of CLFM further indicate this parallel change profiles after week 16, with both p-values are greater than 0.6.

The estimated variances of the random effects in Table 5.4 indicate that there is substantial individual variability in baseline CD4 counts and the rates of change in CD4 counts. For instance, in the triple therapy group, many patients show increases in CD4 counts during the first 16 weeks, but some patients have declining CD4 counts. Specifically, approximately 95 percent of patients are expected to have changes in log CD4 counts from baseline to week 16 between -0.64 and 1.27. Hence, there are approximately 26 percent of patients who are expected to have decreases in CD4 counts during the first 16 weeks of triple therapy, based on the ignorable likelihood approach; by comparison, a larger variability from patient to patient is indicated by CLFM. 95 percent of patients are expected to have changes in log CD4 counts from baseline to week 16 between -1.15 and 1.87, and correspondingly approximately 30 percent of patients are expected to decrease CD4 counts from CLFM. Substantial components of variability due to measurement error are also suggested from all models.

5.3.4 Distributions on Latent Factor

In this study, we have explored under the assumption of a normal distribution on the proposed latent factor $\mathbf{u}$. The normal distribution is a natural starting point for this CLFM, but it also has limitations. The normal distribution implies non-skewed spread on proposed latent factor which may be too simplistic. In this section, we will extend the distribution of latent factor u to more general distribution. Specifically, we

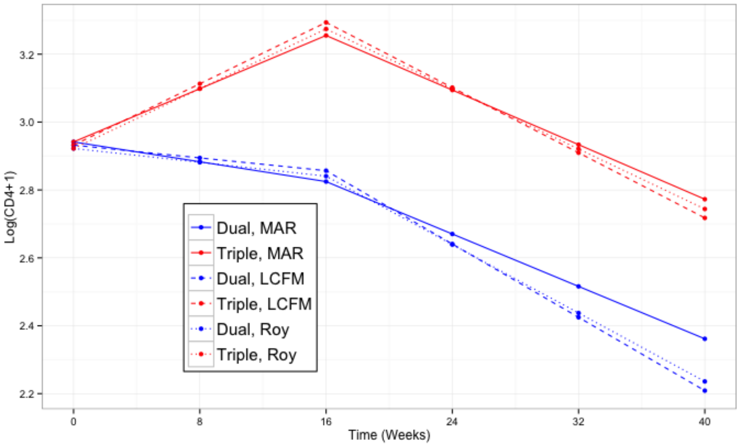


Figure 5.15: Fitted response curve in the dual and triple therapy groups in ACTG study 193A

will give an example of logistic distribution on **b** and compare the estimating results, to demonstrate the flexibility of proposed model, as well as the estimating scheme from Bayesian perspective.

As we described in Chapter 4, a latent factor **u** is proposed to summarize missing patterns and will be used to compensate for the missing information in a repeated-measure model. At the beginning of the investigation, it is natural to choose a normal distribution for **u**, which assumes more information is needed to be filled in the middle of the study. However, some longitudinal studies may experience missing values, which will lead to a heavy tail on the distribution of **u**. In order to fit this senario, a complicate distribution is needed, other than classical normal distribution. Further, the proposed Bayesian estimating scheme allows this extension more straightforward. To present this flexibility on specifying various distribution of the latent factor **u**, we adopted two distribution forms: normal distribution and logistic distribution. In the specification of parameters in logistic distribution, we choose so that the logistic distribution has similar shape with the normal distribution, in order to achieve comparability. Estimation procedure was performed within the Bayesian framework, and the estimation results of parameters including point estimates and standard errors in the linear mixed model are given in Table 5.5. The routine experienced longer time to obtain stable mixed Markov chains when a logistic distribution was used. In detail, we extended the burn-in iterations to $20,000$ and started another $30,000$ iterations to obtain posterior estimates, with thinning size 10. From Table 5.5 one can observe that two distributions produced identical results, due to specified similar distribution shapes. Furthermore, one advantage should be mentioned is that the proposed Bayesian estimating scheme is more flexible in extending distribution of repeated-measures, other than stating different distribution shapes on the latent factor $\mathbf{u}$. In this study, missing data are potentially not ignorable with analyzing a random selected subsample, especially for the first 16 weeks. To evaluate effectiveness of treatment therapies, we compared three approaches, including the ignorable model which assumes missing data are MAR, Roy's model that handles non-ignorable missing data from pattern-mixture perspective, and CLFM with NMAR assumption. Controversial results on change rates of log CD4 counts at dual therapy group during first 16 weeks were obtained, that is, ignorable suggested there is a significant decrease in log CD4 counts, whereas both Roy's model and CLFM indicated this decrease is not substantial. This disagreement is due to those potential non-ignorable missing values. However, all approaches supported that triple therapy has similar change rate on log CD4 counts from week 16 to week 40, compare with dual therapy group. Further, with incorporating missing values, efficacy for both therapy groups is shown to be more substantial from CLFM, which can be seen from the log CD4 counts at week 40. Compared with Roy's model, the proposed CLFM is more flexible in extending the model with a more general distribution.

Table 5.3: Data example on log CD4 counts for four randomly selected subjects from ACTG study 193A

Subject ID	Group	Time	log(CD4 + 1)
56	0	0.0	1.7047
56	0	8.1	1.7981
56	0	16.1	0.6932
56	0	25.4	1.0986
56	0	33.4	0.6932
56	0	39.1	0.6932
529	1	0.0	4.0073
529	1	7.4	3.7136
529	1	16.4	3.5264
529	1	25.4	3.1781
529	1	33.6	3.6636
763	0	0.0	2.8622
763	0	8.0	1.9459
763	0	14.9	1.6094
763	0	21.9	1.7917
777	1	0.0	2.3979
777	1	8.4	1.7918
777	1	10.4	3.0445
777	1	25.3	3.0445

Table 5.4: Estimated regression coefficients (fixed effects) and variance components (random effects) for the log CD4 counts from a MAR model, Roy's model and CLFM in both approaches

Variables	MAR		Roy		MCEM		Bayesian	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.9415	0.0256	2.9223	0.0374	2.9300	0.0250	2.9320	0.0262
t_{ij}	-0.0073	0.0020	-0.0051	0.0056	-0.0040	0.0052	-0.0047	0.0058
$(t_{ij}-16)_+$	-0.0120	0.0032	-0.0201	0.0052	-0.0221	0.0090	-0.0223	0.0092
$Group_i \times t_{ij}$	0.0269	0.0039	0.0271	0.0062	0.0272	0.0105	0.0273	0.0109
$Group_i \times (t_{ij}-16)_+$	-0.0277	0.0062	-0.0240	0.0102	-0.0243	0.0169	-0.0243	0.0177
$Var(b_{1i}) = g_{11}$	585.742	34.754	364.000	49.000	630.050	32.430	640.600	34.7300
$Var(b_{2i}) = g_{22}$	0.923	0.160	1.000	0.500	2.3190	0.9990	2.3230	1.0050
$Var(b_{3i}) = g_{33}$	1.240	0.395	2.000	1.013	37.640	1.9503	38.8600	2.0840
$Cov(b_{1i}, b_{2i}) = g_{12}$	7.254	1.805	-7.106	3.001	-8.6240	3.0500	-8.5240	4.0760
$Cov(b_{1i}, b_{3i}) = g_{13}$	-12.348	2.730	-1.500	3.120	-2.5150	5.3000	-2.5220	6.5000
$Cov(b_{2i}, b_{3i}) = g_{23}$	-0.919	0.236	-6.405	0.892	-7.0130	0.9980	-7.1530	1.0070
$Var(e_i) = \sigma^2$	306.163	10.074	412.000	36.000	500.6300	6.7390	515.3000	9.3570

