Nonlinear Mixed-Effects Models for HIV Viral Load Trajectories Before and After Antiretroviral Therapy Interruption, Incorporating Left Censoring

by

Sihaoyu Gao

B.Sc., The University of British Columbia, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Nonlinear Mixed-Effects Models for HIV Viral Load Trajectories Before and After Antiretroviral Therapy Interruption, Incorporating Left Censoring

submitted by **Sihaoyu Gao** in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** in **Statistics**.

Examining Committee:

Lang Wu, Statistics, UBC
Supervisor

M. Ehsan Karim, School of Population and Public Health, UBC
Additional Examiner

Abstract

Longitudinal studies are common in biomedical research, such as an HIV study. In an HIV study, the viral decay during an anti-HIV treatment and the viral rebound after the treatment is interrupted can be viewed as two longitudinal processes, and they may be related to each other. In this thesis, we investigate if key features of HIV viral decay and CD4 trajectories during antiretroviral therapy (ART) are associated with characteristics of HIV viral rebound following ART interruption. Motivated by a real AIDS dataset, two non-linear mixed effects (NLME) models are used to model the viral load trajectories before and following ART interruption, respectively, incorporating left censoring due to lower detection limits of viral load assays. A linear mixed effects (LME) model is used to model CD4 trajectories. The models may be linked through shared random effects, since these random effects reflect individual characteristics of the longitudinal processes. A stochastic approximation EM (SAEM) method is used for parameter estimation and inference. To reduce the computation burden associated with maximizing the joint likelihood, an easy-to-implement three-step (TS) method is proposed by using SAEM algorithm and bootstrap. Data analysis results show that some key features of viral load and CD4 trajectories during ART (e.g., viral decay rate) are significantly associated with important characteristics of viral rebound following ART interruption (e.g., viral set point). Simulation studies are conducted to evaluate the performances of the proposed TS method and the naive method, which still uses SAEM algorithm but substitutes the censored viral load values with half the detection limit and without bootstrap. It is concluded that the proposed TS method outperforms the naive method.

Lay Summary

Motivated by an HIV study, a major interest of this piece of research is to investigate if key features of viral decay during ART are associated with individual-specific characteristics of viral rebound following ART interruption. Due to computational complexity of joint modelling, this thesis proposed a three-step method by using a stochastic approximation EM algorithm and a bootstrap method to analyze three linked LME and NLME models with censoring. Data analysis results show that some key features of viral load and CD4 trajectories during ART (e.g., viral decay rate) are significantly associated with important characteristics of viral rebound following ART interruption (e.g., viral set point). These findings may provide insights into HIV cure research. In addition, simulation studies show that the three-step method performs reasonably well.

Preface

This thesis is formed based on a paper jointly authored with Professor Lang Wu, Professor Rui Wang, and Tingting Yu. This paper has been submitted for publication.

Chapters 1, 3, 4, 5, and 6 are solely authored, except Section 4.2, which also appears in the submitted paper. Chapter 2 is based on the book written by Wu (2009). All the authors contributed in project formation. Professor Lang Wu contributed to the writing of the paper. Professor Rui Wang and Tingting Yu contributed to manuscript edits. I carried out all the data analyses, simulation studies, and productions of figures and tables, under the supervision of Professor Lang Wu.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Frist of all, I would like to express my deepest thanks and sincere appreciation to my supervisor Dr. Lang Wu, who provided me with his excellent guidance and support throughout my master's program. I am very fortunate to have an adviser who has always been so motivating and patient whenever I needed it most. Without his help and encouragement, I could not have accomplished it.

I would also like to thank Dr. Fidel Vila-Rodriguez for the opportunity to collaborate with a number of excellent students in NINET lab. I would like to thank Estella and Fei for their help with my academic writing skills. Also, I would like to thank my second reader, Dr. Ehsan Karim, for his valuable comments and suggestions on this thesis.

Last but not least, I would like to thank my parents for their unconditional love, support, and encouragement. I also thank my dear friends, Grace, Joy, Zoe, Yibo, Zhan, and Baiji, for making my life so enjoyable and memorable.

To my parents.

Chapter 1

Introduction

1.1 Longitudinal Studies

Longitudinal studies are common in practice, especially in the field of biomedical research. Longitudinal data consists of repeated measurements over time on multiple subjects. For example, in an HIV study, viral loads are repeatedly measured on participating patients over a period of time, and a researcher may be interested in the changes of viral loads over time. Figure 1.1 shows a visualization of longitudinal viral load trajectories. In Figure 1.1, each line represents the viral load trajectory of a patient, and the solid dots on each line represent the repeatedly measured viral loads for this patient. Suppose that these individuals are considered as a random sample from a population. Since the repeated measurements are data collected over time on the same patients, they are likely correlated. For instance, the viral load values from a patient measured at times t_1 and t_2 should be close (i.e., correlated) if t_1 and t_2 are close. Ignoring this correlation may lead to biased results in statistical analysis (Diggle et al., 2002). Therefore, in a longitudinal analysis, it is important to incorporate the correlations between repeated measures.

Sometimes, the values of some variables may be censored, since they are too small or too large to be observed. For example, in an HIV study, the viral loads may be *left-censored* due to a lower detection limit. As shown in Figure 1.1, left-censored values are denoted by triangle dots on the bottom horizontal line with the censored values imputed by the detection limit. Since the (unobserved) true values of the left-censored data are smaller than the detection limit, ignoring the censored data in statistical analyses may lead to biased results (Hughes, 1999). The techniques used to analyze longitudinal data with left censoring are discussed in Section 2.3.

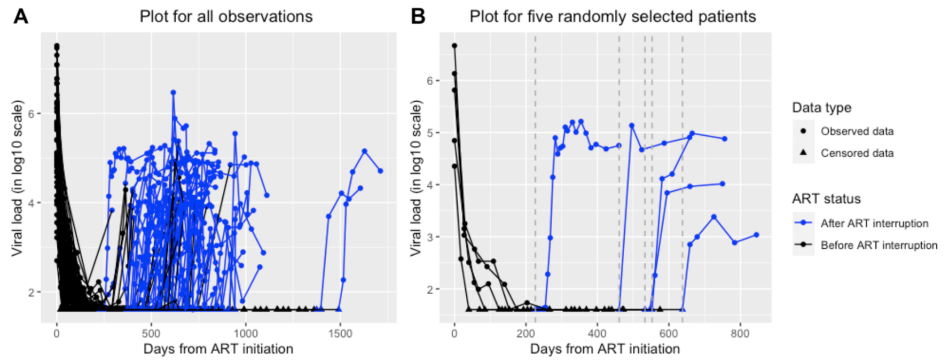

Figure 1.1: Viral load (in $\log_{10}$*-scale) trajectories before and following ART interruption. left-censored values are denoted by triangle dots on the bottom horizontal line with the censored values imputed by the detection limit. Observed values are denoted by solid dots. Data during ART are in black, and data following ART interruption are in blue. The dashed vertical lines in gray indicate times when the ART was interrupted. Figure (A) shows data from all subjects, and Figure (B) shows data from 5 randomly selected subjects.*

1.2 Analysis of Longitudinal Data

Regression models are very useful in longitudinal data analysis, since the covariates in the regression models can partly explain the systematic variabilities in the longitudinal responses. To be more specific, a longitudinal response consists of two types of variations. One is within-individual variation, which reflects variations of the repeated measurements within an individual over time. The other one is between-individual variation, which reflects variations of the response data between individuals at a given time point. That is, the covariates in the regression models may partially explain both types of variations since they can be either time-independent (e.g., gender) or time-dependent (e.g., time). In the longitudinal analysis, the response data are repeated measurements over time on a variable of interest, while the covariates can either be longitudinal (time-dependent) or cross-sectional (time-independent). Three types of regression models are commonly used in longitudinal data analysis, which incorporate the within-individual correlation in different ways. They are mixed effects models, generalized estimating equation (GEE) models, and transitional models.

Mixed effects models assume that the repeated measurements within each individual are correlated since they share some common unobserved

characteristics of the individual. Mixed effects models incorporate these unobserved characteristics by introducing random effects in the models. These random effects not only incorporate the within-individual correlations, but they also reflect individual deviations from population averages. A main advantage of mixed effects models is that they allow for individual-specific inference; that is, model parameters are allowed to be different across individuals. As a result, mixed effects models are preferred if data exhibit large between-individual variations. There are three classes of commonly used mixed effects models for longitudinal data, which are linear mixed effects (LME) models, non-linear mixed effects (NLME) models, and generalized linear mixed models (GLMMs). Some of them are reviewed in Chapter 2. There are other common mixed effects models, such as survival models with random effects (i.e., frailty models), semiparametric or nonparametric mixed effects models. Those models may be studied in future research.

GEE models allow specifying the mean structure and variance-covariance structure separately, without any distributional assumptions. These models are preferred for non-normal data or when the distributional assumptions do not hold. Transitional models assume that the repeated measurements within each individual follow a stochastic process such as a Markov process. Transitional models are preferred if such Markov structure is reasonable, such that the previous response values can be viewed as covariates for the current response value.

Each of the three types of models for longitudinal data has its advantages and limitations. In practice, the choice of the methods for longitudinal data analysis requires both statistical and scientific considerations. This thesis focuses on mixed effects models, especially LME models and NLME models, since the motivating HIV dataset exhibits large between-individual variations.

1.3 Joint Modelling for Longitudinal Data

In practice, two or more longitudinal processes may be associated. For example, in an HIV study, the entire longitudinal process may consist of two separate longitudinal processes: viral decay during antiretroviral therapy (ART) and viral rebound following ART interruption. As shown in Figure 1.1, the HIV viral loads during ART show decreasing trends, which are in black. The viral loads following ART interruption show increasing trends, which are in blue. The dashed vertical lines are used to separate the two longitudinal processes, which indicate times when the ART was interrupted.

We may be interested in studying the association between the key features of viral decay during ART and important characteristics of viral rebound following ART interruption. In this situation, we may need to model both longitudinal processes simultaneously and make full use of the information provided by both processes. Therefore, joint modelling is needed.

When modelling several longitudinal data at the same time, the longitudinal models are often assumed to be linked through shared parameters or random effects, since these random effects reflect individual characteristics of the longitudinal processes. For example, in an HIV study, the viral decay model during ART and the viral rebound model following ART may share the same random effects, which reflect the individual-specific characteristics of the viral decay phase. The individual-specific characteristics of viral decay may be predictive for the individual trajectories of viral rebound, and can be used as "covariates" in the viral rebound model following ART interruption.

We will consider statistical inference methods for joint longitudinal models, such as a naive two-step method and joint likelihood method. These methods are discussed in Chapter 3. Briefly, the naive two-step method works as follows:

- Step 1: fit a model to the observed data in the first longitudinal process, and estimate the random effects;
- Step 2: fit another model to the observed data in the second longitudinal process, substitute the shared random effects in the second model by their estimates from the first step, and then make inference in the second model as if the estimated values were observed values.

Although the naive two-step method is straightforward and easy to implement in statistical software, it may lead to biased results since the uncertainty of estimation in the first step is not incorporated in the second step.

To avoid potential bias in the joint modelling process, we can make statistical inference based on the joint likelihood of all longitudinal data. Maximum likelihood estimates (MLEs) of all parameters are obtained simultaneously by maximizing the joint likelihood. Inference based on the joint likelihood method produces less biased estimates and more reliable standard errors. However, the computation may be time-consuming, since the joint likelihood for longitudinal models is often complicated, and involves high-dimensional and intractable integral due to unobservable random effects and censored data. Linearization methods can be used for approximate inference.

These approximate methods are computationally more efficient, but may offer potential convergence problems. Therefore, we propose *a three-step (TS) method* for joint analysis of two longitudinal models with shared parameters.

The TS method is modified based on the naive two-step method to address its potential problem of biased results. It incorporates the estimation uncertainty in the first step by bootstrap. The bootstrap step works as follows:

- Step 1: generate longitudinal values from the fitted longitudinal models, with unknown parameters substituted by their estimates;
- Step 2: conduct the naive two-step method to fit the generated data from Step 1 and obtain new estimates for parameters of interest;
- Step 3: repeat step 1 and step 2 for B times, and then use the standard deviations of all the new estimates as the standard errors for the corresponding estimates.

The TS method provides more reliable standard errors than the naive two-step method, since it adjusts the standard errors of the estimates by incorporating the estimation uncertainty using bootstrap. The TS method is also easier to implement in statistical software than joint likelihood method.

In this thesis, a comprehensive data analysis on the motivating HIV data is performed using the TS method. The performance of the TS method is also evaluated using simulation studies.

1.4 Literature Review

There has been active research on NLME models and joint models of longitudinal data in recent years. For NLME models, numerical integration methods and MCEM algorithms for likelihood estimation can be computationally expensive and sometimes may exhibit convergence problems, especially when the dimension of random effects is high. Therefore, a computationally more efficient approximation method for NLME models was proposed by Beal and Sheiner (1982), so-called *first-order method*. This method took a Taylor series expansion of the NLME model about the mean of the random effects, i.e., $\mathbf{b}_i = 0$. An improved method, called *first-order conditional method*, was proposed by Lindstrom and Bates (1990), who took a Taylor series expansion about the empirical Bayes estimates of the random effects, i.e., $\mathbf{b}_i = \hat{\mathbf{b}}_i$. The improved procedure has been proved to perform better than that of Beal and Sheiner (1982). The procedures may also be derived

using Laplace approximations (Wolfinger, 1993; Wolfinger and Lin, 1997). Rather than the approximation methods, Delyon et al. (1999) proposed a stochastic approximation version of the EM (SAEM) algorithm for NLME models.

In practice, values of some variables may be censored, since they are too large or too small to be observed. The proportion of censored data may not be small in longitudinal studies, so failure to account for the censoring in the statistical analysis may lead to biased results in the parameter estimates and inference (Hughes, 1999). Hughes (1999) proposed a MCEM algorithm for LME models with censored responses. Fitzgerald et al. (2002) extended Huges' method to NLME models with censored responses. Wu (2002) considered NLME models with both censored responses and covariate measurement errors. Samson et al. (2006) also extended the SAEM algorithm for NLME models with left-censored data. The above methods assume that the censored data follow the same distribution as the observed data. However, such an assumption is not testable based on the observed data. Yu et al. (2018) considered an approach by treating the censored values as point masses.

Statistical inference methods for jointly modelling longitudinal data and survival data have received much attention in the literature. Lawrence Gould et al. (2015) provided a review of this field. For example, Wu et al. (2008) considered an NLME model for the longitudinal process and a Cox proportional hazards model for the time-to-event process, where the individual characteristics of the repeated measures may predict for the time to an event. Rizopoulos et al. (2009) proposed a Laplace approximation approach for joint models of continuous longitudinal response and time-to-event outcome. Although there is a rich literature on joint models of longitudinal data and survival data, there is relatively fewer literature on joint models of two longitudinal processes. To our knowledge, there is no previous research on joint NLME models with left censoring. A comprehensive statistical analysis of HIV study requires us to link two NLME models with censoring, since the two NLME models share some random effects. Therefore, a new statistical method is in demand.

1.5 Outline

In this thesis, we consider two NLME models with left-censored responses: one NLME model for viral dynamics during ART and another NLME model for viral rebounds following ART interruption. We also consider a linear

mixed effects (LME) model for CD4 data. The three models are linked through some shared parameters. We fit the three models separately based on a three-step (TS) method, using the SAEM algorithm. The performance of the TS method is evaluated in simulation studies.

The contributions of this thesis are: (i) to our knowledge, this work is the first to study the relationship between viral decay during ART and viral rebound following ART based on NLME models; (ii) the proposed TS method is simple, and easy to implement in statistical software; (iii) the proposed TS method is based on exact likelihood method, so there is no concern about approximation accuracies as in other approximation methods such as linearization methods, and it is also computationally efficient; and (iv) the proposed TS method performs reasonably well, as shown in simulation studies, and clearly outperforms a common naive two-step method that uses an imputed value for censored observations and model-based standard errors.

This thesis is organized as follows. Chapter 2 reviews mixed effects models for longitudinal data, including LME models, NLME models, and NLME models with left censoring. For inference, we review the Monte Carlo EM (MCEM) algorithm and linearization method, as well as the Stochastic Approximation EM (SAEM) algorithm. Simulation studies are conducted to compare the performances of the linearization method (`nlme` package) and the SAEM algorithm (`saemix` package) in R (R Core Team, 2013; Pinheiro et al., 2019; Comets et al., 2017). In Chapter 3, we describe different joint inference methods, such as a naive two-step method and joint likelihood method, for two NLME models with shared parameters. Since the two inference methods have some limitations, we have proposed a TS method based on the SAEM algorithm for inference. In Chapter 4, real data analysis is conducted on an HIV study by using the naive two-step method and the proposed TS method. In Chapter 5, simulation studies are conducted to compare the performances of the naive two-step method and the proposed TS method under different settings. Conclusions and future research are discussed in Chapter 6.

Chapter 2

NLME Models with left-censored Responses

In this chapter, we briefly review LME models and NLME models, including their expressions and inference. We also discuss the NLME models when there are left-censored responses. Comprehensive descriptions of these models can be found in Wu (2009).

2.1 LME Models

In this section, we first briefly review LME models, and then we extend LME models to NLME models in the next section. Detailed descriptions of these models can be found in Wu (2009), among others.

Suppose there are n individuals in a longitudinal study and n_i is the number of repeated measurements within individual i. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$ be the n_i repeated measurements of the response variable y on individual i, $i = 1, 2, \cdots, n$. A general form of LME models can be written as (Laird and Ware, 1982)

$$\mathbf{y}_i = X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, 2, \cdots, n, \tag{2.1}$$

$$\mathbf{b}_i \sim N(0, D), \quad \mathbf{e}_i \sim N(0, R_i), \tag{2.2}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is a $(p+1) \times 1$ vector of fixed effects, $\mathbf{b}_i = (b_{i0}, b_{i1}, \cdots, b_{iq})^T$ is a $(q+1) \times 1$ vector of random effects for individual i, $\mathbf{e}_i = (e_{i1}, e_{i2}, \cdots, e_{in_i})^T$ are random errors of the repeated measurements within individual i, D is a $(q+1) \times (q+1)$ covariance matrix of the random effects, and R_i is a $n_i \times n_i$ covariance matrix of the within-individual random errors.

The diagonal elements of D are the variances of the random effects $\mathbf{b}_i$, which measure the variability of longitudinal trajectories between individuals. The diagonal elements of R_i are the variances of the random errors e_{ij}'s, which measure the variability of repeated measurements within each individual. The design matrices X_i and Z_i have dimensions $n_i \times (p+1)$ and

$n_i \times (q+1)$ respectively, and Z_i is often a submatrix of X_i. The two design matrices often contain covariates of individual i (i.e., time and age) and can be written as

$$X_i = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_i1} & \cdots & x_{n_ip} \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1q} \\ 1 & z_{21} & \cdots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n_i1} & \cdots & z_{n_iq} \end{pmatrix}.$$

Note that, the repeated measurements $\{y_{i1}, y_{i2}, \cdots, y_{in_i}\}$ within each individual can be taken at different time points t_{ij} for different individuals, and the number of repeated measurements n_i may also vary across individuals. This is an advantage of LME models, such that they allow unbalanced data in the response.

There are standard assumptions for LME models (2.1) and (2.2), including: (i) the individuals are independent, and (ii) the random errors $\mathbf{e}_i$ and the random effects $\mathbf{b}_i$ are independent and both are normally distributed with mean zero. Under the above assumptions, the marginal distribution of the response $\mathbf{y}_i$ is a multivariate normal distribution, which can be written as

$$\mathbf{y}_i \sim N(X_i\beta, Z_iDZ_i^T + R_i). \tag{2.3}$$

The marginal mean of the response $\mathbf{y_i}$ is given by $E(\mathbf{y}_i) = X_i\beta$, and the variance covariance matrix of the response $\mathbf{y_i}$ is given by

$$Cov(\mathbf{y}_i) = Z_iDZ_i^T + R_i = V_i,$$

where V_i is an $n_i \times n_i$ square matrix. Note that, the marginal distribution (2.3) provides population average inference, while the individual-specific inference can be conducted by conditioning on the random effects $\mathbf{b}_i$.

The covariance R_i is often assumed to depend on i only through their dimensions and the within-individual repeated measurements are conditionally independent given the random effects, i.e., it is often assumed that $R_i = \sigma^2 I_{n_i}$, where I_{n_i} is a $n_i \times n_i$ identity matrix. This assumption may be reasonable if the within-individual repeated measurements are far apart or the between-individual variation dominates the within-individual variation. The simplified within-individual covariance structure greatly reduces the number of parameters and may reduce some identifiability problems.

Let's consider the motivating HIV dataset and fit an LME model with first four viral load measurements as responses and time as predictors. The

LME model with random intercept and random slope can be written as

$$\mathbf{y}_i = X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, 2, \cdots, n, \tag{2.4}$$

where

$$X_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix},$$

$\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ contains fixed effect parameters, $\mathbf{b}_i = (b_{0i}, b_{1i})^T$ contains random effects for individual i and follows $N(0, D)$ with D being a 2×2 covariance matrix, and $\mathbf{e}_i$ is within-individual error. Assuming that the random effects $\mathbf{b}_i$ and the random error $\mathbf{e}_i$ are independent, we have $\mathbf{e}_i \sim N(0, \sigma^2 I_{n_i})$. The LME model can also be written as

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + e_{ij}. \tag{2.5}$$

Figure 2.1 shows the observed and fitted viral load trajectories before ART interruption based on the LME model (2.4) for four randomly selected patients. We can see that slope and intercept of the fitted line vary from patient to patient.

Statistical inference for an LME model is typically based on the maximum likelihood method. Let $\boldsymbol{\theta}$ denote all parameters in the LME model (2.1) and (2.2). The likelihood for the observed response $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ is given by

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\theta) \tag{2.6}$$

$$= \prod_{i=1}^{n} \int f(\mathbf{y}_i|\mathbf{b}_i, \beta, R_i) f(\mathbf{b}_i|D) d\mathbf{b}_i, \tag{2.7}$$

where

$$f(\mathbf{y}_i|\mathbf{b}_i, \beta, R_i) = (2\pi)^{-n_i/2}|R_i|^{-1/2} \exp[-(\mathbf{y}_i - X_i\beta - Z_i\mathbf{b}_i)^T R_i^{-1} \times (\mathbf{y}_i - X_i\beta - Z_i\mathbf{b}_i)],$$

$$f(\mathbf{b}_i|D) = (2\pi)^{-q/2}|D|^{-1/2} \exp(-\mathbf{b}_i^T D^{-1}\mathbf{b}_i).$$

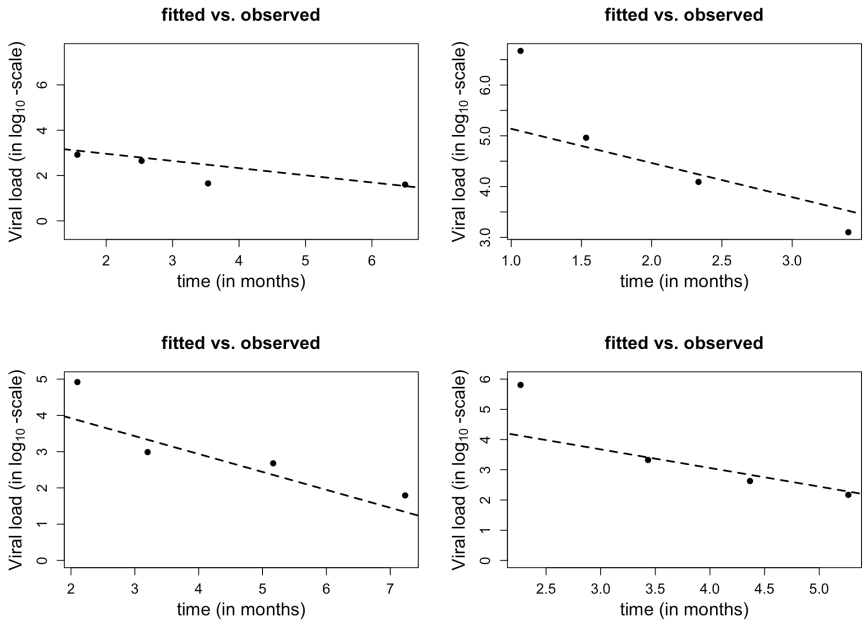

Figure 2.1: Observed and fitted viral load trajectories before ART interruption based on LME model (2.4) for four randomly selected patients.

In the case $R_i = \sigma^2 I$, assuming V_i is known, the maximum likelihood estimates (MLEs) of the fixed effects β and σ^2 are

$$\begin{aligned} \hat{\beta} &= (\sum_{i=1}^{n} X_i^T V_i^T X_i)^{-1} \sum_{i=1}^{n} X_i^T V_i^{-1} \mathbf{y}_i, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - X_i \hat{\beta})^T V_i^{-1} (\mathbf{y}_i - X_i \hat{\beta}). \end{aligned}$$

The MLEs of θ can be obtained by using the expectation-maximization (EM) algorithm. The EM algorithm is a popular iterative method for computing MLEs with incomplete data (Dempster et al., 1977). The EM algorithm alternates between performing an expectation step (E-step) and a maximization step (M-step). The E-step computes the conditional expectation of the "complete-data" log-likelihood evaluated based on the observed data and current estimates for the parameters, and the M-step computes parameters maximizing the expected log-likelihood in the E-step. These parameter estimates are then used to compute the expectation of the log-likelihood in the next E-step. Given starting values of θ, one iterates be-

tween the E-step and the M-step until convergence. At convergence, we obtain (possibly local) maximizers of the observed-data likelihood.

To find MLEs of θ in LME model (2.1) and (2.2), the EM algorithm is used by treating the random effects $\mathbf{b}_i$ as "missing data". Let k index the iterations, where $k = 0$ refers to starting values and $k = \infty$ refers to convergence. Define $Q(\theta|\theta^{(k)})$ as the expected log-likelihood function of θ given $\theta^{(k)}$, where $\theta^{(k)}$ denotes the parameter estimates from the k-th iteration. Then, the E-step of the EM algorithm at iteration $k+1$ computes

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E\left[\log L(\theta|\mathbf{y},\mathbf{b})|\mathbf{y},\theta^{(k)}\right] \\ &= E\left[\sum_{i=1}^{n}\left\{(\log f(\mathbf{y}_i|\mathbf{b}_i,\beta,R_i)+\log f(\mathbf{b}_i|D))|\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\}\right]. \end{aligned}$$

Let $\boldsymbol{\eta}$ denote the vector of all distinct parameters in the variance-covariance matrices D and R_i. Then the estimated sufficient statistics of $\boldsymbol{\eta}$ at iteration k are:

$$\begin{aligned} \sum_{i=1}^{n} E\left(\mathbf{e}_i^T\mathbf{e}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)}\right) &= \sum_{i=1}^{n}\left[\hat{\mathbf{e}}_i^{(k)T}\hat{\mathbf{e}}_i^{(k)}+\mathrm{tr}\left(\mathrm{Cov}(\mathbf{e}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)})\right)\right], \\ \sum_{i=1}^{n} E\left(\mathbf{b}_i^T\mathbf{b}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)}\right) &= \sum_{i=1}^{n}\left[\hat{\mathbf{b}}_i^{(k)}\hat{\mathbf{b}}_i^{(k)T}+\mathrm{Cov}(\mathbf{b}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)})\right], \end{aligned}$$

where

$$\begin{aligned} \hat{\mathbf{e}}_i^{(k)} &= \mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}^{(k)} - Z_i\hat{\mathbf{b}}_i^{(k)}, \\ \hat{\mathbf{b}}_i^{(k)} &= D(\hat{\boldsymbol{\eta}}^{(k)})Z_i^T V_i^{-1}(\hat{\boldsymbol{\eta}}^{(k)})\left(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}^{(k)}\right), \\ V_i(\hat{\boldsymbol{\eta}}^{(k)}) &= Z_i D(\hat{\boldsymbol{\eta}}^{(k)})Z_i^T + R_i(\hat{\boldsymbol{\eta}}^{(k)}). \end{aligned}$$

After E-step, we can obtain an updated parameter estimate $\boldsymbol{\theta}^{(k+1)}$ by maximizing $Q(\theta|\theta^{(k)})$ in the M-step, i.e.,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(k+1)} &= \left[\sum_{i=1}^{n} X_i^T\hat{V}_i^{-1}(\hat{\boldsymbol{\eta}}^{(k)})X_i\right]^{-1}\sum_{i=1}^{n} X_i^T\hat{V}_i^{-1}(\hat{\boldsymbol{\eta}}^{(k)})\mathbf{y}_i, \\ \hat{\sigma}^{(k+1)2} &= \sum_{i=1}^{n} E\left(\mathbf{e}_i^T\mathbf{e}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)}\right)/\sum_{i=1}^{n} n_i, \\ \hat{D}^{(k+1)} &= \sum_{i=1}^{n} E\left(\mathbf{b}_i^T\mathbf{b}_i|\mathbf{y}_i,\hat{\boldsymbol{\theta}}^{(k)}\right)/n. \end{aligned}$$

Alternating the above E-step and M-step until convergence, we obtain the final parameter estimates that maximize the observed-data likelihood (possibly locally), depending on the starting values. EM algorithm performs well when finding the MLEs of $\boldsymbol{\theta}$ for an LME model, since we can obtain closed-form expressions of the E-step and M-step by integrating out the unobserved random effects in the E-step.

2.2 NLME Models

In the previous section, we describe a widely used LME model in longitudinal studies. In general, linear regression models have the advantage of simplicity because they reasonably fit the observed data without understanding of the true relationship between the response and the covariates (i.e., also known as the data-generation mechanism). In order to understand the data-generation mechanism, mechanistic models such as nonlinear regression models, are suggested. Compared to linear models, nonlinear models have some advantages, including: (i) nonlinear models may be able to make better predictions outside the range of the observed data, since they usually have a better understanding of the data-generation mechanism; (ii) the parameters in nonlinear models often have natural scientific interpretations; and (iii) non-linear models often need fewer parameters than linear models to fit the observed data equally well.

In a longitudinal study, random effects can be introduced in the nonlinear models to incorporate the within-individual correlations and between-individual variations. The extended models are called nonlinear mixed effects (NLME) models. Suppose there are n individuals and n_i is the number of repeated measurements within individual i. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$ be the n_i repeated measurements of the response variable y on individual i, $i = 1, 2, \cdots, n$. A general form of NLME models can be written as (Lindstrom and Bates, 1990)

$$
\begin{aligned}
y_{ij} &= g(t_{ij}, \beta_i) + e_{ij}, && (2.8)\\
\beta_i &= h(\mathbf{x}_i, \beta, \mathbf{b}_i), \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i && (2.9)\\
\mathbf{b}_i &\sim N(0, D), \quad \mathbf{e}_i \sim N(0, R_i), && (2.10)
\end{aligned}
$$

where $g(\cdot)$ is a known nonlinear function governing within-individual behavior depending on a $(p \times 1)$ vector of individual-specific parameter β_i, $h(\cdot)$ is a p-dimensional function depending on an $(r \times 1)$ vector of fixed effects β, a $(k \times 1)$ vector of random effects $\mathbf{b}_i$, and a vector of covariates $\mathbf{x}_i$ specific to individual i, $\mathbf{e}_i = (e_{i1}, e_{i2}, \cdots, e_{in_i})^T$ are random errors of within-individual

measurements, D is a covariance matrix of the random effects, and R_i is a covariance matrix of the within-individual random errors. We assume that $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent. Similar as LME models, if the within-individual repeated measurements are far apart or the between-individual variation dominates the within-individual variation, we can assume that $R_i = \sigma^2 I_{n_i}$.

The function $h(\cdot)$ in model (2.9) is often a linear function, as a result, the model can be written more compactly in the form

$$\boldsymbol{\beta}_i = A_i\boldsymbol{\beta} + B_i\mathbf{b}_i, \tag{2.11}$$

where A_i is a design matrix depending on $\mathbf{x}_i$, and B_i is a design matrix typically involving only 0's and 1's depending on the random effects $\mathbf{b}_i$.

Let's consider the motivating HIV dataset in Chapter 1 and fit an NLME model (2.12) for viral decay before ART interruption. The NLME model (2.12) is introduced by Wu and Ding (1999) and it is a widely used viral dynamic model in HIV studies. The model can be written as:

$$\begin{aligned} y_{ij} &= g(t_{ij}, \beta_i) + e_{ij}, \\ &= \log_{10}(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}) + e_{ij}, \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \\ i &= 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i, \end{aligned} \tag{2.12}$$

where $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T \sim N(0, D)$ and $e_{ij} \sim N(0, \sigma^2)$. The model may be represented as in model (2.20) with $\boldsymbol{\beta} = (P_1, P_2, \lambda_1, \lambda_2)^T$, and

$$A_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad B_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Figure 2.2 shows the observed and fitted viral load trajectories before ART interruption based on NLME model (2.12) for four randomly selected patients. Compare to Figure 2.1, we can see that the NLME model fits the observed data better.

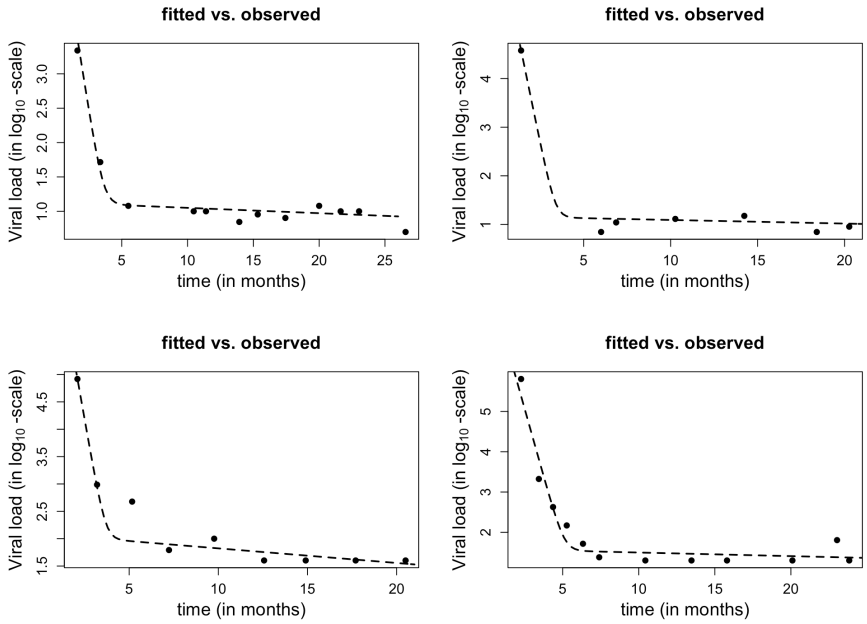

Figure 2.2: Observed and fitted viral load trajectories before ART interruption based on NLME model (2.12) for four randomly selected patients.