Table 5.5: Estimated regression coefficients (fixed effects) and variance components (random effects) for the log CD4 counts from CLFM with normal distribution and logistic distribution

Variables	Normal Distribution		Logistic Distribution	
	Estimate	SE	Estimate	SE
Intercept	2.9320	0.0262	2.9310	0.0258
t_{ij}	-0.0047	0.0058	-0.0048	0.0058
$(t_{ij} - 16)_+$	-0.0223	0.0092	-0.0221	0.0090
$Group_i \times t_{ij}$	0.0273	0.0109	0.0273	0.0111
$Group_i \times (t_{ij} - 16)_+$	-0.0243	0.0177	-0.0241	0.0173
$Var(b_{1i}) = g_{11}$	640.600	34.7300	641.300	35.720
$Var(b_{2i}) = g_{22}$	2.3230	1.0050	2.3210	1.0120
$Var(b_{3i}) = g_{33}$	38.8600	2.0840	38.7900	2.0580
$Cov(b_{1i}, b_{2i}) = g_{12}$	-8.5240	4.0760	-8.5760	4.0790
$Cov(b_{1i}, b_{3i}) = g_{13}$	-2.5220	6.5000	-2.5850	6.4420
$Cov(b_{2i}, b_{3i}) = g_{23}$	-7.1530	1.0070	-7.0980	1.0090
$Var(e_i) = \sigma^2$	515.3000	9.3570	515.2000	9.3880

5.4 Growth of Language and Early Literacy Skills in Preschoolers with Developmental Speech and Language Impairment

In this section, we will apply the proposed model CLFM on the study of growth of language and early literacy skills in preschoolers who have developmental speech and language impairment. A robustness analysis, including comparison of CLFM with the conventional model that ignores missing data will also be conducted, under two different assumptions of missing data mechanism: missing at random (ignorable missing data) and not missing at random (non-ignorable missing data).

5.4.1 Description of Study

U.S. Department of Education data for the Individuals with Disabilities Education Act (IDEA) demonstrate that 13% of four-year olds and five-year olds are receiving special education services in preschool and that 82% of these children indicate developmental speech and language impairment (DSLI) as a primary diagnosis. Young children with DSLI often fail to develop crucial pre-literacy skills, which will place those children at high risk for later literacy difficulties and reading failure. Further researchers also find that preschoolers with DSLI demonstrate persistently depressed academic achievement, greater grade retention, and lower rates of post-secondary school attendance than their normally-developing peers. Due to these potential risks, it is urgent and necessary to address children's oral language and early literacy skills during the preschool years to increase their ability to benefit from reading and writing instruction in elementary school. Small intervention studies have been conducted on children with DSLI, including targeting code-related early literacy skills, inferential language skills and oral language or curriculum supplements based on shared reading. Researchers also perform studies on evaluating the effectiveness of an early childhood curriculum with regard to improving early literacy and oral language skills for young children with DSLI.

In a recent study, researchers are interested in examining the efficacy of "Teaching Early Literacy and Language" (TELL) curriculum in promoting the early literacy and oral language growth trajectories of preschoolers with DSLI. The TELL curriculum (Wilcox and Gray, 2011) includes a series of instructions, scripted teaching activities, materials for implementation of oral language and early literacy activities, and professional development for teachers. It was designed to target both code-focused (phonological awareness, alphabet knowledge, print concepts, and writing) and oral language skills fields (vocabulary and complex language) because these skill sets have been documented as predictors of children's literacy success. In an earlier small randomized controlled trial, the TELL curriculum has shown positive results for promoting gains in early literacy and oral language skills in preschool children with DSLI; with comparison, recent research expands existing interests by examining growth trajectories of early literacy and oral language skills for children with DSLI and comparing those trajectories of children who received the TELL curriculum with those who were randomly assigned to control classes.

In this study, we are interested in one specific item from TELL curriculum, Curriculum Based Measurement (CBM) Themed Vocabulary Total Correct Expressive (VOCE). Data in the study were obtained from 130 preschool children with the average age 53.9 months, including preschool children's demographic variables, as well as parental education level, family income. These children were randomly assigned to offer the TELL curriculum or accept those with business as usual (BAU). All children met inclusion criteria which include a test on hearing within normal limits (WNL), a diagnosis of DSLI as the only disability, a K-ABC score WNL, and ability to produce simple sentences (S+V+O). The efficacy variable, VOCE test score was scheduled to be collected at baseline, as well as six follow-up time points (week 1, 2, 3.3, 6, 7.3, 8.8). As a risk factor, mother's education level score, was also collected and treated as a continuous baseline covariate. The average VOCE standard scores in both TELL Curriculum and control groups, including corresponding standard deviation at each follow-up time visit are given in Table 5.6. On average, children who received TELL curriculum have higher VOCE scores since the first follow-up visit, compared with those accepted BAU (control group). In the next section, a linear mixed model is describe and used to capture the change profile for both groups in terms of VOCE changes from the first follow-up time to the end of study, and robust analysis on the assumptions of different missing data mechanism will be introduced, through two missing data generating models. Meanwhile, the sample size is too small to apply Roy's model.

Table 5.6: VOCE score by group (TELL curriculum vs. control): mean, standard deviations at each scheduled visit

Variables	TELL ($n = 87$)	Control ($n = 43$)
	Mean (SD)	Mean (SD)
VOCE (T1) Standard Scores	2.057 (1.748)	0.326 (0.566)
VOCE (T2) Standard Scores	2.057 (1.450)	0.651 (0.783)
VOCE (T3) Standard Scores	2.448 (1.796)	0.326 (0.606)
VOCE (T4) Standard Scores	3.851 (1.632)	1.698 (1.081)
VOCE (T5) Standard Scores	3.828 (1.819)	1.465 (1.162)
VOCE (T6) Standard Scores	4.333 (1.809)	1.372 (1.310)

5.4.2 Model Specification for Complete Data and Missing Data Generation

In previous studies we conducted, including simulation studies, PPVT study, and ACTG study, the non-ignorable missingness is assumed or strongly suggested by application data. The performance of CLFM have been researched by comparing with two other models, (conventional analysis that does not include missing values and Roy's model that handles non-ignorable missing data) under this assumption. In this section, we want to conduct a robust analysis on CLFM by using this real application data. With the complete information in the TELL study, we will adopt two alternative models to generate missing data with different assumptions: ignorable missing data and non-ignorable missingness. CLFM's performance will be evaluated under both scenarios: correctly specified missing data assumption (non-ignorable) and mis-specified missing data assumptions (ignorable). Figure 5.16 plots loess curves fittings of VOCE changes from first visit to the end of study on both groups. Based on this plot one can obtain the information that Preschool children who were in TELL curriculum group had a higher VOCE scores, compared with those children in the control group. Further, children in TELL group showed a significant benefit from proposed TELL and demonstrated a steady increase in the efficacy variable from baseline to the end of the study; on the contrary, children in control group presented a flat change profile at the beginning of the study period, however, this ability increased from week 3.5 to 6, and tended to be flat or decreasing after week 6. To capture VOCE score changes that differ from within and between groups, we use a linear mixed model to model the efficacy variable. In order to fit changes in VOCE score over time, higher order terms in time will be included in the model; a treatment group term and an interaction term between treatment group and time will be introduced to present both baseline and changing rate differences in VOCE