There are a number of inferential methods for the NLME model. We first introduce the likelihood method. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, D)$ denote all parameters in the NLME model (2.8), (2.9) and (2.10), where $\boldsymbol{\eta}$ is a vector of distinct parameters in R_i. The marginal distribution of the response $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \int f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{b}_i) f(\mathbf{b}_i|D) d\mathbf{b}_i, \tag{2.13}$$

and the likelihood can be written as

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \int f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{b}_i) f(\mathbf{b}_i|D) d\mathbf{b}_i. \tag{2.14}$$

Since the NLME model is nonlinear in the random effects $\mathbf{b}_i$, the integrations in the likelihood (2.14) cannot be done in a closed form. Thus, the evaluation of the integral can be computationally expensive, which leads to a major difficulty of likelihood inference for an NLME model. Three commonly used methods are numerical or Monte Carlo integration methods, EM algorithms, and approximate methods (Davidian and Giltinan, 1995).

An iterative EM algorithm can be used to compute the maximum likelihood estimates. Let's denote $\{\mathbf{y}_i, i = 1, \cdots, n\}$ as the observed data. By

treating the random effects $\mathbf{b}_i$ as missing data, we have "complete data" $\{(\mathbf{y}_i, \mathbf{b}_i), i = 1, \cdots, n\}$. The "complete-data" log-likelihood for individual i is given by

$$l_i^{(c)}(\boldsymbol{\theta}|\mathbf{y}_i, \mathbf{b}_i) = \log f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\eta}) + \log f(\mathbf{b}_i|D).$$

At iteration $k = 1, 2, 3, \cdots$, the E-step computes the conditional expectation

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E\left[\sum_{i=1}^{n} l_i^{(c)}(\theta|\mathbf{y}_i, \mathbf{b}_i)|\mathbf{y}_i, \theta^{(k)}\right] \\ &= E\left[\sum_{i=1}^{n}\left\{(\log f(\mathbf{y}_i|\mathbf{b}_i, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i|D))|\mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right\}\right], \end{aligned}$$

with respect to the conditional distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)})$, where $\boldsymbol{\theta}^{(0)}$ is the starting value. Since the conditional expectation $Q(\theta|\theta^{(k)})$ does not have a closed-form expression, we may consider Monte Carlo simulations or numerical integration methods to evaluate $Q(\theta|\theta^{(k)})$ when the dimension of the random effects $\mathbf{b}_i$ is low. When the E-step of the EM algorithm is evaluated using Monte Carlo simulations, the EM algorithm is called a Monte Carlo EM algorithm (MCEM).

To perform MCEM algorithm, we need to simulate a large sample of the missing data $\mathbf{b}_i$ from the conditional distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)})$ at k-th EM iteration. This sampling can be accomplished using Markov Chain Monte Carlo (MCMC) method such as the Gibbs sampler or a rejection sampling method. The Gibbs sampling is proposed by Geman and Geman (1993) and it is used to obtain random samples from an intractable multidimensional probability distribution by sequentially sampling from lower-dimensional conditional distributions which are easier to sample from. These samples then comprise a Markov chain, whose stationary distribution is the target distribution. The details of the Gibbs sampler are described as follows.

Let $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T, \cdots, \mathbf{u}_q^T)^T$ be a random vector, and each component $\mathbf{u}_i$ may also be a random vector. Note that the component $\mathbf{u}_i$ are usually unobserved values with different dimensions. Suppose that we wish to generate samples from the probability distribution $f(\mathbf{u}|\boldsymbol{\theta})$ and $f(\mathbf{u}|\boldsymbol{\theta})$ is highly intractable. Therefore, we are not able to generate a sample from $f(\mathbf{u}|\boldsymbol{\theta})$ directly, and the Gibbs sampler is introduced. For simplicity, assume that $\boldsymbol{\theta}$ is known. Let

$$\mathbf{u}_{-j} = (\mathbf{u}_1^T, \cdots, \mathbf{u}_{j-1}^T, \mathbf{u}_{j+1}^T, \cdots, \mathbf{u}_q^T)^T, \quad j = 1, 2, \cdots, q,$$

be the subvector of $\mathbf{u}$ without the component $\mathbf{u}_j$. Beginning with the starting values $(\mathbf{u}_1^{(0)}, \cdots, \mathbf{u}_q^{(0)})$, at step $k = 1, 2, \cdots$,

- sample $\mathbf{u}_1^{(k)}$ from $f(\mathbf{u}_1|(\mathbf{u}_2^{(k-1)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- sample $\mathbf{u}_2^{(k)}$ from $f(\mathbf{u}_2|(\mathbf{u}_1^{(k-1)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- $\cdots$;
- sample $\mathbf{u}_q^{(k)}$ from $f(\mathbf{u}_q|(\mathbf{u}_1^{(k-1)}, \mathbf{u}_2^{(k-1)}, \cdots, \mathbf{u}_{q-1}^{(k-1)}, \boldsymbol{\theta})$.

The sampling procedure works well, because generating samples from the lower dimensional conditional distribution $f(\mathbf{u}_j|\mathbf{u}_{j-1}, \boldsymbol{\theta})$, $j = 1, 2, \cdots, q$, is much easier. The sequence $\{(\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)}), k = 1, 2, \cdots\}$ comprises a Markov chain with stationary distribution $f(\mathbf{u}|\boldsymbol{\theta})$. Hence, when k is large, $\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)})^T$ can be viewed as a sample generated from the probability distribution $f(\mathbf{u}|\boldsymbol{\theta})$.

In the E-step of the MCEM algorithm, the Gibbs sampler works when simulate the missing data $\mathbf{b}_i$ from the conditional distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)})$ since

$$f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)}) \propto f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\theta}^{(k)}) f(\mathbf{b}_i|D^{(k)}),$$

where $f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\theta}^{(k)})$ and $f(\mathbf{b}_i|D^{(k)})$ are known distributions. Denote the simulated data as $\{\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(2)}, \cdots, \mathbf{b}_i^{(M)}\}$, where M is the number of Monte Carlo sample. Then, the conditional expectation $Q(\theta|\theta^{(k)})$ can be approximated by the following empirical mean:

$$\tilde{Q}(\theta|\theta^{(k)}) = \frac{1}{M} \sum_{j=1}^{M} \left[\log f(\mathbf{y}_i|\mathbf{b}_i^{(j)}, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i^{(j)}|D) \right].$$

The M-step is then to maximize $\tilde{Q}(\theta|\theta^{(k)})$ to produce updated $\theta^{(k+1)}$ by using standard optimization procedure such as the Newton-Raphson method.

The foregoing Monte Carlo EM algorithm can be arbitrarily accurate by increasing M, but the computational time also grows rapidly as M increases. Alternatively, we can use numerical integration, such as the Gauss-Hermite quadrature. Compared to Monte Carlo integration, numerical integration may reduce computational time without loss of accuracy, especially when the random effects have a low dimension and follow a normal distribution. However, the computational time grows exponentially with the dimension

of the random effects. More details can be found in Davidian and Gallant (1993).

The approximation method is another approach for inference for NLME models. Laplace approximation can be used to directly approximate the likelihood (2.14). In addition, Taylor expansions can be used to linearize the NLME model and iteratively solve the resulting LME models (Lindstrom and Bates, 1990). The linearization method is widely used in standard software such as the `nlme` package in *R*. The idea is to take a first-order Taylor expansion about estimates of parameters and random effects, which leads to a "working" LME model. Then, we update parameter estimates from this LME model iteratively until they converge. To be more specific, the NLME model (2.8) and (2.9) can be written as a single equation

$$y_{ij} = u_{ij}(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{b}_i) + e_{ij}, \quad i = 1, \cdots, n; j = 1, \cdots, n_i,$$

where $u_{ij}(\cdot)$ is a nonlinear function. Let $\mathbf{u}_i = (u_{i1}, \cdots, u_{in_i})^T$. At each iteration, denote the current estimates of $(\boldsymbol{\beta}, \mathbf{b}_i)$ by $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i)$, suppressing the iteration number, where $\widehat{\mathbf{b}}_i$ is the empirical Bayesian estimate of $\mathbf{b}_i$. Then, iteratively solving the following "working" LME model (Wolfinger, 1993)

$$\widetilde{\mathbf{y}}_i = W_i\boldsymbol{\beta} + T_i\mathbf{b}_i + \mathbf{e}_i, \tag{2.15}$$

where

$$\begin{aligned} \widetilde{\mathbf{y}}_i &= \mathbf{y}_i - \mathbf{u}_i(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i) + W_i\widehat{\boldsymbol{\beta}} + T_i\widehat{\mathbf{b}}_i, \\ W_i &= \frac{\partial \mathbf{u}_i(\mathbf{x}_i, \boldsymbol{\beta}, \widehat{\mathbf{b}}_i)}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \quad T_i = \frac{\partial \mathbf{u}_i(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \mathbf{b}_i)}{\partial \mathbf{b}_i^T}\Big|_{\mathbf{b}=\widehat{\mathbf{b}}_i}. \end{aligned}$$

At each iteration, the parameters and random effects from the LME model (2.15) are updated by $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i)$ using standard methods described in Section 2.1. The advantage is that the approximation methods are computationally efficient since they avoid intractable integration. A drawback of the approximation method is that the approximation may not be accurate if the model is very "nonlinear". Another drawback is that there may arise some convergence issues.

Other than the likelihood method and the linearization method, inference for an NLME model also can be based on a two-step method if the number of repeated measurements n_i's are large. In step 1, individual parameter $\boldsymbol{\beta}_i$ are estimated by fitting a nonlinear regression model to the repeated observations within each individual using standard estimation methods for nonlinear models such as the least square method. In step 2, the individual

estimates $\hat{\boldsymbol{\beta}}_i$ are used to estimate the fixed parameters $\boldsymbol{\beta}$ and perform inference based on large-sample asymptotic results. More details can be found in Davidian and Giltinan (1995). The two-step method is simple and requires no distributional assumptions, but it requires a large number of repeated measurements.

2.3 NLME models with left censoring

In practice, the values of some variables may be viewed as censored since their values are too large or too small to observe. For example, in the motivating HIV dataset, some viral loads are (left) censored due to a low detection limit. That is, if the viral loads are lower than the detection limit, these values are unobservable. Figure 1.1 shows entire viral load trajectories during ART and following ART interruption for all subjects and for 5 randomly selected subjects, respectively (for data following ART interruption, we only show the first 36 weeks of data because viral load levels typically stabilize before then). left-censored values are denoted by triangle dots on the bottom horizontal line with the censored values imputed by the detection limit. Observed values are denoted by circle dots. Data during ART are in black, and data following ART interruption are in blue. The dashed vertical lines in gray indicate times when the ART was interrupted. In this section, we mainly focus on NLME models with left censoring.

When data are censored, their true values are unknown but only known to be lower or higher than some thresholds. In the HIV dataset, all the censored viral loads are known to be between 0 and the detection limit, that is, between 0 and $\log_{10} 40$. It is important to take into consideration of censored value in the analysis, since the proportion of censored data is high. Failure to account for the censored data in the statistical analysis may lead to significant biased results (Hughes, 1999).

Paxton et al. (1997) introduced a naive method to address left censoring problem in HIV studies, which imputes the censored values by half the detection limit. A major drawback of this naive method is that this imputation method ignores the uncertainty of the censored values, which may lead to biased results. To better address censoring, we can consider an EM algorithm for likelihood estimation.

Let y_{ij} be the response for individual i at time t_{ij}, $i = 1, \cdots, n$; $j = 1, \cdots, n_i$. Since some observations are censored, we can write y_{ij} as (q_{ij}, c_{ij}),

where q_{ij} is the observed value and c_{ij} is the censoring indicator such that

$$y_{ij} = \begin{cases} q_{ij} & \text{if } c_{ij} = 0, \\ \leq d & \text{if } c_{ij} = 1, \end{cases}$$

where d is a known constant such as a detection limit in an HIV study. Denote $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^T$, $\mathbf{q}_i = (q_{i1}, \cdots, q_{in_i})^T$, and $\mathbf{c}_i = (c_{i1}, \cdots, c_{in_i})^T$. The observed data are $\{(\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i), i = 1, \cdots, n\}$, where $\mathbf{z}_i$ is a collection of covariates.

Let $f(\cdot)$ and $F(\cdot)$ denote a generic density function and the corresponding cumulative density function (cdf), respectively. Assume that $y_{i1}, \cdots, y_{in_i}$ are conditionally independent given the random effects $\mathbf{b}_i$, and $\mathbf{b}_i \sim N(0, D)$, where D is an unknown covariance matrix. Let $\boldsymbol{\theta}$ be the collection of all unknown parameters. The likelihood for the observed data $\{(\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i), i = 1, \cdots, n\}$ can be written as

$$L_o(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int \{\prod_{j=1}^{n_i} (f(y_{ij}|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}))^{1-c_{ij}} (F(d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}))^{c_{ij}}\} \times f(\mathbf{b}_i|B) d\mathbf{b}_i, \tag{2.16}$$

where

$$F(d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\sigma}) \equiv P(Y_{ij} < d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\sigma}),$$

$\boldsymbol{\beta}$ contains mean parameter, $\boldsymbol{\sigma}$ contains variance-covariance parameters, and Y_{ij} is the random version of y_{ij}. The likelihood (2.16) does not have an analytic expression. Hughes (1999) proposed a Monte Carlo EM algorithm to find the MLEs of the parameters $\boldsymbol{\theta}$ for LME models with censored responses. The same method can also be extended to NLME models with censored responses (Wu, 2002).

Treating the censored values in $\mathbf{y}_i$ and the random effects $\mathbf{b}_i$ as missing data, we have "complete data" $\{(\mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i), i = 1, 2, \cdots, n\}$. The log-likelihood of the complete data can be written as

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_c^{(i)}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [\log f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}) + \log f(\mathbf{b}_i|\boldsymbol{\theta})].$$

The E-step is to compute the conditional expectation of the complete data log-likelihood given the observed data and current parameter estimates. The

conditional expectation for the i-th observation at the $(k+1)$-th EM iteration can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\{l_c(\boldsymbol{\theta})|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}\} \\ &= \int\int [\log f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}^{(k)}) + \log f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)})] \\ &\quad \times f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) d\mathbf{y}_{cen,i} d\mathbf{b}_i \end{aligned} \tag{2.17}$$

where $\mathbf{y}_{cen,i}$ is a vector of censored responses.

Since the density function $f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}^{(k)})$ is nonlinear in the random effects $\mathbf{b}_i$, the random effects $\mathbf{b}_i$ cannot be integrate out in the conditional expectation (2.17), so we are unable to obtain a closed-form expression for $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. To implement this E-step, we can use Monte Carlo methods to simulate the missing data $(\mathbf{y}_{cen,i}, \mathbf{b}_i)$ from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, and approximate the conditional expectation $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ by an empirical mean based on the simulated missing data. In other words, the censored values are assumed to follow the same distribution as the distribution assumed for the observed data. The simulation step can be done by using the Gibbs sampler along with rejection sampling methods.

As discussed in Section 2.2, the key idea of the Gibbs sampler is to generate samples from the target high dimensional distribution by sequentially sampling from lower dimensional conditional distributions. In general, it is easier to sample from these lower dimensional conditional distributions than the target high dimensional distribution. However, sometimes, sampling from the lower dimensional conditional distributions may not easy either. In this case, we may need to combine the Gibbs sampler with rejection sampling methods to sample from these conditional distributions.

Suppose that we wish to generate a sample from $f(x)$, but $f(x)$ is too complicated to sample from directly. However, suppose that we know how to sample from the density distribution $h(x)$, and there is a known constant c, such that

$$f(x) \leq ch(x), \quad \text{for all } x.$$

Then, a rejection sampling works as follows:

- generate a value x^* from the density distribution $h(x)$;
- generate a value u from the uniform distribution on $(0, 1)$;
- accept x^*, if $u < \dfrac{f(x^*)}{ch(x^*)}$,
  reject x^*, otherwise.

Repeating the above steps and only retaining the accepted x^*. Then, x_1^*, x_2^*, $\cdots$, is a sample from the target distribution $f(x)$.

In the E-step of the MCEM algorithm for NLME models with censored responses, we can generate samples from $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$ by iteratively sampling from the conditional distributions $f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$ and $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$ based on the Gibbs sampler. Note that

$$f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) \propto f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) \times f(\mathbf{c}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}), \quad (2.18)$$

$$f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) \propto f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)}) \times f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(k)}), \quad (2.19)$$

so we only need to generate samples from the right-hand sides of (2.18) and (2.19), which can be accomplished using rejection sampling methods since the density functions on the right-hand sides of (2.18) and (2.19) are known. The resulting samples constitute a Markov chain which will converge to a stationary distribution $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$. Repeating this process many times, we obtain many independent samples $(\mathbf{y}_i, \mathbf{b}_i)$, say $\{(\mathbf{y}_i^{(1)}, \mathbf{b}_i^{(1)}), (\mathbf{y}_i^{(2)}, \mathbf{b}_i^{(2)}), \cdots, (\mathbf{y}_i^{(M)}, \mathbf{b}_i^{(M)})\}$, from $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$. Then, the conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ in the E-step can be approximated by its empirical mean

$$\tilde{Q}(\theta|\theta^{(k)}) = \frac{1}{M}\sum_{j=1}^{M}\left[\log f(\mathbf{y}_i^{(j)}|\mathbf{z}_i, \mathbf{b}_i^{(j)}, \boldsymbol{\theta}^{(k)}) + \log f(\mathbf{b}_i^{(j)}|\boldsymbol{\theta}^{(k)})\right],$$

with the unobserved $(\mathbf{y}_{cen,i}, \mathbf{b}_i)$ substituted by their simulated values. The M-step is then to maximize the approximated conditional expectation in order to obtain the updated parameters. The M-step is like a complete-data maximization, so standard complete-data optimization procedures may be used to update the parameter estimates, such as the Newton-Raphson method. Iterating between the E-step and M-step until convergence, we obtain an MLE of $\boldsymbol{\theta}$ or a local maximum of the observed-data likelihood.

Since the parameters and the random effects enter the NLME models in a nonlinear fashion, there are two major difficulties arise for the Monte Carlo EM algorithm. Firstly, the E-step becomes very complicated since the random effects cannot be integrated out. Secondly, analytic expressions for the E-step and the M-step are no longer available, so iterative algorithms are needed. Since the exact likelihood estimation for NLME models is computationally expensive, approximate methods based on Taylor or Laplace approximations can be used (Lindstrom and Bates, 1990). The linearization method is computationally efficient, but its performance may be less satisfactory in some cases. Multiple imputation methods may also be used when

analyzing censored longitudinal data. For a multiple imputation method, the choice of imputation models only affects the imputed values, so if the proportion of censored values is small, a multiple imputation method should perform well.

For parameter estimation and inference for NLME models, an MCEM algorithm may be time-consuming while the linearization method may be inaccurate and may have convergence issues. Hence, a stochastic approximation expectation-maximization (SAEM) algorithm is proposed by Delyon et al. (1999). The SAEM algorithm is computationally more efficient than the MCEM method, since in the E-step, only one value is simulated instead of many values as in MCEM method. More details about SAEM algorithm will be discussed in the next section.

2.4 Review of SAEM algorithm

For NLME models, an MCEM algorithm may be computationally intensive, while the linearization method may not be accurate and may have convergence issues. Therefore, Delyon et al. (1999) proposed a stochastic approximation version of the EM algorithm (SAEM) for NLME models with no censored values. As shown in Delyon et al. (1999), the SAEM algorithm is computationally more efficient than an MCEM algorithm and it converges to a (local) maximum of the likelihood theoretically. The SAEM algorithm has also proven to perform better than the linearization method (Girard and Mentré, 2005). It has been implemented in the *Monolix* software (Lavielle, 2014), and also in the *R* software through the `saemix` package (Comets et al., 2017). Samson et al. (2006) has extended the SAEM algorithm to estimate parameters of NLME models with left-censored data. The SAEM algorithm and the extended SAEM algorithm will be reviewed in this section.

As reviewed in Section 2.2, the linearization and EM algorithm are two commonly used inferential methods for NLME models. The linearization method is to obtain an approximation of the likelihood, which is maximized through Newton-Raphson minimization. There are different approximations to the likelihood. Lindstrom and Bates (1990) proposed a first-order conditional method, which uses Taylor expansions to linearize the NLME model and iteratively solve the resulting LME models. This approximation method is implemented in statistical software *R*, where the `nlme` package is widely used in NLME modelling. However, the linearization methods have notable shortcomings, such as convergence issues and possibly limited with complex models. An alternative to the linearization method is the EM algorithm,

which is developed based on missing data (Dempster et al., 1977).

Recall that the EM algorithm computes the maximum likelihood estimates in two steps. At the k-th iteration, the E-step is the evaluation of the conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, where $\boldsymbol{\theta}$ is a vector of all parameters in the model. The M-step then updates $\boldsymbol{\theta}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. For NLME models, we need to simulate a large sample of the missing data $\mathbf{b}_i$ at the k-th EM iteration. The simulation of the missing data $\mathbf{b}_i$ can be accomplished using MCMC method, such as the Gibbs sampler or a rejection sampling method. This EM algorithm is also known as MCEM algorithm. Since the MCEM algorithm requires simulating a large number of the missing data, the computation may be expensive. The SAEM algorithm, as an alternative to the MCEM algorithm, combines a stochastic approximation with an EM algorithm. It makes use of a stochastic approximation procedure for estimating the conditional expectation of the complete data log-likelihood. Unlike the MCEM algorithm, in the E-step, the SAEM algorithm only simulates random effects once, rather than many times, which greatly reduces computational costs. Delyon et al. (1999) proves that the SAEM algorithm is very efficient for NLME models, and also the convergence of the SAEM algorithm under general conditions if the complete data log-likelihood belongs to the regular curved exponential family.

Let $\{\mathbf{y}_i, i = 1, \cdots, n\}$ denote the observed data and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, D)$ denote all parameters in the NLME model

$$\begin{aligned} y_{ij} &= g(t_{ij}, \beta_i) + e_{ij}, \\ \beta_i &= h(\mathbf{x}_i, \beta, \mathbf{b}_i), \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i \\ \mathbf{b}_i &\sim N(0, D), \quad \mathbf{e}_i \sim N(0, R_i), \end{aligned}$$

where $\boldsymbol{\eta}$ is a vector of distinct parameters in R_i. Recall that the marginal distribution of the response $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \int f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{b}_i) f(\mathbf{b}_i|D) d\mathbf{b}_i,$$

and the likelihood can be written as

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \int f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{b}_i) f(\mathbf{b}_i|D) d\mathbf{b}_i,$$

which does not have an analytical expression. By treating the random effects $\mathbf{b}_i$ as missing data, the complete-data log-likelihood for individual i can be

written as

$$l_i^{(c)}(\boldsymbol{\theta}|\mathbf{y}_i, \mathbf{b}_i) = \log f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\eta}) + \log f(\mathbf{b}_i|D).$$

At iteration k, the SAEM algorithm proceeds as follows. The E-step is divided into a simulation step (S step) and a stochastic approximation step (SA step). In the S step, $\{\mathbf{b}_i^{(1)}, \cdots, \mathbf{b}_i^{(m_k)}\}$ are simulated from the conditional distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)})$, then in the SA step, $Q_k(\theta)$ is updated based on

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\frac{1}{m_k}\sum_{j=1}^{m_k} \log f(\mathbf{y}_i|\mathbf{b}_i^{(j)}, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i^{(j)}|D) - Q_{k-1}(\theta)),$$

where $\{\gamma_k\}_{k\geq 1}$ is a sequence of positive step size. The choice of γ_k will be discussed later. The M-step is then to maximize $Q_k(\theta)$ to produce updated $\boldsymbol{\theta}^{(k+1)}$.

In practice, if the M-step is much faster than the simulation step, the number of simulations $m_k = 1$ may set for all the iterations. This is also the case in the `saemix` package (Delyon et al., 1999; Comets et al., 2017). The simplified E-step is to simulate only a single vector of $\mathbf{b}_i$, and update $Q_k(\theta)$ according to

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k([\log f(\mathbf{y}_i|\mathbf{b}_i, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i|D)] - Q_{k-1}(\theta)).$$

The SAEM algorithm improves the computing time greatly, since the number of simulations M in MCEM is very large.

The convergence of the SAEM algorithm depends on the sequence of the step size γ_k. As discussed in Delyon et al. (1999), γ_k should be a decreasing sequence with a rate slower than 1 and converging to 0. In practice, in the first few initial iterations (default by 6 in `saemix`), the step size γ_k is set to 0, since computing the expectation $Q(\theta)$ is uninterested during the run-in sequence (Comets et al., 2017). During the first K_1 iterations, γ_k is set to 1. It allows the algorithm to explore the parameter space without memory, and to converge to a neighbourhood of the MLE quickly. During the final K_2 iterations, γ_k is set to $1/(k - K_1 + 1)$ to ensure that the estimator is almost sure converged. The convergence of the SAEM algorithm is also proved under general conditions (Kuhn and Lavielle, 2004). The two main conditions are (i) for any $\boldsymbol{\theta} \in \Theta$, the Gibbs algorithm generates a uniformly ergodic chain whose invariant probability is $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta})$; and (ii) for $k = 1, 2, \cdots$, the step size $\gamma_k \in [0, 1]$, $\sum_{k=1}^{\infty} \gamma_k = \infty$, and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.

Compared to the MCEM algorithm, the SAEM algorithm uses the simulated missing values more efficiently. At each iteration of the MCEM algorithm, an entire set of missing values needs to be simulated and all the previously simulated missing values are dropped. This can be easily seen from the E-step of the MCEM algorithm. Recall that to perform MCEM algorithm, at iteration k, we need to simulate a large sample of the missing data $\{\mathbf{b}_i^{(1)}, \cdots, \mathbf{b}_i^{(M)}\}$ from the conditional distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(k)})$. In the E-step, the conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ can be approximated by the following empirical mean:

$$\tilde{Q}(\theta|\theta^{(k)}) = \frac{1}{M}\sum_{j=1}^{M}\left[\log f(\mathbf{y}_i|\mathbf{b}_i^{(j)}, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i^{(j)}|D)\right].$$

At iteration $k+1$, another set of missing data $\{\mathbf{b}_i^{(M+1)}, \cdots, \mathbf{b}_i^{(2M)}\}$ need to be simulated, and the new conditional expectation

$$\tilde{Q}(\theta|\theta^{(k+1)}) = \frac{1}{M}\sum_{j=M+1}^{2M}\left[\log f(\mathbf{y}_i|\mathbf{b}_i^{(j)}, \beta, \boldsymbol{\eta}) + \log f(\mathbf{b}_i^{(j)}|D)\right]$$

is independent of the simulated data at iteration k. However, in the SAEM algorithm, all the simulated missing values contribute to $Q(\theta|\theta^{(k)})$ since $Q(\theta|\theta^{(k)})$ is calculated from $Q(\theta|\theta^{(k-1)})$. Previously simulated missing values weighted less in $Q(\theta|\theta^{(k)})$, with a factor inversely proportional to the step size γ_k. Therefore, the SAEM algorithm converges more quickly than the MCEM algorithm in terms of the number of simulations and the SAEM algorithm is much more computationally efficient.

Samson et al. (2006) extends the SAEM algorithm to NLME models with left-censored responses, based on simulating the left-censored values $\mathbf{y}_{i,cen}$ from a right-truncated Gaussian distribution $f(\mathbf{y}_{i,cen}|\mathbf{y}_{i,obs}, \mathbf{b}_i, \boldsymbol{\theta}^{(k)})$ based on the Gibbs sampling in the E-step of the SAEM algorithm.

For NLME models with left-censored data, the unobserved values are $(\mathbf{y}_{i,cen}, \mathbf{b}_i)$, with $\mathbf{y}_{cen,i}$ being the left-censored data vector and $\mathbf{b}_i$ being the random effects. The simulation step of the SAEM algorithm is the simulation of the missing data $(\mathbf{y}_{i,cen}, \mathbf{b}_i)$ from the conditional distribution $f(\mathbf{y}_{i,cen}, \mathbf{b}_i|\mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, which can be performed by using Gibbs sampling method. At the k-th iteration of the SAEM algorithm, the Gibbs sampling procedure can be divided into two steps:

1. Simulate $\mathbf{b}_i^{(k)}$ by using a Metropolis-Hastings(M-H) algorithm constructing a Markov Chain $\mathbf{b}_i^{(k)}$ with $f(\mathbf{b}_i^{(k)}|\mathbf{y}_{i,obs}, \mathbf{y}_{i,cen}^{(k-1)}; \boldsymbol{\theta}^{(k-1)})$ as the unique stationary distribution,

2. Simulate $\mathbf{y}_{i,cen}^{(k)}$ with the posterior right-truncated Gaussian distribution $f(\mathbf{y}_{i,cen}^{(k)}|\mathbf{y}_{i,obs}, \mathbf{b}_i^{(k)}; \boldsymbol{\theta}^{(k-1)})$.

Then the E-step and M-step are similar as the regular SAEM algorithm discussed earlier in this section. Under the two assumptions and general additional conditions, the estimate sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ produced by the extended SAEM algorithm converges towards a (local) maximum of the likelihood $L_o(\boldsymbol{\theta})$ in (2.16).

The SAEM algorithm, combining a stochastic approximation to the likelihood with an EM algorithm, is shown to converge much faster to the maximum likelihood estimators than the MCEM algorithm (Delyon et al., 1999) and perform much better than the linearization methods in the sense of less non-convergence (Girard and Mentré, 2005). Another advantage of using SAEM is that there is developed software, such as "*Monolix*" (Lavielle, 2014). Comets et al. (2017) also provides an implementation of the SAEM algorithm in the *R* software through the `saemix` package. All the data analysis and simulation studies in this thesis are conducted in *R*.

2.5 A Simulation Study

In this section, a simulation study is conducted to evaluate the performances of the SAEM and the linearization method for NLME models, based on the `saemix` package, compared to regular `nlme` package for NLME model fitting in software *R* under different settings. The package `nlme` uses linearization method (Lindstrom and Bates, 1990), while the package `saemix` uses SAEM method for parameter estimations (Comets et al., 2017).

The performances of different methods are compared based on the relative biases and the relative mean square errors (rMSEs) of the estimates, the computational time, and the frequency of convergence problems. For a parameter β and its estimate $\hat{\beta}^{(i)}$ in the i-th simulation, the relative bias and mean square error (MSE) of the parameter estimates are defined as follows:

- relative bias (%) of $\hat{\beta} = \left| \dfrac{\sum_{i=1}^{N}(\hat{\beta}^{(i)} - \beta)}{N\beta} \right| \times 100\%$,
- relative MSE (%) of $\hat{\beta} = \dfrac{\sum_{i=1}^{N}(\hat{\beta}^{(i)} - \beta)^2}{N\,|\beta|} \times 100\%$,

where N is the number of simulation repetitions. The computational time is in hours, and the frequency of convergence problems is presented by the

number of non-converged runs before there are enough successful estimates. The simulation study is designed as follows:

- Step 1: the observed data are generated from the following nonlinear mixed effects model:

$$\begin{aligned} y_{ij} &= \log_{10}(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}) + e_{ij}, \qquad (2.20) \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \\ i &= 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i, \end{aligned}$$

  where y_{ij} is the observed value for individual i at time t_{ij}, $\{P_1, P_2, \lambda_1, \lambda_2\}$ are the fixed effects, $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$ is a vector of random effects, and $\mathbf{e}_i = (e_{i1}, e_{i2}, \cdots, e_{in_i})^T$ is the measurement errors. Assume that $e_{ij} \sim N(0, \sigma^2)$ and $\mathbf{b}_i \sim N(0, D)$. To keep the procedure simple, assume that there are no censored observations in the simulated dataset.

- Step 2: fit the simulated data using the `nlme` package and the `saemix` package, respectively. If there is a non-convergence problem, re-simulate the dataset and count for 1 non-converged run.

- Step 3: repeat the above steps for N times.

Simulation Study I

In Setting I, the true values of the model parameters are set to be similar to the estimates in a real data analysis, where $P_1 = 17$, $P_2 = 2.6$, $\lambda_1 = 4.0$, and $\lambda_2 = 0.05$, $\sigma = 0.5$, and

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.03 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 0.0001 \end{pmatrix}.$$

The sample size is set to be $n = 50$ individuals. The within-individual longitudinal measurements are set to be 10 repeated measurements: $t = (0.5, 1.7, 2.3, 3.0, 4.6, 6.5, 7.6, 11.2, 14.9, 19.1)$. The simulations are repeated $N = 100$ times.

The simulation results based on model (2.20) are shown in Table 2.1. We can see that the computational time for `nlme` is much longer than `saemix`, while `nlme` seems more accurate than `saemix` based on the bias and rMSE, especially for parameter P_2 and λ_2 in the second phase of viral decay where

Table 2.1: Simulation results of comparing the linearization method and SAEM method based on model (2.20).

Method	Time	NC	Parameter	True value	Bias	rMSE
linearization	2.0	18	P_1	17.0	-0.6	6.7
			λ_1	4.0	-2.5	10.4
			P_2	2.6	-2.4	11.0
			λ_2	0.05	-14.4	6.6
SAEM	0.1	0	P_1	17.0	-19.7	97.1
			λ_1	4.0	-23.0	55.9
			P_2	2.6	121.0	233.9
			λ_2	0.05	1587.4	430.5

Note: Time is the computational time in hours, and NC is the number of non-convergence problems. Bias and rMSE are percentage relative bias and percentage relative mean square error respectively, units in %.

the variation in the data is larger than that in the first phase. This is unsurprising because the SAEM algorithm only simulates one set of unobserved data in the E-step. The SAEM algorithm is computationally more efficient, but may produce inaccurate results. The `nlme` package uses the linearization method for parameter estimations. The linearization method produces more accurate results, but relatively computationally expensive. It may also sometimes cause convergence problems.

Different Variations of Data

The performances of the linearization method and the SAEM algorithm are also compared under different conditions. Table 2.2 compares the two methods when the variation of data is larger, while the models and the rest of the parameter values we consider are similar to those in Setting I. The within-individual standard deviation σ is increased from 0.5 to 2, and the between-individual variance is increased by changing the covariance matrix D to

$$D = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0.0005 \end{pmatrix}.$$

By comparing Table 2.1 and 2.2, we can see that when the variation of data is increased significantly, the computational time for `nlme` is increased, and `nlme` produces worse results based on the bias and rMSE.

Table 2.2: Simulation results when the between-individual variation and within-individual variation increase based on model (2.20).

			between individual variation increased				within individual variation increased			
Method	Par	TV	Time	NC	Bias	rMSE	Time	NC	Bias	rMSE
linearization	P_1	17.0	4.4	36	-0.6	9.5	49.3	411	-3.1	22.7
	λ_1	4.0			-2.3	13.4			-12.5	32.9
	P_2	2.6			-3.5	12.8			-17.3	46.0
	λ_2	0.05			-20.2	7.9			-95.7	29.1
SAEM	P_1	17.0	0.1	0	-17.8	88.3	0.1	0	4.8	27.2
	λ_1	4.0			-20.2	49.0			12.2	36.5
	P_2	2.6			112.7	217.2			27.3	58.6
	λ_2	0.05			1442.6	397.9			-74.7	27.3

Note: Par represents parameter, TV represents the true values of the parameters, and NC represents the number of non-convergence problems. Time is the computational time in hours. Bias and rMSE are percentage relative bias and percentage relative mean square error respectively, units in %.