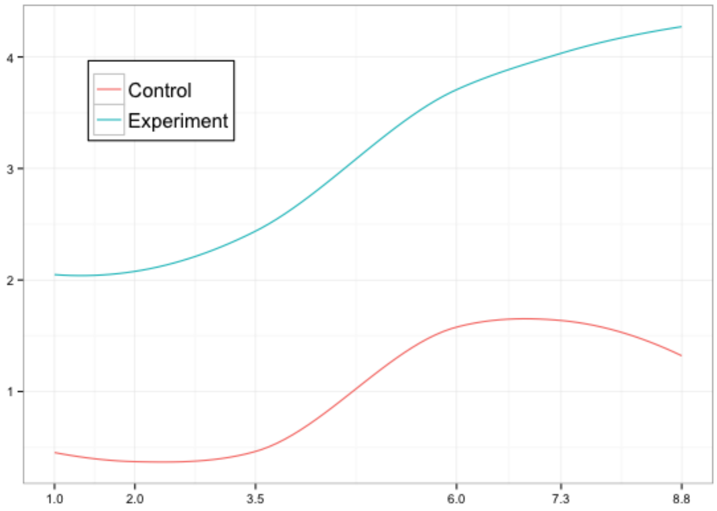


Figure 5.16: Lowess smoothed curves of VOCE standard scores against time (in weeks), for subject in the TELL curriculum and control groups

score. A risk factor, mother education level, will also be investigated.

Let t_{ij} be the time since the first visit for the jth visit time on the ith subject, with $t_{ij} = 1$ at first follow-up visit. The following linear mixed effects model is considered:

$$Y_{ij} = \beta_1 + \beta_2\ t_{ij} + \beta_3\ t_{ij}^2 + \beta_4\ t_{ij}^3 + \beta_5\ Group_i + \beta_6\ Group_i \times t_{ij} +$$

$$+ \beta_7\ Matedu + b_{0i} + b_{1i}\ t_{ij} + \epsilon_{ij}$$

where Y_{ij} is the VOCE standard score at jth visit for subject i ($i = 1, 2, \ldots, 130$; $j = 1, 2, \ldots, 6$); $t_i = (1.0,\ 2.0,\ 3.5,\ 6.0,\ 7.3,\ 8.8)$ is a visiting time vector on ith subject. Random effects terms b_{0i} and b_{1i} are used to describe individual variabilities. With the complete data information, the above model can be fitted by conventional likelihood and resulting parameter estimates and corresponding standard errors can be treated as underlying truth in a robust analysis. The following two models are applied to generate missing values with different assumptions:

$$Pr(r_{ij} = 1) = \frac{exp(\alpha + \xi_{1j} Y_{i(j-1)})}{1 + exp(\alpha + \xi_{1j} Y_{i(j-1)})}$$

where r_{ij} represents a binary missing indicator for response Y_{ij}. Here missing values are only allowed to be appear after second visit, i.e. $j = 2, 3, \ldots, 6$; in the above expression, we set $\alpha = -1$, and assume ξ_1 is a 1×5 vector with elements $\xi_1 = (0.2, 0.2, 0.3, 0.4, 0.5)$. With these settings, the above model will produce ignorable missing values and missing probabilities keep change from time to time. Specifically, a response tends to be missing with a higher chance at a later phase of the study. Similarly, the following model is used to generate non-ignorable missing values:

$$Pr(r_{ij} = 1) = \frac{exp(\alpha + \xi_{1j} Y_{ij} + \xi_2 Y_{i(j-1)})}{1 + exp(\alpha + \xi_{1j} Y_{ij} + \xi_2 Y_{i(j-1)})}$$

where $\alpha = -1$, $\xi_1 = (0.2, 0.2, 0.3, 0.4, 0.5)$, and $\xi_2 = 0.1$. Figure 5.17 describes the observed proportions for both simulated missing data.

As mentioned above, complete data will be fitted by using the linear mixed model, and estimated parameters will be used to evaluate the models' performance on fitting TELL data with ignorable or non-ignorable missing values. For TELL data with missing values, two models will be compared: the conventional model that excludes missing data, and CLFM that handles non-ignorable missingness. In CLFM, we incorporate information on missing values through modeling missing indicator matrix $\mathbf{R}$ with time location parameters $\mathbf{d} = (d_2, d_3, \ldots, d_6)$ starting from second visit, (the reason for excluding the first time is due to no missingness at that point) as well as a continuous latent factor $\mathbf{u}$. The continuous latent factor $\mathbf{u}$ is linked with intercept and slope related parameters ($\mathbf{b}_0$ and $\mathbf{b}_1$) in the mixed model. To estimate CLFM, we implement the MCEM algorithm to obtain likelihood estimates and the full Bayesian approach with conjugate priors. The posterior means and corresponding standard deviations are summarized as point estimates and standard errors.

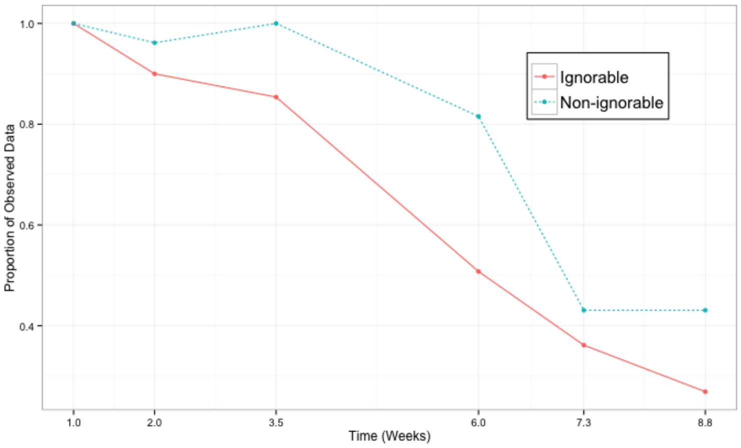


Figure 5.17: Proportions of observed responses in both ignorable and non-ignorable missing data generation in TELL study

5.4.3 Results

In this section we compare and summarize parameter estimation in the linear mixed model which is of primary interest in modeling longitudinal changes on VOCE standard scores. Complete data were analyzed using the linear mixed model which examined growth curve trends for children in the TELL and control conditions at six time points. The model included random effects for intercept, linear slope and fixed effects for covariates and experimental condition (TELL, control). Mother education was also included as a covariate. Since we conducted the analysis from the first follow-up visit, that is, we exclude baseline information, a main effect of the TELL treatment on the intercept was anticipated. The estimation results are given in both Tables 5.7 and 5.8, under the 'Complete' column. The corresponding standard errors were given in the parentheses. In Table 5.7 the parameter estimates on ignorable missing data were also given: conventional likelihood estimates which ignores missing values and standard errors were given under the column 'MAR'; both likelihood estimates and bayesian estimates from CLFM were given in the columns 'MCEM' and 'Bayesian', respectively. By comparison, Table 5.8 listed parameter estimates when missing data cannot be ignored. When all complete data were considered in the model, there was a strong evidence to show that a significant linear time by condition interaction. Time by condition interaction results suggest that while the control condition reached a plateau in terms of VOCE score growth, the TELL condition showed evidence of continued growth by the end of the school year, which is indicating positive effects of TELL curriculum for DSLI preschool children. Results suggested that the TELL treatment produced a shift in the growth curve for VOCE scores where the peak for the control condition was achieved around week 7, and the VOCE score kept increasing until the end of the study.