However, `saemix` produces similar results when data has larger variations. Especially, the performances of `nlme` and `saemix` are quite close when the within-individual variation is larger, while `saemix` is much faster than `nlme`.

Different Sample Size and Number of Repeated Measurements

Table 2.3 shows the simulation results when the sample size is increased and when there are more frequent repeated measurements. In Table 2.3, the sample size n is increased from 50 to 200, and the number of repeated measurements n_i is increased from 10 to 14, where the new set of repeated measurements $t = (0.4, 0.5, 1.2, 2.1, 3.1, 4.1, 5.5, 6.7, 8.2, 9.8, 11.6, 13.5, 15.5, 18)$. The rest of the true values of parameters remain the same as in Setting I.

By comparing Table 2.1 and 2.3, we can see that when n is increased, the performances of both `nlme` and `saemix` are improved based on the bias and rMSE. The performance of `nlme` improves slightly, while the performance of `saemix` improves significantly. However, `nlme` takes much longer to complete the simulation and the number of non-convergence problems increases. The results are expected because a larger sample size generally results in better estimations.

When there are more frequent repeated measurements, the computational time of `nlme` is slightly slower, but the performance of `nlme` does not

Table 2.3: Simulation results when the sample size n is increased to 200 and the number of repeated measurements n_i is increased to 14 based on model (2.20).

				n increased			n_i increased			
Method	Par	TV	Time	NC	Bias	rMSE	Time	NC	Bias	rMSE
linearization	P_1	17.0	29.63	106	-0.4	4.0	4.2	47	-4.1	6.6
	λ_1	4.0			-1.8	6.3			-1.5	9.6
	P_2	2.6			-2.3	8.8			-4.2	11.6
	λ_2	0.05			-15.0	5.1			-22.8	7.4
SAEM	P_1	17.0	0.3	0	-7.0	41.9	0.1	0	-24.4	105.2
	λ_1	4.0			-7.5	24.1			157.2	264.9
	P_2	2.6			44.2	101.1			-28.3	59.7
	λ_2	0.05			530.9	176.0			2201.1	519.5

Note: Par represents parameter, TV represents the true values of the parameters, and NC represents the number of non-convergence problems. Time is the computational time in hours. Bias and MSE are percentage relative bias and percentage relative mean square error respectively, units in %.

seem to improve significantly based on the bias, rMSE, and number of non-convergence problems. However, `saemix` shows a slightly worse performance compared to Table 2.1. In general, the two methods need a large number of repeated measurements to perform well, since more repeated measurements provide more information about the longitudinal covariate process. The results do not show a significant difference probably because the change of the number of repeated measurements is not large enough. We may consider an even larger number of repeated measurements to investigate the influence of the number of repeated measurements within individuals on parameter estimation.

Simulation Study II

To check the sensitivity of the performances of the linearization method and the SAEM algorithm to different models, we conduct another simulation study based on another model

$$\begin{aligned} y_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij}+\exp(\beta_{2i}-\beta_{3i}t_{ij})}+\beta_{4i}+\xi_{ij}, \qquad (2.21)\\ \beta_{1i} &= \beta_1+\tau_{1i}, \quad \beta_{2i}=\beta_2+\tau_{2i}, \quad \beta_{3i}=\beta_3+\tau_{3i}, \quad \beta_{4i}=\beta_4+\tau_{4i},\\ i &= 1,2,\cdots,n, \quad j=1,2,\cdots,n_i, \end{aligned}$$

where vector $\beta_i = (\beta_{1i}, \cdots, \beta_{4i})^T$ contains individual-specific parameters, vector $\beta = (\beta_1, \cdots, \beta_4)^T$ contains fixed effect parameters and represents some key features of viral rebound trajectory, $\tau_i = (\tau_{1i}, \cdots, \tau_{5i})^T \sim N(0, G)$ contains random effects with G being a covariance matrix, and ξ_{ij} is within-individual random error. We assume that the random effects τ_i and the random error ξ_{ij} are independent, and ξ_{ij} are i.i.d. $\sim N(0, \omega^2)$. This model is used for fitting viral rebound trajectories, and details about this viral rebound model are described in Section 4.2.3. The entire simulation process is similar as in Simulation Study I. Note that there are no censored values in the simulated data for simplicity.

The true values of the model parameters are set to be similar to the estimates in a real data analysis, where $\beta_1 = 3.2$, $\beta_2 = 5.6$, $\beta_3 = 10$, and $\beta_4 = 1$, $\omega = 0.5$, and

$$G = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0.006 \end{pmatrix}.$$

The sample size is set to be $n = 50$ individuals. The within-individual longitudinal measurements are set to be 11 repeated measurements: $t = (0.2, 0.7, 1.1, 1.6, 2.1, 2.5, 3.0, 3.5, 4.0, 4.4, 4.9)$. The simulations are repeated $N = 100$ times.

Table 2.4 shows the simulation results of the `nlme` package and `saemix` package based on model (2.21). We can see that `saemix` performs better than `nlme`, as the bias and rMSE are smaller, computing time is less, and the number of non-convergence problems is fewer. The results are similar to those in Simulation Study I, which indicate that the performances of the linearization method and the SAEM algorithm may not be sensitive to the models.

Different Variations of Data

Table 2.5 shows the simulation results when the variations of data is increased. In Table 2.5, the within-individual standard deviation ω is increased from 0.5 to 2, and the variance covariance matrix G is changed to

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & 0.02 \end{pmatrix}.$$

Table 2.4: Simulation results of comparing the linearization method and SAEM method based on model (2.21).

Method	Time	NC	Parameter	True value	Bias	rMSE
linearization	39.3	447	β_1	3.2	40.1	203.7
			β_2	5.6	-82.1	201.6
			β_3	10	-73.2	241.0
			β_4	1	-64.0	126.0
SAEM	0.10	0	β_1	3.2	-1.8	8.5
			β_2	5.6	55.3	169.6
			β_3	10	51.0	211.3
			β_4	1	4.2	9.7

Note: Time is the computational time in hours, and NC is the number of non-convergence problems. Bias and rMSE are percentage relative bias and percentage relative mean square error respectively, units in %.

Table 2.5: Simulation results when ω is increased to 2 and the diagonal elements in G is increased based on model (2.21).

			between individual variation increased				within individual variation increased			
Method	Par	TV	Time	NC	Bias	rMSE	Time	NC	Bias	rMSE
linearization	β_1	3.2	3.57	303	24.1	1114.9	39.38	307	-4.2	42.7
	β_2	5.6			-82.1	228.9			-24.2	28.8
	β_3	10			-77.3	280.5			-11.1	92.8
	β_4	1			59.4	140.8			-24.8	103.5
SAEM	β_1	3.2	0.11	0	-26.5	49.3	0.093	0	-7.6	48.0
	β_2	5.6			2126.7	5687.5			-13.9	30.5
	β_3	10			1760.5	6251.1			16.1	72.7
	β_4	1			88.5	91.0			287.6	116.2

Note: Par represents parameter, TV represents the true values of the parameters, and NC represents the number of non-convergence problems. Time is the computational time in hours. Bias and rMSE are percentage relative bias and percentage relative mean square error respectively, units in %.

Table 2.6: Simulation results when the sample size n is increased to 200 and the number of repeated measurements n_i is increased to 21 based on model (2.21).

			n increased				n_i decreased			
Method	Par	TV	Time	NC	Bias	rMSE	Time	NC	Bias	rMSE
linearization	β_1	3.2	116.9	358	24.0	128.3	47.8	332	7.9	39.3
	β_2	5.6			-80.9	197.2			-36.4	93.6
	β_3	10			-71.3	233.1			-34.7	119.6
	β_4	1			-47.5	101.0			-12.1	18.9
SAEM	β_1	3.2	0.29	0	-2.4	5.7	0.13	0	1.0	6.9
	β_2	5.6			32.1	88.6			-0.4	31.4
	β_3	10			27.2	100.9			-1.7	38.2
	β_4	1			6.0	7.2			-0.6	5.5

Note: Par represents parameter, TV represents the true values of the parameters, and NC represents the number of non-convergence problems. Time is the computational time in hours. Bias and rMSE are percentage relative bias and percentage relative mean square error respectively, units in %.

Table 2.5 shows that when within-individual standard deviation ω is increased, the performance of `saemix` is much worse than `nlme`, as the bias and rMSE are larger. This may be due to inaccurate single draw in the SAEM when the model is more nonlinear and within-individual variation is larger. It may lead to biased results as the model is complex and the data has large variations. In addition, when the variance-covariance matrix G is increased, the performance of `nlme` and `saemix` are quite close based on the bias and rMSE, while `nlme` has a large number of non-convergence issues.

Different Sample Size and Number of Repeated Measurements

In Table 2.6, we increase the sample size and the number of repeated measurements to examine the effect of sample size and number of repeated measurements on the results. We consider a larger sample size $n = 200$. We can see that both `nlme` and `saemix` perform better as the sample size increases, which is expected.

In addition, we increase the number of repeated measurements from 11 to 21: $t = (0.1, 0.3, 0.6, 0.8, 1.1, 1.3, 1.5, 1.8, 2.0, 2.2, 2.5, 2.9, 3.3, 3.7, 4.1, 4.5, 5.0, 5.5, 6.0, 6.8, 7.6)$. Table 2.6 shows that when the number of repeated measurements is increased, both `nlme` and `saemix` perform better based on relative bias and rMSE. The results are expected since more repeated measurements provide more information about the longitudinal covariate process,

and the two methods result in more accurate parameter estimates.

Simulation Summaries

In general, `nlme` (the linearization method) produces more accurate estimates than `saemix` (the SAEM algorithm), but with longer computational time and more non-convergence problems. The performances of the two methods may depend on the variations of data, sample size, and number of repeated measurements. For example, the two methods perform better as the sample size and the number of repeated measurements is increased. The two methods usually have similar performances as the variations of data are increased. The performances of `nlme` and `saemix` seem to be insensitive to different models. However, the number of simulation repetitions $N = 100$ may not be large enough to produce reliable results. We may consider a larger number of simulation repetitions in future research. In conclusion, it is important to choose a suitable parameter estimation method for NLME models by weighing the accuracy of the estimates against the computational cost.

Chapter 3

Simultaneous Inference for Joint NLME Models with left-censored Responses

As reviewed in the previous chapter, the SAEM algorithm makes use of a stochastic approximation procedure for parameter estimations and inference for the NLME models. As an alternative to the MCEM algorithm and the linearization method of Lindstrom and Bates (1990), the SAEM algorithm converges much faster to a local maxima of the likelihood function and performs reasonably well.

In practice, there may be two NLME models which are associated. For example, in the motivating HIV dataset, an NLME model can be fitted for viral decay during ART, and a separate NLME model can be fitted for viral rebound data following ART interruption. The two NLME models are linked through some shared parameters, as will be described later. For parameter estimations and inference, we may consider the MCEM or SAEM algorithm for both models simultaneously based on the joint likelihood of all observed data. However, such a joint likelihood method can be computationally extremely intensive. Therefore, we propose a *three-step (TS) method* to reduce the computation burden.

3.1 Joint Modelling for Longitudinal Data

In practice, we may need to model several longitudinal processes jointly. For example, in modelling the HIV viral loads, we want to study the association between the key features of viral decay during ART and important characteristics of viral rebound following ART interruption. As shown in Figure 1.1, the measurements before the gray dashed vertical lines belong to the first longitudinal process, representing the viral decay before ART interruption. The measurements after the gray vertical dashed lines are the viral rebound following ART interruption, which belong to the second

longitudinal process.

Two NLME models are used to model the viral decay and viral rebound respectively, and the two models share some random effects, which characterize the individual-specific features of viral decay. Denote y_{ij} as the observed value of viral load at t_{ij} during ART and w_{ij} as the observed value of viral load at t^*_{ij} following ART interruption. Let's consider the following viral decay NLME model proposed by Wu and Ding (1999)

$$\begin{aligned} y_{ij} &= \log_{10}(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}) + e_{ij}, \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \\ i &= 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i, \\ \mathbf{b}_i &\sim N(0, D), \quad \mathbf{e}_i \sim N(0, R_i), \end{aligned}$$

and the following viral rebound NLME model proposed by Wang et al. (2020)

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t^*_{ij}}{t^*_{ij} + \exp(\beta_{2i} - \beta_{3i}t^*_{ij})} + \beta_{4i} + \xi_{ij}, \\ \beta_i &= R_i\beta + \tau_i, \qquad\qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n^*_i. \end{aligned}$$

Parameters λ_1 and λ_2 represent the viral decay rate during ART, β_1 represents set point after a rebound, β_2 and β_3 represent rates of rise in viral load during rebound, and β_4 represents initial viral load value at the start of rebound. Detailed justifications of both models will be discussed in Chapter 4. The two models are linked through shared parameters, since the individual specific characteristics $\mathbf{b}_i$ from viral decay may be predictive for the individual trajectories of viral rebound. For example, if we want to study the association between the viral decay rate and the set point in the second longitudinal process, we may consider the following second-stage model for viral rebound

$$\beta_{1i} = \beta_1 + \gamma_{13}\beta_{3i} + \tau_{1i}, \quad \beta_{ki} = \beta_k + \tau_{ki}, \quad k = 2, 3, 4.$$

In this case, we need to model the two longitudinal processes simultaneously.

Statistical inference of several models with shared parameters may be based on separate inference or joint inference. For separate inference of two joint models, a simple approach is a two-step method:

- Step 1: estimate the shared variables or parameters in one model based on the observed data.

- Step 2: estimate the parameters in the other model separately, with the shared variables or parameters substituted by the estimated values from Step 1.

This two-step method is simple and naive. Standard statistical software, such as R, can be readily used. However, this simple approach may lead to biased estimation and under-estimated standard errors of the parameter estimates. The estimation may be biased, especially when the two longitudinal processes are strongly associated. In addition, the standard errors of the parameter estimates in the second NLME model may be under-estimated because the uncertainty of estimation in the first step is not incorporated in the second step.

Statistical inference for joint models can also be based on the (joint) likelihood of all the observed data. Joint likelihood means the likelihood for the two joint models. The MLEs of all model parameters can be obtained simultaneously by maximizing the joint likelihood. Compared to separate inferences, the MLEs from the joint likelihood approach are more efficient. Since the likelihood method is a standard approach for inference in mixed effects models, the joint likelihood method appears to be a natural choice for inference in joint mixed effects models.

Motivated from the viral decay model and the viral rebound model in an HIV study, let's consider two NLME models for modelling the longitudinal processes with censored responses. Let y_{ij} be the response in the first longitudinal process for individual i at time t_{ij}, which follows the NLME model

$$\begin{aligned} y_{ij} &= g_1(t_{ij}, \beta_i) + e_{ij}, && (3.1)\\ \beta_i &= h_1(\mathbf{x}_i, \beta, \mathbf{b}_i), \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i, && (3.2)\\ \mathbf{b}_i &\sim N(0, D), \quad e_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2). && (3.3) \end{aligned}$$

The observed responses y_{ij} can be written as (q_{ij}, c_{ij}), and

$$y_{ij} = \begin{cases} q_{ij} & \text{if } c_{ij} = 0, \\ \leq d & \text{if } c_{ij} = 1, \end{cases}$$

where c_{ij} is the censoring indicator and d is a known constant such as a detection limit. Let $f(\cdot)$ and $F(\cdot)$ denote a generic density function and the corresponding cumulative density function, respectively. Let $f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\sigma})$ be the density function of the above NLME model (3.1)−(3.3) for the first longitudinal process, given random effects $\mathbf{b}_i$, mean parameters $\boldsymbol{\beta}$, and variance-

covariance parameters $\boldsymbol{\sigma}$. Assume that $y_{i1}, \cdots, y_{in_i}$ are conditionally independent given the random effect $\mathbf{b}_i$ and $\mathbf{b}_i \sim N(0, D)$, where D is an unknown covariance matrix.

Similarly, let w_{ij} be the response in the second longitudinal process for individual i at time t^*_{ij}, which follows the NLME model

$$\begin{aligned} w_{ij} &= g_2(t^*_{ij}, \beta^*_i) + \xi_{ij}, && (3.4)\\ \beta^*_i &= h_2(\mathbf{x}^*_i, \beta^*, \boldsymbol{\tau}_i), \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n^*_i, && (3.5)\\ \boldsymbol{\tau}_i &\sim N(0, G), \quad \xi_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \omega^2). && (3.6) \end{aligned}$$

The observed responses w_{ij} can be written as (q^*_{ij}, c^*_{ij}), and

$$w_{ij} = \begin{cases} q^*_{ij} & \text{if } c^*_{ij} = 0, \\ \leq d^* & \text{if } c^*_{ij} = 1. \end{cases}$$

Assume that $w_{i1}, \cdots, w_{in^*_i}$ are conditionally independent given the random effect $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_i \sim N(0, G)$, where G is an unknown covariance matrix. Let $f(\mathbf{w}_i|\mathbf{z}^*_i, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\beta}^*, \boldsymbol{\sigma}^*)$ be the density function of the above NLME model (3.4)−(3.6) for the second longitudinal process, given random effects $\boldsymbol{\tau}_i$, mean parameters $\boldsymbol{\beta}^*$, variance-covariance parameters $\boldsymbol{\sigma}^*$, and the random effects $\mathbf{b}_i$ from the first longitudinal process. The individual-specific characteristics of the first longitudinal process, represented by $\mathbf{b}_i$, are included in the NLME model (3.4)−(3.6) since they may be predictive for the individual trajectories of the second longitudinal process. For example, in the motivating example, individual-specific viral decay rates during ART may be associated with individual specific viral rebound rates or set points following ART interruption. In this case, joint modelling is needed.