Missing data which are ignorable were designed to conduct a robust analysis for CLFM. That is, CLFM is evaluated when the missing data assumption was misspecified for this model. From Table 5.7 we can observe that identical estimates were obtained from both estimation techniques: MCEM approach and Bayesian framework. These results were expected since Bayesian theories tell us when non-informative priors were provided, the resulting estimates are identical to those obtained from classic frequentist perspectives. In estimating CLFM from the Bayesian approach, we applied conjugate priors for all parameters that were contained in the model, by providing large variances in the prior distribution. Compared with estimates from complete data, the conventional model with assumption of ignorable missing data produced closer estimates and had a better overall performance. However, the proposed CLFM can still generate plausible results, especially in evaluating those terms which are of primary interest, including overall treatment efficacy, as well as treatment effects' changes in terms of VOCE scores across the whole study period. However, a noticeable departure on estimating linear time trend was also observed for all approaches.

Table 5.7: Ignorable missing data: point estimates (standard error) on regression coefficients (fixed effects) and variance components (random effects) for the VOCE standard scores from a MAR model and CLFM in both approaches

Variables	Complete	Ignorable Missing		
		MAR	MCEM	Bayesian
Intercept	2.218 (0.238)	2.498 (0.251)	2.578 (0.317)	2.578 (0.318)
t_{ij}	-0.386 (0.184)	-0.678 (0.202)	-0.652 (0.209)	-0.6526 (0.211)
t_{ij}^2	0.178 (0.042)	0.261 (0.049)	0.302 (0.047)	0.303 (0.048)
t_{ij}^3	-0.012 (0.003)	-0.019 (0.003)	-0.021 (0.003)	-0.020 (0.003)
Treatment	-1.441 (0.234)	-1.582 (0.246)	-1.563 (0.294)	-1.563 (0.295)
Treatment x t_{ij}	-0.157 (0.033)	-0.116 (0.039)	-0.115 (0.058)	-0.116 (0.058)
Matedu	0.192 (0.065)	0.211 (0.067)	0.208 (0.072)	0.209 (0.073)
$\sigma_{b_0}^2$	0.924 (0.160)	0.992 (0.178)	0.900 (0.175)	0.901 (0.176)
$\sigma_{b_1}^2$	0.010 (0.003)	0.008 (0.005)	0.235 (0.030)	0.235 (0.032)
σ_{ϵ}^2	0.990 (0.061)	1.010 (0.074)	0.940 (0.063)	0.941 (0.064)

With the assumed missing data, both models tended to produce biased estimates. That is, the decreasing rate on VOCE score with linear time trend seemed to be exaggerated in both the conventional model and CLFM. Table 5.8 included the case where missing values cannot be ignored and should be considered in the analysis. With the correct missing data assumption, CLFM provided more accurate estimates compared with the model that excluded missing values in the analysis. Two warnings may be noticed in this case: conventional model cannot estimate variance component for the random slope term, due to singularity of Hessian matrix; the other is we observed that the model excludes missing values cannot estimate correctly the interaction term between condition group and time. The complete data analysis told that this interaction term should have a negative value, which was also supported by CLFM; however, the conventional model that does not include these non-ignorable missing data produced an estimate with the opposite sign.

Table 5.8: Non-ignorable missing data: point estimates (standard error) on regression coefficients (fixed effects) and variance components (random effects) for the VOCE standard scores from a MAR model and CLFM in both approaches

Variables	Complete	Non-ignorable Missing		
		MAR	MCEM	Bayesian
Intercept	2.218 (0.238)	2.517 (0.228)	2.349 (0.315)	2.350 (0.316)
t_{ij}	-0.386 (0.184)	-0.640 (0.197)	-0.353 (0.218)	-0.353 (0.219)
t_{ij}^2	0.178 (0.042)	0.189 (0.049)	0.180 (0.049)	0.181 (0.050)
t_{ij}^3	-0.012 (0.003)	-0.015 (0.004)	-0.012 (0.003)	-0.013 (0.004)
Treatment	-1.441 (0.234)	-1.830 (0.206)	-1.422 (0.232)	-1.422 (0.233)
Treatment x t_{ij}	-0.157 (0.033)	0.129 (0.038)	-0.160 (0.035)	-0.161 (0.036)
Matedu	0.192 (0.065)	0.174 (0.052)	0.181 (0.071)	0.180 (0.071)
$\sigma_{b_0}^2$	0.924 (0.160)	0.547 (0.128)	0.907 (0.192)	0.908 (0.192)
$\sigma_{b_1}^2$	0.010 (0.003)		0.008 (0.004)	0.008 (0.004)
σ_{ϵ}^2	0.990 (0.061)	0.827 (0.066)	0.942 (0.059)	0.942 (0.065)

With the complete information from VOCE study, the linear mixed model suggested several terms were significant in accounting for VOCE changes for DSLI children, including mother education, as well as linear, quadratic and cubical time treands. A significant interaction between linear time with condition was also concluded, which was further confirmed the efficacy of TELL curriculum for DSLI preschool children. Based on these findings, a robust analysis on CLFM was explored, with two specified missing data assumptions: ignorable missing data and non-ignorable missingness. Each dataset was simulated from Diggle-Kenward model, with different model settings. The proposed CLFM and classical approach were adopted to fit each simulated senario. As we expected, CLFM performed better in the case of which non-ignorable missing data are a feature of the study; furthermore, CLFM also provided robust estimates, even with a misspecified missing data mechanism, that is, ignorable missing data.

In this chapter, CLFM was investigated in detail and the corresponding effectiveness was further confirmed through series of simulation studies and three applications. In estimating parameters of CLFM from the MCEM approach, a heavy computational burden was involved. This computation burden sometimes can be alleviated by specifying different initial values for MCEM algorithm. When using estimates from ignorable likelihood as initial values in AIDS Clinical Trial study, the computation time was reduced by 1/3 from the arbitrary specified initial values. As proposed in Chapter 4, missing patterns were estimated by a latent factor model, by incorporating with a continuous latent factor $\mathbf{u}$, and time location parameters. The constant variability across each time is assumed among all applications we presented in this chapter. However, there maybe cases that heterogeneous variability will produce a better fit. A likelihood ratio test can be used to determine which model is better. Table 5.9 gives the log-likelihood values on fitting latent factor models with constant or heterogeneous variability, as well as p-values on likelihood ratio test on each application.

From Table 5.9 we observed that continuous latent factor models with heterogeneous slope parameters were preferred for the first two studies: Peabody Vocabulary Test and ACTG studies. A separated study was conducted for ACTG case, and for

Table 5.9: Log-likelihood values and likelihood ratio tests on latent factor models with constant variability (Constraint) or heterogeneous variability (Full)

Study	Constraint	Full	df	p-value
PIAT	-639.7661	-622.9483	3	< 0.0001
ACTG	-3649.173	-3629.591	5	< 0.0001
Growth NMAR	-259.2228	-257.1922	4	0.3978
Growth MAR	-218.1306	-215.8879	4	0.3443

the primary parameters of interest were similar to those obtained from the continuous latent factor model with homogeneous slope.