Let $\boldsymbol{\theta}$ be the collection of all unknown parameters. The joint likelihood for both longitudinal processes can be written as

$$\begin{aligned} L_o(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int \int \{ \prod_{j=1}^{n_i} (f(y_{ij}|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}))^{1-c_{ij}} (F(d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}))^{c_{ij}} \} \\ \times \{ \prod_{k=1}^{n^*_i} (f(w_{ik}|\mathbf{z}^*_i, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}))^{1-c^*_{ij}} (F(d^*|\mathbf{z}^*_i, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}))^{c^*_{ij}} \} \\ \times f(\mathbf{b}_i|B) f(\boldsymbol{\tau}_i|G) d\mathbf{b}_i d\boldsymbol{\tau}_i, \qquad (3.7) \end{aligned}$$

where

$$\begin{aligned} F(d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}) &\equiv P(Y_{ij} < d|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}), \\ F(d^*|\mathbf{z}^*_i, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}) &\equiv P(W_{ij} < d^*|\mathbf{z}^*_i, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}), \end{aligned}$$

Y_{ij} is the random version of y_{ij}, and W_{ij} is the random version of w_{ij}. The likelihood (3.7) does not have an analytic expression.

The Monte Carlo EM algorithm can be used to find the MLEs of the parameters $\boldsymbol{\theta}$ for the two joint NLME models with censored responses. Treating the censored values in $\mathbf{y}_i$ and $\mathbf{w}_i$, and the random effects $\mathbf{b}_i$ and $\boldsymbol{\tau}_i$ as "missing data", we have "complete data" $\{(\mathbf{y}_i, \mathbf{w}_i, \mathbf{z}_i, \mathbf{z}_i^*, \mathbf{b}_i, \boldsymbol{\tau}_i), i = 1, 2, \cdots, n\}$. The log-likelihood of the complete data can be written as

$$\begin{aligned} l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_c^{(i)}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [\log f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}) + \log f(\mathbf{b}_i|\boldsymbol{\theta}) \\ &\quad + \log f(\mathbf{w}_i|\mathbf{z}_i^*, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}) + \log f(\boldsymbol{\tau}_i|\boldsymbol{\theta})]. \end{aligned}$$

The E-step is to compute the conditional expectation of the complete data log-likelihood given the observed data and current parameter estimates. The conditional expectation for the i-th observation at the $(k+1)$-th EM iteration can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\{l_c(\boldsymbol{\theta})|\mathbf{q}_i, \mathbf{q}_i^*, \mathbf{c}_i, \mathbf{c}_i^*, \mathbf{z}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}\} \\ &= \int\int\int\int [\log f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}^{(k)}) + \log f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)}) \\ &\quad + \log f(\mathbf{w}_i|\mathbf{z}_i^*, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}^{(k)}) + \log f(\boldsymbol{\tau}_i|\boldsymbol{\theta}^{(k)})] \\ &\quad \times f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\tau}_i|\mathbf{q}_i, \mathbf{q}_i^*, \mathbf{c}_i, \mathbf{c}_i^*, \mathbf{z}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}) d\mathbf{y}_{cen,i} d\mathbf{b}_i d\mathbf{w}_{cen,i} d\boldsymbol{\tau}_i, \end{aligned}$$

where $\mathbf{y}_{cen,i}$ and $\mathbf{w}_{cen,i}$ are vectors of censored responses.

Since the above conditional expectation $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is extremely complicated, we are unable to obtain a closed-form expression for $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, and evaluate the expectation analytically. Therefore, we can use a Monte Carlo method to generate large samples of the "missing data" $(\mathbf{y}_{cen,i}, \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\tau}_i)$ from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\tau}_i|\mathbf{q}_i, \mathbf{q}_i^*, \mathbf{c}_i, \mathbf{c}_i^*, \mathbf{z}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$ at each iteration, and then approximate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ by an empirical mean. The sampling procedure can be done using the Gibbs sampler by iteratively sampling from the full conditionals $f(\mathbf{y}_{cen,i}|\mathbf{b}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, $f(\mathbf{w}_{cen,i}|\mathbf{b}_i, \boldsymbol{\tau}_i, \mathbf{q}_i^*, \mathbf{c}_i^*, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$, and $f(\boldsymbol{\tau}_i|\mathbf{b}_i, \mathbf{w}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$. Sampling from the full conditionals can be done using rejection sampling methods. The MCEM algorithm can be computationally intensive since the dimension of the missing data $(\mathbf{y}_{cen,i}, \mathbf{w}_{cen,i}, \mathbf{b}_i, \boldsymbol{\tau}_i)$ is very high, so simulating large numbers from the conditional distribution can be very slow.

We may consider the linearization method based on Lindstrom and Bates (1990) for the two joint NLME models, with some modification. Briefly, we can rewrite NLME models (3.1)-(3.6) as a single equation of z_{ij}, where z_{ij} is

the observed response for individual i at time t_{ij} for the entire study period. Then, z_{ij} can be modelled by using two nonlinear functions $u_{ij}(\mathbf{x}_i, \boldsymbol{\beta}_i, \mathbf{b}_i)$ and $u^*_{ij}(\mathbf{x}^*_i, \boldsymbol{\beta}^*_i, \boldsymbol{\tau}_i)$ at the same time. As discussed in Chapter 2, statistical inference for an NLME model based on the linearization method may offer potential convergence problems. Hence, there may be more frequent convergence problems when using the linearization method for the two joint NLME models. In addition, the approximation may be less accurate if the model of z_{ij} is very "nonlinear". Hence, the linearization method may be more computationally efficient than the MCEM algorithm for joint NLME models, but the parameter estimates may be less accurate, and may offer potential convergence problems. Details about the linearization method for the two joint NLME models are provided in Section 6.2 of future research.

In addition, we may consider the SAEM algorithm for the two joint NLME models. Recall that the SAEM algorithm replaces the E-step of the MCEM algorithm by a single draw from the conditional distribution based on an MCMC method, and then uses a stochastic approximation to update the expectation. Although the SAEM algorithm for joint mixed effects models based on the joint likelihood may be computationally more efficient than the MCEM algorithm, it may still be computationally expensive, since the dimension of the missing data $(\mathbf{y}_{cen,i}, \mathbf{w}_{cen,i}, \mathbf{b}_i, \boldsymbol{\tau}_i)$ is very high. In addition, parameter estimates for the joint mixed effects models when using SAEM algorithm may be less accurate than the linearization method according to the simulation results in Section 2.5. More details about the SAEM algorithm for the two joint NLME models are discussed in Section 6.2.

In conclusion, for joint inference of two NLME models, the MCEM algorithm is extremely computationally expensive, the linearization method may have convergence issues, and the SAEM algorithm may not be accurate. The simulation studies in Section 2.5 may be viewed as a reference. There are also some drawbacks for joint likelihood approaches. First of all, the joint likelihood approaches can be quite time-consuming, since the joint likelihood may involve high-dimensional and intractable integrals. In addition, model or parameter identifiability may be a potential problem due to a possibly large number of unknown parameters in joint models. This arises a non-identifiable problem, such that two sets of different parameters may lead to the same likelihood. Although we can use *Monolix* and the `nlme` package in *R* for the joint NLME models with SAEM and linearization, the implementation and computation may take a long time. Hence, we propose to use SAEM for each NLME model with left censoring separately to reduce the computation burden, so-called a *three-step (TS) method*. This method is easy to implement using the existing software and it is computationally

efficient. For computational simplicity, the joint likelihood methods based on the MCEM, linearization, and SAEM are not considered in this thesis. The performance of different joint likelihood methods can be studied in future research. More details about the three-step method are discussed in the next section.

3.2 A Three-Step Method

As discussed in Section 3.1, a joint likelihood method is computationally extremely intensive and the naive two-step method may lead to biased estimation and under-estimated standard errors of the parameter estimates. Therefore, we propose a new method, *three-step (TS) method* for two longitudinal processes.

The *three-step (TS) method* uses SAEM for each NLME model with left censoring separately to simplify the computation process and the implementation can be done easily:

- Step 1: For the first longitudinal process, fit an NLME model with left censoring using the SAEM algorithm and then obtain the maximum likelihood estimates (MLEs) of the fixed parameters and the empirical Bayes estimates of the random effects $\hat{\mathbf{b}}_i$;
- Step 2: For the second longitudinal process, fit an NLME model with left censoring using the SAEM algorithm, with the random effects $\mathbf{b}_i$ in the covariates model substituted by their empirical Bayes estimates $\hat{\mathbf{b}}_i$ from Step 1;
- Step 3: Obtain the standard errors of the parameter estimates based on a (parametric) bootstrap method.

In general, the above three-step method can be applied to any two or more linked longitudinal processes. It can also be used in an HIV study, with the viral loads and CD4 counts during ART and the viral loads following ART interruption as three longitudinal processes. The correspondingly modified three-step method is discussed in section 4.3.

In the three-step method, the random effects $\mathbf{b}_i$ in the second step are substituted by their empirical Bayes estimates $\hat{\mathbf{b}}_i$. The resulting standard errors of the parameter estimates are not accurate. Therefore, a parametric bootstrap method is used to obtain the standard errors, which incorporates the estimation uncertainty of random effect estimates $\hat{\mathbf{b}}_i$ in Step 1. The parametric bootstrap method works as follows:

1. Simulate data with left censoring, based on the fitted NLME models using the above three-step method, where the model parameters are replaced by their estimates.

2. For the simulated data, fit all models again using the above three-step method and obtain all parameter estimates.

3. Repeat the above process B times (say, $B = 100$), we obtain B estimates for each parameter. The sample standard deviation of these B estimates of each parameter is the parametric *bootstrap estimate* of the standard error of the corresponding parameter estimate.

The above bootstrap method incorporates the estimation uncertainty of the parameter and random effect estimates in the TS method with separate model fitting, so it should produce more reliable standard errors of the parameter estimates than those from separate model fitting.

The TS method is much faster than the joint likelihood method and it can be easily implemented in statistical software. Therefore, the TS method is applied to the motivating HIV dataset to study the association of the viral loads before and following ART interruption. Data analysis, including study background, model descriptions, and results, are discussed in the following chapter.

Chapter 4

Data Analysis

In the previous chapter, we propose a three-step method for parameter estimations in the joint NLME models with left censoring based on the SAEM algorithm. The proposed method can be easily implemented in statistical software (e.g., R), and it is much faster than the joint likelihood method. Moreover, it produces more reliable standard errors of the parameter estimates than those from separate model fitting. In this chapter, we describe the two NLME models motivated by the real HIV dataset, and we perform a comprehensive data analysis based on the three-step method. We also obtain results from a *naive method*, which still uses SAEM algorithm, but the censored values are substituted by half the detection limit, and uses model SE without bootstrapping.

4.1 Data Description and Objective

The dataset is from the Zurich Primary HIV Infection Study (Aceto et al., 2005). The study consists of acutely or recently HIV-1 infected individuals between November 2002 and July 2008. The participants were offered immediate standard first-line antiretroviral therapy (ART) independently of clinical indication and laboratory values (plasma viral load and $CD4^+$ T-cells). After one year of viral suppression below detection limits (< 50 HIV-1 RNA copies/ml of plasma), the participants could choose to stop therapy. Successful ART suppresses viral replication in patients infected with HIV-1, reduces HIV-1 RNA in plasma to levels below the detection limit of standard assays, and results in a reduction of mortality and morbidity. More details can be found in Gianella et al. (2011).

The dataset contains repeatedly measured viral loads (in copies/mL) and CD4 values (in cells/mm^3) during ART and following ART interruption, as well as an indicator of left-censored viral loads, i.e., the indicator equals 1 when viral loads below the detection limit, 0 otherwise. To keep the analysis simple and focused, some variables such as age, gender, and assay types are not included in our analyses. They may be useful in future research. Figure 4.1 (same as Figure 1.1 in Chapter 1) shows the entire viral load

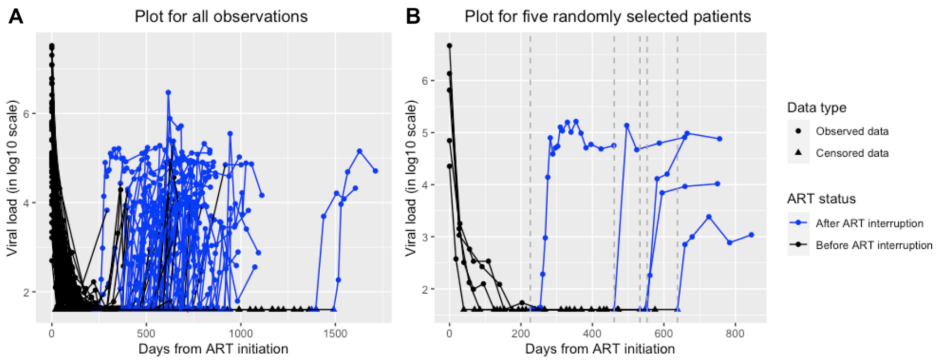

Figure 4.1: Viral load (in $\log_{10}$*-scale) trajectories before and following ART interruption. left-censored values are denoted by triangle dots on the bottom horizontal line with the censored values imputed by the detection limit. Observed values are denoted by circle dots. Data during ART are in black, and data following ART interruption are in blue. The dashed vertical lines in gray indicate times when the ART was interrupted. Figure (A) shows data from all subjects, and Figure (B) shows data from 5 randomly selected subjects.*

trajectories during ART and following ART interruption for all subjects and for 5 randomly selected subjects respectively. For data following ART interruption, we only show the first 36 weeks of data because viral load levels typically stabilize before then.

From Figure 4.1, we can see that viral loads decline rapidly during ART and then may rebound quickly following ART interruption, and that viral loads after reaching peak points during rebound exhibit large variations between subjects without clear patterns. There are two patients in Figure 4.1 (A) exhibit distinct patterns from the other patients. The two patients, who are separated, are not outliers because the duration of ART varies across patients, and they have a long duration of ART. If we only look at viral rebound data (with time since ART interruption), all the curves roughly follow the typical pattern. Our main objective is to study if key features of viral decay during ART are associated with important characteristics of viral rebound following ART interruption. For example, we may be interested in studying whether individual-specific viral decay rates during ART are associated with individual-specific viral rebound rates or set points following ART interruption.

4.2 The Models

Mixed effects models are well-suited for the HIV dataset because there are large variations between individual viral load trajectories, and random effects in the models can be used to incorporate the between-individual variations, as well as individual-specific inference. To model viral load trajectories during ART, Wu and Ding (1999) proposed NLME models based on reasonable biological arguments. The proposed exponential decay models have been shown to fit viral decay trajectories well. For viral rebound trajectories following ART interruption, recently Wang et al. (2020) proposed a different NLME model where the key features of viral rebounds are represented by the model parameters. In this section, we use these two NLME models to fit viral loads before and after ART interruption respectively, as well as an LME model for fitting the CD4 data. The three models are linked through shared random effects, because we are interested in studying the association between individual-specific characteristics of viral decay during ART and important features of viral rebound following ART interruption, and the individual-specific characteristics can be represented by random effects in the models.

Specifically, an NLME model for viral decay and an LME model for CD4 trajectories during ART are considered at the first step. Then, the random effects in these two models, which summarize individual-specific CD4 and viral load trajectories, are used as "covariates" in the viral rebound NLME model following ART interruption. In both NLME models, the left-censored viral loads are incorporated in model fitting, as shown in Chapter 3. Our goal is to exam if the individual-specific viral rebound characteristics following ART interruption are associated with individual-specific CD4 and viral load profiles during ART. To reduce the computation burden, we fit the three models separately based on a three-step method, using the SAEM algorithm. We use a parametric bootstrap method to obtain standard errors of all parameter estimates since a bootstrap method may incorporate estimation uncertainty from separate model fittings. We also obtain results from a naive method, which still uses SAEM algorithm, but substitutes the censored viral load values with half the detection limit. The corresponding SE is obtained based on the model SE, without bootstrapping. The performances of the two methods are further compared in Chapter 5.

4.2.1 An NLME model for viral load during ART

Many deterministic HIV-1 dynamic models have been developed to describe the viral load trajectories after the initiation of potential antiviral treatments. However, most of the models are too complicated and contain too many unknown parameters to be used to analyze real clinical data. There are also several simplified models proposed in the literature (Ho et al., 1995; Wei et al., 1995), but various assumptions have to be made in order to make simplification. Even though these models can be applied, the estimation methods used are not flexible enough to deal with sparse data. As a result, Wu and Ding (1999) introduced random effects in the deterministic models to obtain parametric NLME models and simplified them in different phases of virological responses based on some reasonable biological arguments. To be more specific, the viral load trajectories during ART is shown to typically exhibit exponential decay patterns. A mathematical model for HIV dynamics is proposed first by considering several cell and virus compartments. This model is then simplified based on different phases of virological response, as shown in Figure 4.2. Next, we introduce the NLME model used to fit viral decay data during ART.

Let Y_{ij} be the ($\log_{10}$-transformed) viral load value (in copies/mL) of individual i measured at time t_{ij} during ART. Let y_{ij} be the observed value of Y_{ij}, and let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$, $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i$. We use the similar notations for other variables. The values of Y may be left-censored due to a low detection limit. For the viral load data during ART, as shown in Figure 4.1, let's consider the following viral dynamic NLME model proposed by Wu and Ding (1999)

$$\begin{aligned} y_{ij} &= \log_{10}(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}) + e_{ij}, \qquad (4.1)\\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i},\\ i &= 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n_i,\\ \mathbf{b}_i &\sim N(0, D), \quad \mathbf{e}_i \sim N(0, R_i), \end{aligned}$$

where $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$ are random effects, $\mathbf{e}_i = (e_{i1}, \cdots, e_{in_i})^T$ are random errors of within-individual measurements, D is a covariance matrix of the random errors, and R_i is a covariance matrix of the within-individual random errors. We assume that the random effects $\mathbf{b}_i$ and the within-individual random errors $\mathbf{e}_i$ are independent. We also assume that the within-individual errors are conditionally independent given the random effects, i.e., $R_i = \sigma^2 I_{n_i}$, where I_{n_i} is a $n_i \times n_i$ identity matrix. Note that Y_{ij} is log-transformed in order to satisfy the assumptions of normality and constant variance.

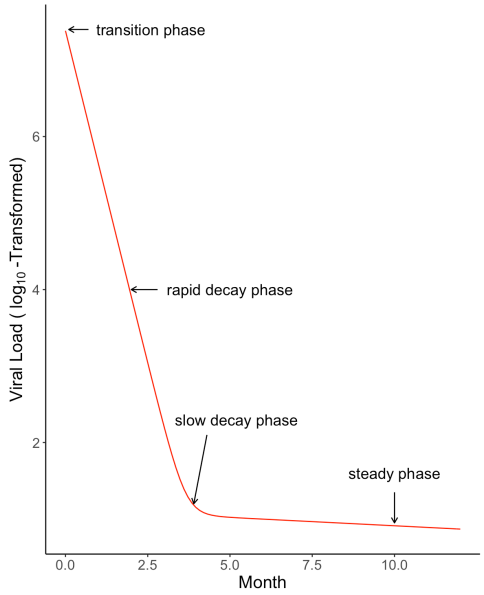

Figure 4.2: Typical viral dynamic profiles during ART based on model (4.1).

The four model parameters represent different phases of plasma viral dynamics: (1) λ_1 is the first-phase viral decay rate, which corresponds to the rapid decay phase reflecting decay of productively, long-lived and/or latently infected cells; (2) λ_2 is the second-phase viral decay rate during ART, which corresponds to the slow decay phase reflecting decay of long-lived and/or latently infected cells and other residual infected cells; and (3) $\log_{10}(P_1+P_2)$ is population viral load value at the start of ART. Figure 4.2 shows the viral decline profile for a typical subject based on model (4.1). Random effects are introduced to each parameter to incorporate large variations between individuals.

As shown in Figure 4.1, some viral load values are left-censored or below the detection limit. In estimating the parameters in the above NLME model, the censored viral load values must be taken into account to avoid biased results (Hughes, 1999; Wu, 2002). The SAEM algorithm simulates the left-censored responses based on the Gibbs sampling in the E-step. Alternatively, we may use a naive method by substituting the censored values by half the detection limit. Since the naive method ignores the uncertainty in the censored values, it may produce more biased results than using the SAEM algorithm.

Note that many mathematical models, such as nonparametric smooth

functions and polynomial functions may fit the viral load dynamic trajectories equally well or even better. However, their biological interpretations may not be valid. The NLME model may be used to predict unmeasured or censored viral load values since it is based on the underlying data generation mechanism. This is one of the main advantages of NLME model over empirical models, which only describe observed data well but may be poor for predictions. The model (4.1) proposed by Wu and Ding (1999) is preferred when analyzing the HIV-1 dynamics after initiation of potential antiviral treatments because it has the following features: (i) good fit to the observed data; (ii) reasonable biological interpretation; (iii) conceptual and computational simplicity.

4.2.2 An LME model for CD4 during ART

Due to the known association between CD4 and viral load, the CD4 values during ART may also be considered in the analysis. The observed CD4 values reflect both short-term biological variation and measurement error, and we may fit a model on the observed CD4 values empirically to address measurement errors in the observed CD4 values. To extract individual-specific characteristics of CD4 trajectories, we may use an LME model.

Specifically, let z_{ij} be the observed CD4 value (in cells/mm^3) of individual i measured at time t_{ij}, $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m_i$. We may consider the following general LME model for CD4 data

$$z_{ij} = \mathbf{u}_{ij}^T \alpha + \mathbf{v}_{ij}^T \mathbf{a}_i + \epsilon_{ij} \equiv z_{ij}^* + \epsilon_{ij}, \quad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m_i, \tag{4.2}$$

where vectors $\mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ contain covariates including time, vector α contains fixed effect parameters, vector $\mathbf{a}_i$ contains random effects with $\mathbf{a}_i \sim N(0, A)$, z_{ij}^* is the (unobserved) true CD4 value whose corresponding observed error-prone value is z_{ij}, and ϵ_{ij} is the measurement error, with ϵ_{ij}'s are i.i.d. $\sim N(0, \delta^2)$. We may take appropriate transformations of the observed CD4 values, such as a log-transformation, or a $\sqrt{z_{ij}}$-transformation to make the normality assumption reasonable.

Note that the general LME model (4.2) includes nonparametric mixed effects models. They may be useful if the CD4 trajectories are complicated without clear patterns, since we can use a basis-based approach to approximate the nonparametric mixed model by an LME model (Wu, 2009). Thus, the general LME model (4.2) is quite flexible for modelling complex CD4 longitudinal data by choosing different fixed and random effects.

For the motivating dataset shown in Section 4.1, due to the lack of biologically justified mechanistic models, we considered several empirical poly-

nomial LME models for CD4 data during ART. We find that the following simple empirical LME model fits the CD4 data reasonably well

$$z_{ij} = \alpha_{1i} + \alpha_{2i} t_{ij} + \epsilon_{ij}, \qquad \alpha_{1i} = \alpha_1 + a_{1i}, \qquad \alpha_{2i} = \alpha_2 + a_{2i}, \tag{4.3}$$

where $\alpha = (\alpha_1, \alpha_2)^T$ are fixed effects and $\mathbf{a}_i = (a_{1i}, a_{2i})^T$ are random effects. The random effects a_{1i} and a_{2i} represent the individual-specific intercept and slope of the CD4 trajectory respectively. More complex models, such as a quadratic model

$$\begin{aligned} z_{ij} &= \alpha_{1i} + \alpha_{2i} t_{ij} + \alpha_{3i} t_{ij}^2 + \epsilon_{ij}, \\ \alpha_{1i} &= \alpha_1 + a_{1i}, \qquad \alpha_{2i} = \alpha_2 + a_{2i}, \qquad \alpha_{3i} = \alpha_3 + a_{3i}, \end{aligned}$$

do not appear to improve the fit significantly based on the likelihood ratio test. Hence, model (4.3) is used to fit the CD4 values, since the model fits the CD4 data well and easy to implement.

4.2.3 An NLME model for viral rebound following ART interruption

From Figure 4.1, we can see that the viral loads rebound quickly following ART interruption, and that viral loads after reaching peak points during rebound start to decrease slightly and exhibit large variations between subjects. There are many previously proposed nonlinear parametric models for HIV-1 viral load dynamics, but these models either work for the rise phase or the decay phase, followed by a possible rebound. Therefore, Wang et al. (2020) proposed a new parametric model for the current setting where the viral rebound after treatment interruption may rise rapidly to the peak point followed by a decrease to a viral set point. The new parametric model characterizes finer features of viral load rebound after ART interruption that does not require assumptions about mechanisms of viral dynamics. A flexible functional form is chosen to capture the shapes of viral rebound trajectories and to provide biological insights regarding the rebound process. Each parameter can incorporate a random effect to allow for variations in parameters across individuals. Some key features of viral rebound, such as rate of rise and set point, are represented by the parameters in the model. The NLME model proposed by Wang et al. (2020) is described as follows.

Let w_{ij} be the $\log_{10}$-transformed viral load value of individual i measured at time t^*_{ij}, $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n^*_i$, where the time t^*_{ij} is the time since ART interruption (not since start of ART). Following Wang et al.

(2020) (with minor modification), we consider the following NLME model for modelling viral rebounds following ART interruption

$$w_{ij} = \beta_{1i}\frac{t^*_{ij}}{t^*_{ij} + \exp(\beta_{2i} - \beta_{3i}t^*_{ij})} + \beta_{4i} + \xi_{ij}, \quad (4.4)$$

$$\beta_i = R_i\beta + \tau_i, \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n^*_i, \quad (4.5)$$

where vector $\beta_i = (\beta_{1i}, \cdots, \beta_{5i})^T$ contains individual-specific parameters, vector $\beta = (\beta_1, \cdots, \beta_q)^T$ contains fixed effect parameters, R_i is a $5 \times q$ design matrix contains covariates, and $\tau_i = (\tau_{1i}, \cdots, \tau_{5i})^T \sim N(0, G)$ contains random effects with G being a covariance matrix, and ξ_{ij} is within-individual random error. We assume that the random effects τ_i and the random error ξ_{ij} are independent, and ξ_{ij} are i.i.d. $\sim N(0, \omega^2)$.

Note that the above NLME model (4.4) - (4.5) may be viewed as a two-stage model: In stage 1, model (4.4) describes the viral rebound trajectories within an individual; and in stage 2, model (4.5) assumes that the between-individual variations in the individual-specific parameters in model (4.4) may be partially explained by covariates in R_i as well as random effects τ_i.

The parameters in NLME model (4.4) have the following attractive interpretations (Wang et al., 2020): parameter β_1 represents *set point* after a rebound, parameter β_2 and β_3 respectively represent first and second phases of the *rates of rise* in viral load during rebound, and parameter β_4 denotes *initial viral load value* at the start of a rebound. Figure 4.3 shows the viral load rebound profile for a typical subject based on model (4.4). Each of the four parameters denotes an important characteristic of the viral rebound trajectories following ART interruption.

Recall that our main objective is to exam if key features of viral load or CD4 trajectories during ART are associated with important characteristics of viral rebounds following ART interruption. We may use the second-stage model (4.5) to evaluate such possible associations. Note that the random effects $\mathbf{b}_i$ in NLME model (4.1) for viral load data during ART may be viewed as individual-specific characteristics of the viral load trajectories during ART. Thus, we may use the random effects $\mathbf{b}_i$ as "covariates" in the rebound model (4.5) to see if these "covariates" may partially explain the large variations in the individual-specific parameters β_i during viral rebound. Similarly, we may also consider the random effects $\mathbf{a}_i$ in the CD4 model (4.3) during ART, and use them as possible "covariates" in the NLME model (4.5) for viral load following ART interruption.

Specifically, in the NLME model (4.4) for viral rebound, we may consider

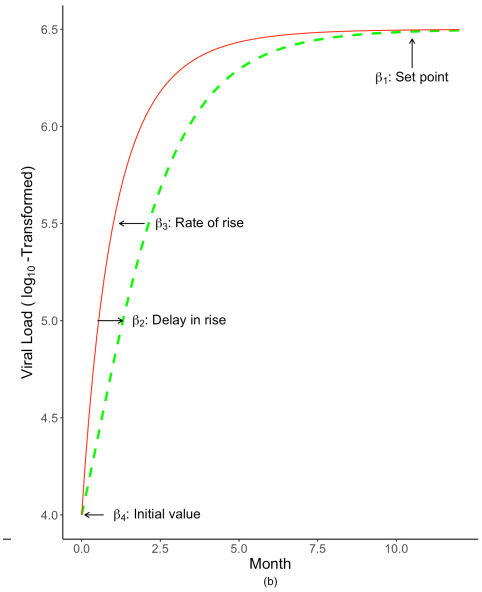

Figure 4.3: Typical viral rebound profiles following ART interruption based on model (4.4).

the following second-stage model for (4.5)

$$\begin{aligned} \beta_{ki} &= \beta_k + \gamma_{k1} b_{1i} + \gamma_{k2} b_{2i} + \gamma_{k3} b_{3i} + \gamma_{k4} b_{4i} \\ &\quad + \gamma_{k5} a_{1i} + \gamma_{k6} a_{2i} + \gamma_{k7} \mathbf{v}_i + \tau_{ki}, \\ \beta_{ji} &= \beta_j + \tau_{ji} \qquad\qquad j \neq k, \quad k = 1, \cdots, 4, \quad i = 1, \cdots, n, \end{aligned} \tag{4.6}$$

where $\mathbf{a}_i = (a_{1i}, a_{2i})$ are the random effects in the CD4 model (4.3), $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})$ are the random effects in the viral decay model (4.1) during ART, β_k's are fixed effects parameters, γ_{kl}'s are fixed effect parameters associated with the corresponding random effects respectively, and $\mathbf{v}_i$ denote other baseline covariates. For example, if $k = 1$, testing $H_0 : \gamma_{13} = 0$ versus $H_1 : \gamma_{13} \neq 0$ allows us to check possible association between b_{3i} and β_{1i}. If the null hypothesis is rejected, we can conclude that there is a significant association between viral decay rate during ART and viral set point following ART interruption.

4.3 A Three-Step (TS) method

As reviewed in Section 2.4, the SAEM method performs well in NLME models with left censoring. Although the SAEM algorithm produces less

accurate estimates than the linearization method, it is much more computationally efficient. Therefore, SAEM method may be used for parameter estimations in the NLME model (4.1) for viral decay during ART, as well as in the NLME model (4.4) and (4.5) for viral rebound data following ART interruption. In fact, we may consider the SAEM algorithm for all three models simultaneously based on the joint likelihood of all observed data. However, such a joint likelihood method may be computationally extremely intensive, since the dimension of the "missing data" $(\mathbf{y}_{cen,i}, \mathbf{w}_{cen,i}, \mathbf{b}_i, \tau_i, \mathbf{a}_i)$ is very high, so even a single simulation using an MCMC method may be computationally over-whelming. Note that $\mathbf{w}_{cen,i}$ denotes the censored components of $\mathbf{w}_i = (w_{i1}, w_{i2}, \cdots, w_{in_i^*})^T$. Separate inference, such as the naive two-step method, may also be applied to parameter estimations in joint modelling. However, the naive two-step method may lead to biased estimations, since the uncertainty of estimation in the first step is not incorporated in the second step. Details about the joint likelihood method and the naive two-step method are discussed in Section 3.1. Since the above two methods have some drawbacks, we propose to use SAEM algorithm for each NLME model with left censoring separately to reduce the computation burden, so-called a *three-step (TS) method*.

We have discussed the TS method in Section 3.2. Here we make minor modifications to the TS method for the specific NLME models for the HIV data as described above.

- Step 1: For data during ART, fit the NLME model (4.1) for viral load with censoring using the SAEM algorithm and fit the LME model for CD4 using the standard method respectively, and then obtain the maximum likelihood estimates (MLEs) of the fixed parameters and the empirical Bayes estimates of the random effects, $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$, respectively;

- Step 2: For viral rebound data following ART interruption, fit the NLME model (4.4) with left censoring using the SAEM algorithm, with the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ in the second-stage model (4.6) substituted by their empirical Bayes estimates $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$ from Step 1;

- Step 3: Obtain the standard errors of the parameter estimates based on a (parametric) bootstrap method, which incorporates the estimation uncertainty of random effect estimates $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$ in Step 1.

The parametric bootstrap method works as follows:

1. Simulate CD4 and viral load data with left censoring, based on the fitted LME and two NLME models using the above three-step method, where the model parameters are replaced by their estimates.

2. For the simulated CD4 and viral load data, fit all three models again using the above three-step method and obtain all parameter estimates.

3. Repeat the above process B times (say, $B = 100$), we obtain B estimates for each parameter. The sample standard deviation of these B estimates of each parameter is the parametric *bootstrap estimate* of the standard error of the corresponding parameter estimate.

The above bootstrap method incorporates the estimation uncertainty of the parameter and random effect estimates in the TS method with separate model fitting, so it should produce more reliable standard errors of the parameter estimates than those from separate model fitting.

4.4 Data Analysis Results

In this section, we analyze the dataset shown in Figure 4.1 using the proposed TS method and a naive (NV) method, which still uses the SAEM algorithm but the censored values are substituted by half the detection limit and without bootstrap. The goal of data analysis is to study if key features of viral decay during ART are associated with individual-specific characteristics of viral rebound following ART interruption.

There are 75 patients in the study. Viral loads and CD4 are repeatedly measured on patients during ART and following ART interruption. After ART interruption, viral load usually increases to a peak within 6–10 weeks, then decreases to a stable level over a time scale of months. Therefore, we restrict our attention to data within week 36 (9 months) after interruption of ART. From Figure 4.1, we see that the viral load trajectories during ART exhibit clear patterns of viral decay. Following ART interruption, the viral loads rebound quickly, but their trajectories become complicated after reaching peak points, with substantial between-subject variations. We excluded individuals with 2 or less repeated measurements either during ART or following ART interruption ($n = 2$) or individuals with unclear viral load patterns following ART interruption ($n = 8$). Some facts about the initial and remaining dataset are summarized in Table 4.1, including the duration of the study, number of repeated measurements, and the proportion of censored measurements during ART and following ART interruption for the initial dataset and the remaining dataset after removing the outliers.

Table 4.1: Summary of repeated measurements in the initial HIV dataset and the remaining dataset after removing outliers.

		Initial dataset			Remaining dataset		
		during	following	overall	during	following	overall
Duration (in months)	min	6.97	4.37	16.77	7.83	4.37	16.73
	max	47.97	87.9	107.77	47.97	14.33	58.4
	mean	18.52	34.20	59.19	18.55	9.08	27.46
Number of repeated measurement	min	5	1	9	5	3	10
	max	20	23	48	20	23	34
	mean	9.72	6.41	24.64	10	6.49	16.5
Proportion of censored measurements (%)		63.5%	25.2%	32.7%	62.6%	27.7%	48.9%

Note: "during", "following", and "overall" stand for during ART, following ART interruption, and the entire dataset.