Chapter 6

DISCUSSION

6.1 Conclusions

In a longitudinal study, an incomplete dataset does not contain information that enables us to identify underlying a missing mechanism, unless extra unverifiable assumptions can be made. In the last two decades, researchers have investigated the implications of NMAR missing data by fitting selection models and pattern-mixture models. However, these models include difficulties to implement in a real case. Selection models make unverifiable assumptions for the missing mechanism, while pattern-mixture models tend to have over-parameterization issues, as well as conditional independence assumptions. In this thesis, we developed a non-ignorable model based on the idea of continuous latent factor of response behavior (missing behavior), and argue that this model excludes most implementing difficulties and is a useful alternative to a standard analysis with MAR assumption.

We believe that this new approach will avoid untestable missing mechanism assumptions from selection models, and also believe that the new model will be more appealing to social behavioral and clinical researchers than pattern-mixture models,because the new model eliminates over-parameterizations issues. Further, the continuous latent factor provides an intuitive description of the response patterns in the study, and offers a feasible way to test conditional independence assumptions. For researchers who are interested in implementing CLFM model, we encourage them to compare latent factor models on missing indicator matrix with either constant slope or heterogeneous slopes and choose the one with better fitting in CLFM, based on information criteria or the likelihood ratio test. Lastly, CLFM is more feasible for small samples.

With the truth that the underlying missing mechanism for missing data is unknown, (that is whether missingness is due to MAR or NMAR), we take this new method primarily as a tool for sensitivity analysis. In the case that a researcher cannot determine the distribution of missing data, the most responsible and objective approach to proceed is to explore and present alternative results from different plausible models.

6.2 Future Work

In this thesis, we have explored the proposed CLFM under the assumption of a multivariate normal distribution for the complete data. The normal model is an intuitive and natural starting point for this method, but it also has limitations. Many longitudinal studies will have discrete responses, such as measuring the total number of bleeding counts in a Hemophilia study; or even binary responses. In the future, we will be extending our method to more flexible models for multivariate discrete responses. One promising approach is the Bayesian estimation approach which allows these extensions more straightforward.

To achieve an in-depth understanding of our method's properties, it is desirable to perform more simulation studies to compare this method to existing MAR and NMAR alternatives under a variety of missing data mechanisms. Only one robust analysis has been done in this thesis, and we are expected to conduct more simulation studies on this topic. Some might regard them as artificial, because in each realistic example the true mechanism is unknown. Nevertheless, it would be interesting to explore whether the proposed model performs better or worse than other methods when its assumptions are violated.

In proposing CLFM, we have a fundamental assumption which is conditional independence. Unlike models that belong to pattern mixture family, this assumption is feasible to be tested in CLFM. As another future work, we will explore the assessment on this assumed conditional independence in the CLFM from the fitted residuals. One approach is to calculate the residual from both the longitudinal and missing pattern models. When these residuals can be treated as approximately iid normal, a correlation coefficient close to 0 will indicate the conditional independence. For a more complicated distribution, some graphical approaches may be useful and could be applied as auxiliary tools.

REFERENCES

Adams, W. M. W. M. L., R. J., “Multilevel item response models: An approach to errors in variables regression”, Journal of Educational and Behavioral Statistics **22**, 47–76 (1997).

Aitkin, A. D. H. J., M., “Statistical modeling of data on teaching styles (with discussion)”, Journal of Royal Statistics Society Series A **144**, 419–461 (1981).

Aitkin, R. D., M., “Estimation and hypothesis testing in finite mixture models”, Journal of Royal Statistics Society Series B **47**, 67–75 (1985).

Bartholomew, D. J., *Latent variable models and factor analysis* (Oxford University Press, 1987).

Bock, R. and M. Aitkin, “Marginal maximum likelihood estimation of item parameters: Application of an em algorithm”, Psychometrika **46**, 443–458 (1981).

Bozdogan, H., “Model selection and akaikes information criterion (aic): the general theory and its analytic extensions”, Psychometrika **52**, 345–370 (1987).

Clogg, C., *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, chap. 6, pp. 311–359 (Plenum Press, 1995).

Dempster, A. P., N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm”, Journal of the Royal Statistical Society. Series B (Methodological) **39**, 1, pp. 1–38 (1977).

Diggle, K., P. Liang and S. Zeger, *Analysis of Longitudinal Data* (Oxford University Press, 1994).

Diggle, P. and M. Kenward, “Informative drop-out in longitudinal data analysis”, Applied Statistics **43**, 49–73 (1994a).

Diggle, P. and M. Kenward, “Pattern-mixture models for multivariate incomplete data”, Journal of the American Statistical Association **88**, 125–134 (1994b).

Draper, D., “Assessment and propagation of model uncertainty”, Journal of Royal Statistics Society Series B **57**, 45–97 (1995).

Embretson, S. E. and S. P. Reise, *Item response theory for psychologists* (Mahwah, NJ: Erlbaum, 2000).

Fitzmaurice, G. and N. M. Laird, *Applied Longitudinal Analysis* (Wiley Series in Probabiity and Statistics, 2004).

Fitzmaurice, L. N., G.M. and J. Ware, *Applied Longitudinal Analysis* (New York: John Wiley and Sons., 2004).

Garrett, E. S. and S. L. Zeger, “Latent class model diagnosis”, Biometrics **56**, 4, 1055–1067 (2000).

Geman, S. and D. Geman, “Stochastic relaxation, gibbs distributions, and the bayesian restoration of images”, IEEE Transactions on Pattern Analysis and Machine Intelligence **6**, 721–741 (1984).

Goodman, L., *Analyzing qualitative/categorical data* (Abt Books, 1978).

Guo, W., S. J. Ratcliffe and T. T. T. Have, “A random pattern-mixture model for longitudinal data with dropouts”, Journal of the American Statistical Association **99**, 468, pp. 929–937 (2004).

Hannan, Q. B., E.J., “The determination of the order of an autoregression”, Journal of Royal Statistics Society Series B **41**, 190–195 (1979).

Hastings, W., “Monte carlo sampling methods using markov chains and their application”, Biometrika **57**, 97–109 (1970).

Haughton, D., “On the choice of a model to fit data from an exponential family”, Annal Statistics **16**, 342–355 (1988).

Henry, K. and A. Erice, “A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies for the treatment of advanced aids”, Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology **19**, 3, 339–349 (1998).

Horton, N. J. and G. M. Fitzmaurice, “Maximum likelihood estimation of bivariate logistic models for incomplete responses with indicators of ignorable and non-ignorable missingness”, Journal of the Royal Statistical Society. Series C (Applied Statistics) **51**, 3, pp. 281–295 (2002).

Hurvich, T. C., C.M., “Regression and time series model selection in small samples”, Biometrika **76**, 297–307 (1989).

Jung, S. J., Hyekyung and B. Seo, “A latent class selection model for nonignorably missing data”, Computational Statistics; Data Analysis **55**, 1, 802 – 812 (2011).

Laird, N. M. and J. H. Ware, “Random-effects models for longitudinal data”, Biometrics **38**, 4, pp. 963–974 (1982).

Lazarsfeld and P.F., *The interpretation and mathematical foundation of latent structure analysis*, pp. 413–472 (Princeton University Press, Princeton, 1950a).

Lazarsfeld and P.F., *The logical and mathematical foundation of latent structure analysis*, pp. 362–412 (Princeton University Press, Princeton, 1950b).

Lee, S.-Y. and X.-Y. Song, “Maximum likelihood estimation and model comparison for mixtures of structural equation models with ignorable missing data”, Journal of Classification **20**, 221–255 (2003).

Lin, H., C. E. McCulloch and R. A. Rosenheck, “Latent pattern mixture models for informative intermittent missing data in longitudinal studies”, Biometrics **60**, 2, 295–305 (2004).

Little, R. J. A., “Modeling the drop-out mechanism in longitudinal studies”, Journal of the American Statistical Association **90**, 1112–1121 (1995).

Little, R. J. A. and D. B. Rubin, *Statistical Analysis with Missing Data* (Wiley Series in Probability and Statistics, 2002).

Lord, F., “A theory of test scores”, Psychometric **No. 7** (1952).

Lord, F., “The relation of test score to the trait underlying the test”, Educational and Psychological Measurement **13**, 517–548 (1953).

Lord, F., *Applications of item response theory to practical testing problems* (Hillsdale, NJ: Erlbaum, 1980).

Louis, T., “Finding the observed information matrix when using the em algorithm”, Journal of the Royal Statistical Society, Series B. **44**, 226–233 (1982).

Lunn, D. and D. Spiegelhalter, “Winbugs a bayesian modelling framework: concepts, structure, and extensibility”, Statistics and Computing **10**, 325–337 (2000).

McCulloch, C. E. and S. R. Searle, *Generalized, Linear, and Mixed Models* (New York: Wiley, 2001).

McHugh, R., “Efficient estimation and local identification in latent class analysis”, Psychometrika **21**, 331–47 (1956).

Meng, X. and S. Schilling, “Fitting full-information item factor models and an empirical investigation of bridge sampling”, Journal of American Statistical Association **91**, 1254–1267 (1996).

Muthen, B., “Contributions to factor analysis of dichotomous variables”, Psychometrika **43**, 551–560 (1978).

Muthen, L. and B. Muthen, *Mplus User’s Guide. Fifth Edition.* (Los Angeles, CA, 1998-2011).

Muthn, B., B. Jo and C. H. Brown, “Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city [with comment]”, Journal of the American Statistical Association **98**, 462, pp. 311–314 (2003).

Pirie, M. D., P.L. and R. Leupker, “Smoking prevalence in a cohort of adolescents, including absentees, dropouts, and transfers”, American Journal of Public Health **78**, 176–178 (1988).

Rasch, G., *Probabilistic models for some intelligence and attainment tests* (Chicago, IL: University of Chicago Press., 1960).

Raudenbush, J. C. S. R., S. W., “A multivariate, multilevel rasch model with applications to self-reported criminal behavior”, Sociological Methodology **33**, 169–211 (2003).

Rijmen, T. F. D. B. P. K. P., F., “A nonlinear mixed model framework for item response theory”, Psychological Methods **8**, 185–205 (2003).

Robert, C. and G. Casella, *Introducing Monte Carlo Methods with R* (Springer, 2010).

Roy, J., “Modeling longitudinal data with nonignorable dropouts using a latent dropout class model”, Biometrics **59**, 4, 829–836 (2003).

Roy, J., “Latent class models and their applications to missing-data patterns in longitudinal studies”, Statistical Methods in Medical Research **16**, 441–456 (2007).

Rubin, D., “Inference and missing data”, Biometrika **63**, 581–592 (1976).

Rubin, D. B., *Multiple imputation for survey nonresponse* (J. Wiley Sons, New York, 1987).

Rusakov, D. and D. Geigerm, “Asymptotic model selection for naive bayesian networks”, Journal of Machine Learning Research **6**, 1–35 (2005).

Schafer, J., *Analysis of Incomplete Multivariate Data* (Chapman and Hall, New York, 1997).

Schwarz, “Estimating the dimension of a model”, Annal Statistics **6**, 461–464 (1978).

S.E. Fienberg, A. Y., P.Hersh, “Maximum likelihood estimation in latent class models for contingency table”, (2007).

Settimi, R. and J. Smith, “Geometry, moments and conditional independence trees with hidden variables”, Annals of Statistics **28**, 1179–1205 (2005).

Smith, J. and J. Croft, “Bayesian networks for discrete multivariate data: an algebraic approach to inference”, Journal of Multivariate Analysis **84**, 387–402 (2003).

Sturtz, S. and A. Gelman, “R2winbugs a package for running winbugs from r”, Journal of Statistical Software **12**, 3, 1–16 (2005).

Takane, d. L. J., Y., “On the relationship between item response theory and factor analysis of discretized variables”, Psychometrica **52**, 393–408 (1987).

Verbeke, G. and G. Molenberghs, *Linear Mixed Models for Longitudinal Data* (New York: Springer, 2000).

Weber, A. M., “Peabody picture vocabulary test”, (2007).

Wei, G. and M. Tanner, “A monte carlo implementation of the em algorithm and the poor mans data augmentation algorithms”, Journal of the American Statistical Association **85**, 699–704 (1990).

Wilcox, M. and S. Gray, “Efficacy of the tell language and literacy curriculum for preschoolers with developmental speech and/or language impairment”, Early Childhood Research Quarterly **26**, 278–294 (2011).

Woodruffe, M., “On model slection and the arcsine laws”, Annal Statistics **10**, 1182–1194 (1982).

Wu, C., “On the convergence of properties of the em algorithm”, Annal Statistics **11**, 95–103 (1983).

Yang., C., “Evaluating latent class analysis models in qualitative phenotype identification”, Computational Statistics and Data Analysis **50**, 1090–1104 (2006).

Zhang, Z. and F. Hamagami, “Bayesian analysis of longitudinal data using growth curve models”, International Journal of Behavioral Development **31 (4)**, 374–383 (2007).

APPENDIX A

MORE SIMULATION STUDIES ON TOPIC I

Table A.1: Number of latent class tallies on MCAR simulation

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	1 (0.001)	112 (0.112)	0 (0.00)	883 (0.887)
BIC	0 (0.00)	978 (0.982)	18 (0.018)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	989 (0.993)	7 (0.007)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	894 (0.898)	102 (0.102)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	593 (0.595)	368 (0.369)	0 (0.00)	35 (0.035)
HT	0 (0.00)	2 (0.002)	160 (0.161)	0 (0.00)	834 (0.837)
BICa	0 (0.00)	536 (0.538)	403 (0.405)	0 (0.00)	57 (0.057)
CAICa	0 (0.00)	549 (0.551)	396 (0.398)	0 (0.00)	51 (0.051)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = 0$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.2: Number of latent class tallies on NMAR simulation (low missing probability)

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	1 (0.001)	112 (0.112)	0 (0.00)	883 (0.887)
BIC	0 (0.00)	978 (0.982)	18 (0.018)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	989 (0.993)	7 (0.007)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	894 (0.898)	102 (0.102)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	593 (0.595)	368 (0.369)	0 (0.00)	35 (0.035)
HT	0 (0.00)	2 (0.002)	160 (0.161)	0 (0.00)	834 (0.837)
BICa	0 (0.00)	536 (0.538)	403 (0.405)	0 (0.00)	57 (0.057)
CAICa	0 (0.00)	549 (0.551)	396 (0.398)	0 (0.00)	51 (0.051)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = 0.2$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.3: Number of latent class tallies on NMAR simulation (different missing probability at different time points)

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	7 (0.007)	360 (0.364)	430 (0.435)	192 (0.194)
BIC	746 (0.754)	224 (0.226)	19 (0.019)	0 (0.00)	0 (0.00)
CAIC	856 (0.866)	132 (0.133)	1 (0.001)	0 (0.00)	0 (0.00)
DBIC	458 (0.463)	400 (0.404)	127 (0.128)	4 (0.004)	0 (0.00)
HQ	104 (0.105)	355 (0.359)	439 (0.444)	90 (0.091)	1 (0.001)
HT	0 (0.00)	12 (0.012)	397 (0.401)	424 (0.429)	156 (0.158)
BICa	82 (0.083)	318 (0.322)	471 (0.476)	116 (0.117)	2 (0.002)
CAICa	89 (0.090)	330 (0.334)	460 (0.465)	109 (0.110)	1 (0.001)

*Latent class models are fitted without incorporating covariates. $\alpha_j = 1$, $\gamma_1 = [0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6]$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.4: Number of latent class tallies on NMAR simulation (different missing probability at different time points)

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	1 (0.001)	112 (0.112)	0 (0.00)	883 (0.887)
BIC	0 (0.00)	978 (0.982)	18 (0.018)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	989 (0.993)	7 (0.007)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	894 (0.898)	102 (0.102)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	593 (0.595)	368 (0.369)	0 (0.00)	35 (0.035)
HT	0 (0.00)	2 (0.002)	160 (0.161)	0 (0.00)	834 (0.837)
BICa	0 (0.00)	536 (0.538)	403 (0.405)	0 (0.00)	57 (0.057)
CAICa	0 (0.00)	549 (0.551)	396 (0.398)	0 (0.00)	51 (0.051)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = [0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6]$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.5: Number of latent class tallies for low missing probability on previous responses

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	20 (0.020)	80 (0.080)	249 (0.251)	654 (0.649)
BIC	0 (0.00)	991 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	991 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	960 (0.966)	34 (0.034)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	762 (0.767)	198 (0.199)	30 (0.030)	4 (0.004)
HT	0 (0.00)	29 (0.029)	122 (0.123)	285 (0.288)	558 (0.561)
BICa	0 (0.00)	711 (0.715)	220 (0.221)	53 (0.053)	10 (0.01)
CAICa	0 (0.00)	721 (0.725)	218 (0.219)	45 (0.045)	10 (0.01)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = 0.2$, $\gamma_2 = 0.1$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.6: Number of latent class tallies for high missing probability on previous responses

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	0 (0.00)	3 (0.003)	286 (0.286)	710 (0.711)
BIC	0 (0.00)	293 (0.293)	508 (0.509)	198 (0.198)	0 (0.00)
CAIC	0 (0.00)	293 (0.293)	508 (0.509)	198 (0.198)	0 (0.00)
DBIC	0 (0.00)	31 (0.031)	537 (0.534)	388 (0.388)	43 (0.043)
HQ	0 (0.00)	1 (0.001)	249 (0.249)	516 (0.517)	233 (0.233)
HT	0 (0.00)	0 (0.00)	4 (0.004)	313 (0.313)	682 (0.682)
BICa	0 (0.00)	1 (0.001)	212 (0.212)	517 (0.518)	269 (0.269)
CAICa	0 (0.00)	1 (0.001)	216 (0.216)	520 (0.521)	262 (0.262)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table A.7: Number of latent class tallies for high missing probability on previous and current responses

Information Criterion	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	0 (0.00)	2 (0.002)	164 (0.165)	827 (0.833)
BIC	0 (0.00)	1 (0.001)	953 (0.960)	39 (0.039)	0 (0.00)
CAIC	0 (0.00)	1 (0.001)	953 (0.960)	39 (0.039)	0 (0.00)
DBIC	0 (0.00)	0 (0.00)	724 (0.729)	257 (0.259)	12 (0.012)
HQ	0 (0.00)	0 (0.00)	307 (0.309)	524 (0.528)	162 (0.163)
HT	0 (0.00)	0 (0.00)	4 (0.004)	211 (0.212)	778 (0.783)
BICa	0 (0.00)	0 (0.00)	261 (0.263)	535 (0.539)	197 (0.198)
CAICa	0 (0.00)	0 (0.00)	271 (0.273)	530 (0.534)	192 (0.193)

*Latent class models are fitted with incorporating covariates. $\alpha_j = 1$, $\gamma_1 = 0.4$, $\gamma_2 = 0.4$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

APPENDIX B

SIMULATION RESULTS FOR CLFM

Table B.1: Parameter estimation in linear mixed model for convectional model (MAR),latent class model (Roy), and CLFM. In this simulation study with 80 individuals, half of them complete the study and the average missing proportion is 13 percent. CLFM was fitted by Bayesian framework

Variables	True	MAR			Roy			CLFM		
		Estimate	SE	RMSE	Estimate	SE	RMSE	Estimate	SE	RMSE
Intercept	3	3.021	0.164	0.186	3.215	0.243	0.780	3.008	0.209	0.191
t_{ij}	1	0.980	0.054	1.981	0.950	0.088	0.331	0.987	0.081	0.089
Age	2	2.004	0.097	0.100	1.942	0.091	0.125	2.001	0.086	0.087
Group	1	1.008	0.191	0.193	0.988	0.203	0.201	1.008	0.174	0.175
$Var(b_{0i}) = \sigma^2_{b_0}$	1	0.940	0.231	0.252	0.632	0.196	0.458	1.008	0.201	0.232
$Var(b_{1i}) = \sigma^2_{b_1}$	0.2	0.191	0.037	0.039	0.124	0.029	0.090	0.196	0.039	0.038
$Cov(b_{0i}, b_{1i}) = \sigma_{b_0 b_1}$	-0.3	-0.283	0.079	0.083	-0.199	0.064	0.147	-0.291	0.056	0.060
$Var(e_i) = \sigma^2_e$	0.5	0.497	0.044	0.045	0.457	0.119	0.133	0.488	0.050	0.056

Table B.2: Parameter estimation in linear mixed model for convectional model (MAR),latent class model (Roy), and CLFM. In this simulation study with 300 individuals, half of them complete the study and the average missing proportion is 20 percent. CLFM was fitted by Bayesian framework

Variables	True	MAR			Roy			CLFM		
		Estimate	SE	RMSE	Estimate	SE	RMSE	Estimate	SE	RMSE
Intercept	3	3.025	0.086	0.099	2.913	0.160	0.218	3.002	0.094	0.100
t_{ij}	1	0.982	0.028	1.982	0.999	0.057	0.217	0.992	0.032	0.069
Age	2	1.998	0.050	0.051	1.924	0.054	0.092	2.001	0.043	0.040
Group	1	0.996	0.099	0.109	0.955	0.097	0.114	0.999	0.097	0.100
$Var(b_{0i}) = \sigma^2_{b_0}$	1	0.968	0.121	0.131	0.926	0.124	0.146	1.008	0.130	0.132
$Var(b_{1i}) = \sigma^2_{b_1}$	0.2	0.191	0.019	0.022	0.154	0.019	0.052	0.195	0.010	0.017
$Cov(b_{0i}, b_{1i}) = \sigma_{b_0 b_1}$	-0.3	-0.287	0.041	0.045	-0.264	0.044	0.057	-0.303	0.047	0.038
$Var(e_i) = \sigma^2_e$	0.5	0.500	0.023	0.024	0.472	0.067	0.068	0.503	0.016	0.023

Table B.3: Parameter estimation in linear mixed model for convectional model (MAR),latent class model (Roy), and CLFM. In this simulation study with 80 individuals, few of them complete the study and the average missing proportion is 70 percent. CLFM was fitted by Bayesian framework

Variables	True	MAR			Roy			CLFM		
		Estimate	SE	RMSE	Estimate	SE	RMSE	Estimate	SE	RMSE
Intercept	3	3.278	0.197	0.322	3.225	0.560	0.338	3.168	0.280	0.246
t_{ij}	1	0.721	0.100	0.289	0.749	0.225	0.319	0.807	0.192	0.217
Age	2	1.965	0.114	0.126	1.951	0.165	0.148	1.969	0.158	0.123
Group	1	0.986	0.323	0.386	0.953	0.334	0.314	1.003	0.460	0.294
$Var(b_{0i}) = \sigma^2_{b_0}$	1	0.913	0.519	0.473	0.370	1.119	0.752	0.934	0.515	0.554
$Var(b_{1i}) = \sigma^2_{b_1}$	0.2	0.179	0.093	0.098	0.027	0.075	0.175	0.195	0.050	0.042
$Cov(b_{0i}, b_{1i}) = \sigma_{b_0 b_1}$	-0.3	-0.276	0.208	0.336	-0.040	0.295	0.290	-0.294	0.120	0.163
$Var(e_i) = \sigma^2_e$	0.5	0.515	0.162	0.177	0.600	0.065	0.268	0.578	0.084	0.132

Table B.4: Parameter estimation in linear mixed model for convectional model (MAR),latent class model (Roy), and CLFM. In this simulation study with 300 individuals, few of them complete the study and the average missing proportion is 70 percent. CLFM was fitted by Bayesian framework

Variables	True	MAR			Roy			CLFM		
		Estimate	SE	RMSE	Estimate	SE	RMSE	Estimate	SE	RMSE
Intercept	3	3.278	0.097	0.282	2.809	0.178	0.501	3.103	0.218	0.227
t_{ij}	1	0.722	0.060	0.289	0.842	0.110	0.249	0.871	0.201	0.234
Age	2	1.945	0.074	0.096	1.923	0.066	0.137	1.993	0.070	0.059
Group	1	0.965	0.132	0.167	0.990	0.159	0.166	0.997	0.193	0.196
$Var(b_{0i}) = \sigma^2_{b_0}$	1	1.278	0.252	0.322	0.783	0.344	0.404	1.067	0.243	0.305
$Var(b_{1i}) = \sigma^2_{b_1}$	0.2	0.179	0.053	0.208	0.144	0.049	0.078	0.193	0.067	0.126
$Cov(b_{0i}, b_{1i}) = \sigma_{b_0 b_1}$	-0.3	-0.476	0.108	0.236	-0.246	0.120	0.133	-0.307	0.112	0.141
$Var(e_i) = \sigma^2_e$	0.5	0.504	0.062	0.087	0.730	0.485	0.613	0.621	0.145	0.137

Table B.5: Parameter estimation in linear mixed model for convectional model (MAR),latent class model (Roy), and CLFM. In this simulation study with 300 individuals, treatment group has higher missing proportion, compared with control group. CLFM was fitted by Bayesian framework

Variables	True	MAR			Roy			CLFM		
		Estimate	SE	RMSE	Estimate	SE	RMSE	Estimate	SE	RMSE
Intercept	3	3.015	0.071	0.103	2.828	0.917	1.047	3.007	0.075	0.081
Slope	-1	-1.041	0.035	0.056	-1.372	0.349	0.587	-1.001	0.043	0.043
Age	2	1.987	0.051	0.072	2.048	0.278	0.258	1.998	0.064	0.0669
Group x Slope	-0.5	-0.472	0.039	0.029	0.015	0.128	0.551	-0.495	0.055	0.058
$Var(b_{0i}) = \sigma^2_{b_0}$	1	0.972	0.127	0.177	0.516	0.522	0.679	1.013	0.134	0.202
$Var(b_{1i}) = \sigma^2_{b_1}$	0.2	0.190	0.019	0.026	0.171	0.049	0.053	0.201	0.024	0.064
$Cov(b_{0i}, b_{1i}) = \sigma_{b_0 b_1}$	-0.3	-0.288	0.042	0.059	-0.206	0.188	0.195	-0.301	0.047	0.055
$Var(e_i) = \sigma^2_e$	0.5	0.500	0.025	0.032	0.627	0.175	0.203	0.589	0.117	0.200

APPENDIX C

REGULARITY CONDITIONS

Given complete data $\mathbf{Y}$ and $\mathbf{R}$, and parameter vector $\theta \in \Theta$ for the proposed parametric model (4.2) - (4.4), the regularity conditions for discussing asymptotic properties of maximum likelihood estimators (MLE) can be stated as follows:

1. Both variables $(\mathbf{Y}_i,\ \mathbf{R}_i,\ i = 1, 2, \cdots$ are independent and identically distributed with density function $f(\mathbf{Y}, \mathbf{R}; \theta)$.

2. The parameter space Θ is compact, and there exists a $\theta_0 \in Int(\Theta)$ (i.e. θ_0 is an interior point of Θ) such that $\theta_0 = \underset{\theta \in \Theta}{argmax}\ \mathbf{E}_{\theta_0}\ log\ f(\mathbf{Y}_i, \mathbf{R}_i; \theta)$.

3. The probability distribution is identifiable, i.e. for different values of θ, the probability distributions are distinct.

4. The log-likelihood function

$$l(\mathbf{Y}, \mathbf{R}; \theta) = \sum_{i=1}^{n} log\ f(\mathbf{Y}_i, \mathbf{R}_i; \theta)$$

   is continuous at θ.

5. $\mathbf{E}_{\theta_0}\ log\ f(\mathbf{Y}_i, \mathbf{R}_i; \theta)$ exists.

6. The log-likelihood function satisfies that $\frac{1}{n} l(\mathbf{Y}, \mathbf{R}; \theta)$ converges almost surely to $\mathbf{E}_{\theta_0}\ log\ f(\mathbf{Y}_i, \mathbf{R}_i; \theta)$ uniformly in $\theta \in \Theta$, i.e.,

$$\sup_{\theta \in \Theta} \left| \tfrac{1}{n} l(\mathbf{Y}, \mathbf{R}; \theta) - \mathbf{E}_{\theta_0}\ log\ f(\mathbf{Y}_i, \mathbf{R}_i; \theta) \right| < \delta \text{ almost surely for some } \delta > 0.$$

7. The log-likelihood function $l(\mathbf{Y}, \mathbf{R}; \theta)$ is twice continuously differentiable in a neighborhood of θ_0.

8. Integration and differential operators are interchangeable.

9. The information matrix

$$I(\theta_0) = \mathbf{E}_{\theta_0} \left(\frac{\partial^2 log\ f(\mathbf{Y}, \mathbf{R}; \theta_0)}{\partial\theta\partial\theta^T} \right)$$

   exists and non-singular.