Model fitting was performed using the R package "`saemix`" (Comets et al., 2017). The "`saemix`" package implements the SAEM algorithm for parameter estimation in (non)linear mixed effects models. It computes the maximum likelihood estimator of the population parameters, without any approximation of the model such as linearization, using the SAEM algorithm and provides standard errors for the maximum likelihood estimator.

For viral load data during ART, we fit the NLME model (4.1) of Wu and Ding (1999), with left-censored data addressed by the SAEM method. The bi-exponential decay NLME model (4.1) fits the viral load data very well. Figure 4.5(A) (top four figures) shows the fitted values versus the corresponding observed values for four randomly selected subjects during ART.

For the CD4 data during ART, the CD4 trajectories do not appear to exhibit clear patterns, as shown in Figure 4.4 (A). CD4 values show large between-subject variations, which possibly due to substantial measurement errors. However, there seems an overall upward trend, as shown in Figure 4.4 (B). Thus, we fit the LME model (4.3) to the CD4 data, which at least captures a rough upward trend before ART interruption. A more complex model, such as a quadratic model, does not seem to improve the fit.

For viral load data following ART interruption, we consider the following

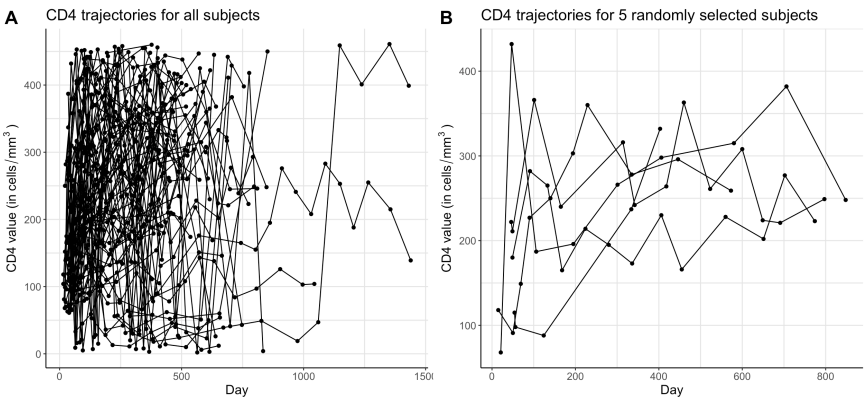

Figure 4.4: Figure (A) shows CD4 value (in cells/mm^3) trajectories for all subjects, and Figure (B) shows CD4 value trajectories for 5 randomly selected subjects.

NLME model

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij}+\exp(\beta_{2i}-\beta_{3i}t_{ij})}+\beta_{4i}+\xi_{ij}, \\ \beta_{ki} &= \beta_k+\gamma_{k1}b_{1i}+\gamma_{k2}b_{2i}+\gamma_{k3}b_{3i}+\gamma_{k4}b_{4i}+\gamma_{k5}a_{1i}+\gamma_{k6}a_{2i}+\tau_{ki}, \\ \beta_{ji} &= \beta_j+\tau_{ji}, \quad k=1,2,3,4, \quad j\neq k, \end{aligned} \tag{4.7}$$

where β_{ki} is the individual-specific viral rebound feature following ART interruption that we are interested in, and the random effects $(\mathbf{a}_i, \mathbf{b}_i)$ are defined in model (4.6) (e.g., b_{3i} is the random effect associated with the initial viral decay rate λ_{1i} during viral decay before ART interruption). For example, to study the association between individual-specific characteristics of viral decay during ART and *viral set point* following ART interruption, we may consider the NLME model:

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij}+\exp(\beta_{2i}-\beta_{3i}t_{ij})}+\beta_{4i}+\xi_{ij}, \\ \beta_{1i} &= \beta_1+\gamma_{11}b_{1i}+\gamma_{12}b_{2i}+\gamma_{13}b_{3i}+\gamma_{14}b_{4i}+\gamma_{15}a_{1i}+\gamma_{16}a_{2i}+\tau_{1i}, \\ \beta_{ji} &= \beta_j+\tau_{ji}, \; j>1, \end{aligned} \tag{4.8}$$

where β_{1i} represents the viral set point during rebound.

We again address left-censored viral rebound data by the SAEM method. We use the parametric bootstrap method with $B = 100$ to estimate the standard errors of all the fixed effect parameter estimates. The NLME model (4.8) for viral rebound also fits the data reasonably well. Figure 4.5 (B) (bottom four figures) shows the fitted values versus the corresponding observed values for four randomly selected subjects during viral rebound.

Table 4.2: Parameter estimates with a second-stage model for setpoint β_{1i}.

Parameter	Estimate	Naive SE	Bootstrap SE	z-value	p-value
P_1	11.258	0.283	0.167	67.449	0.000
λ_1	4.790	0.380	0.362	13.244	0.000
P_2	3.271	0.117	0.206	15.872	0.000
α_1	0.087	0.007	0.017	5.178	0.000
α_2	24.080	0.517	0.438	54.985	0.000
λ_2	0.225	0.027	0.019	11.788	0.000
β_1	2.828	0.161	0.243	11.651	0.000
γ_{11}	0.144	0.073	0.075	1.911	0.056
γ_{12}	-0.027	0.559	0.755	-0.035	0.972
γ_{13}	-0.252	0.079	0.116	-2.177	0.029
γ_{14}	-97.574	50.944	26.294	-3.711	0.000
γ_{15}	-0.026	0.033	0.035	-0.728	0.467
γ_{16}	0.811	1.290	4.291	0.189	0.850
β_2	1.588	1.308	0.694	2.289	0.022
β_3	3.360	1.614	0.828	4.056	0.000
β_4	0.783	0.119	0.151	5.195	0.000

Note: Naive SE is the standard error based on separate model fitting without bootstrap, z-value is the ratio of Estimate/Bootstrap SE, and p-value is based on the z-value and the standard normal tail probability for a two-sided test.

Table 4.2 shows parameter estimation results for models (4.1), (4.3), and (4.8) (the unit of time in data analysis is month). We see that parameters γ_{13} and γ_{14}, which link the initial viral decay rates λ_{1i} and the second-phase viral decay rates λ_{2i} (or b_{i3} or b_{i4}) during ART to viral setpoints β_{1i} following ART interruption for individual i, appears to be statistically significant at 5% level. In other words, the initial viral decay rate during ART appears to be negatively associated with the viral setpoints following ART interruption: the faster the viral decay after the start of ART, the lower the setpoints following ART interruption. In addition, the second-phase viral decay rates during ART also appear to be negatively associated with the viral setpoints following ART interruption. The naive method produces similar estimates but different standard errors, so we only show naive SE in Table 4.2. We will evaluate the two methods via simulation in the next chapter.

As another example, we may also consider the following two-stage model to study the association between individual-specific features of viral decay

Table 4.3: Parameter estimates with a second-stage model for rebound rate β_{3i}.

Parameter	Estimate	Naive SE	Bootstrap SE	z-value	p-value
P_1	11.258	0.283	0.204	55.176	0.000
λ_1	4.790	0.380	0.236	20.321	0.000
P_2	3.271	0.117	0.167	19.575	0.000
α_1	0.087	0.007	0.011	7.765	0.000
α_2	24.080	0.517	0.518	46.456	0.000
λ_2	0.225	0.027	0.019	11.627	0.000
β_1	2.946	0.375	0.127	23.196	0.000
β_2	1.542	1.040	0.419	3.680	0.000
β_3	3.280	–	0.448	7.327	0.000
γ_{31}	0.837	0.053	0.211	3.971	0.000
γ_{32}	-4.589	0.177	2.691	-1.705	0.088
γ_{33}	-1.110	0.008	0.317	-3.501	0.000
γ_{34}	-221.564	12.712	116.409	-1.903	0.057
γ_{35}	-0.106	0.023	0.109	-0.969	0.333
γ_{36}	-0.090	0.445	8.458	-0.011	0.991
β_4	0.772	0.448	0.057	13.625	0.000

Note: Naive SE is the standard error based on separate model fitting without bootstrap, z-value is the ratio of Estimate/Bootstrap SE, and p-value is based on the z-value and the standard normal tail probability for a two-sided test.

and *viral rate of rise* following ART interruption:

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij}+\exp(\beta_{2i}-\beta_{3i}t_{ij})}+\beta_{4i}+\xi_{ij}, \\ \beta_{3i} &= \beta_3+\gamma_{31}b_{1i}+\gamma_{32}b_{2i}+\gamma_{33}b_{3i}+\gamma_{34}b_{4i}+\gamma_{35}a_{1i}+\gamma_{36}a_{2i}+\tau_{3i}, \\ \beta_{ji} &= \beta_j+\tau_{ji},\ j\neq 3, \end{aligned} \tag{4.9}$$

where β_{3i} represents the rate of viral rise during rebound. The analysis results are presented in Table 4.3. We see that the rate of rise during viral rebound following ART interruption (β_{3i}) appears to be negatively correlated with initial viral load values during ART (γ_{31}) and initial viral decay rate (γ_{33}). That is, the higher the initial viral loads during ART, or the faster the initial viral decline during ART, the slower the viral rising following ART interruption. Note that some standard errors for the population parameter are not available. It may be because the model is too complex and the parameters are unidentifiable.

Similar analysis can be done by linking the random effects in model

Table 4.4: Parameter estimates with a second-stage model for the delay in rise β_{2i}.

Parameter	Estimate	Naive SE	Bootstrap SE	z-value	p-value
P_1	17.179	0.504	0.294	58.488	0.000
λ_1	3.922	0.269	0.183	21.381	0.000
P_2	2.460	0.184	0.260	9.444	0.000
α_1	0.036	0.013	0.014	2.519	0.012
α_2	23.886	0.549	0.431	55.364	0.000
λ_2	0.217	0.027	0.025	8.657	0.000
β_1	3.243	0.187	0.180	18.023	0.000
β_2	7.609	1.277	1.193	6.375	0.000
γ_1	1.437	0.911	0.714	2.013	0.044
γ_2	10.318	9.960	6.881	1.500	0.134
γ_3	0.869	0.629	0.728	1.195	0.232
γ_4	117.176	68.364	224.765	0.521	0.602
γ_5	0.134	0.398	0.164	0.817	0.414
γ_6	3.416	11.115	13.095	0.261	0.794
β_3	2.919	0.461	0.461	6.327	0.000
β_4	0.437	0.244	0.078	5.613	0.000

Note: Naive SE is the standard error based on separate model fitting without bootstrap, z-value is the ratio of Estimate/Bootstrap SE, and p-value is based on the z-value and the standard normal tail probability for a two-sided test.

(4.1) and (4.3) to the delay of rise during rebound β_{2i} and the initial viral load value at the start of rebound β_{4i} separately. The analysis results are presented in Table 4.4 and Table 4.5. From the tables, we can see that the higher the initial viral loads during ART, the lower the initial value and more delay in rise following ART interruption. In addition, the faster the viral decay rate after the start of ART, the lower the initial viral load following ART interruption.

In summary, the analysis results show that some key characteristics of the viral load trajectories during ART, especially the initial viral decay rates after the start of ART, appear to be associated with some important features of the viral rebound following ART interruption, such as viral setpoints and initial viral loads. Note that CD4 data during ART seems not associated with important features of the viral rebound following ART interruption, which may occur due to large measurement error. To further compare and evaluate the performances of the TS method and the naive method, a simulation study is conducted in the next chapter.

Table 4.5: Parameter estimates with a second-stage model for initial value β_{4i}.

Parameter	Estimate	Naive SE	Bootstrap SE	z-value	p-value
P_1	17.179	0.504	0.515	33.326	0.000
λ_1	3.922	0.269	0.168	23.300	0.000
P_2	2.460	0.184	0.217	11.319	0.000
α_1	0.036	0.013	0.015	2.400	0.016
α_2	23.886	0.549	0.348	68.727	0.000
λ_2	0.217	0.027	0.022	10.020	0.000
β_1	3.045	0.118	0.129	23.531	0.000
β_2	16.906	2.859	3.797	4.452	0.000
β_3	5.931	1.096	1.203	4.929	0.000
β_4	0.729	0.003	0.060	12.243	0.000
γ_1	-0.190	0.003	0.082	2.327	0.020
γ_2	-0.796	0.020	1.056	0.754	0.451
γ_3	-0.158	0.002	0.070	2.253	0.024
γ_4	-28.549	0.136	24.028	1.188	0.235
γ_5	-0.037	0.001	0.085	0.432	0.666
γ_6	-0.358	0.027	10.958	0.033	0.974

Note: Naive SE is the standard error based on separate model fitting without bootstrap, z-value is the ratio of Estimate/Bootstrap SE, and p-value is based on the z-value and the standard normal tail probability for a two-sided test.

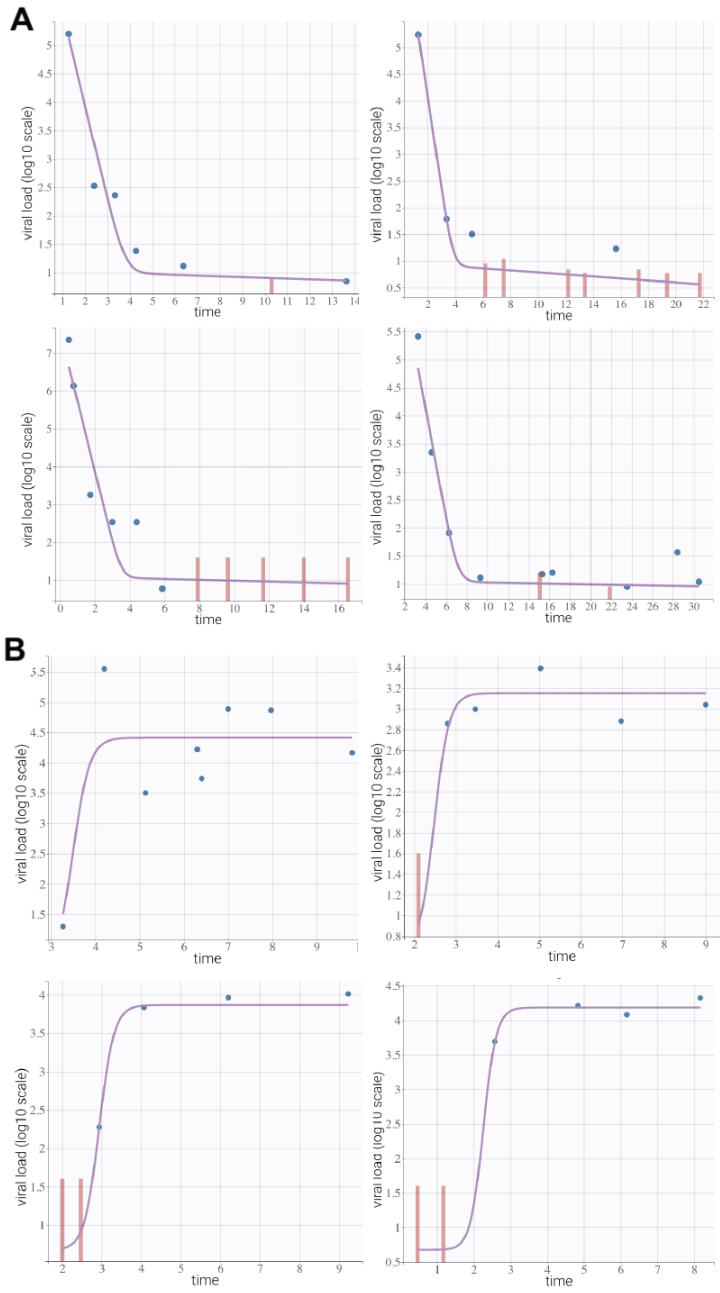

Figure 4.5: Observed and fitted viral load trajectories before (top four figures) and following (bottom four figures) ART interruption for 4 randomly selected subjects respectively. The red vertical bars represent left-censored viral loads.

Chapter 5

A Simulation Study

In this chapter, we conduct extensive simulations to evaluate the proposed TS method and compare it with the naive method used in data analysis. We choose similar NLME models to those in the data analysis section, but we omit the CD4 model for simplicity. The true values of the model parameters are set to be similar to those estimated in the data analysis section.

The viral load data during ART are generated based on the following NLME model

$$\begin{aligned} y_{ij} &= \log_{10}(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}) + e_{ij}, \qquad (5.1) \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_1 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \end{aligned}$$

where e_{ij} i.i.d. $\sim N(0, \sigma_1^2)$ and $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T \sim N(0, D)$. The viral load data following ART interruption are generated based on the following NLME model

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij} + \exp(\beta_{2i} - \beta_{3i}t_{ij})} + \beta_{4i} + \xi_{ij}, \qquad (5.2) \\ \beta_{1i} &= \beta_1 + b_{3i}\gamma_3 + \tau_{1i}, \quad \beta_{2i} = \beta_2 + \tau_{2i}, \quad \beta_{3i} = \beta_3 + \tau_{3i}, \quad \beta_{4i} = \beta_4 + \tau_{4i}, \end{aligned}$$

where b_{3i} is the random effect from model (5.1), $\xi_{ij} \sim N(0, \sigma_3^2)$, and $\tau_i \sim N(0, G)$.

The performance of MLEs of mixed effects models may depend on the number of repeated measurements, the variations of data, the sample size, and the detection limit. Intuitively, when there are more frequent repeated measurements, more information about the longitudinal covariate process is provided. Both of the TS method and the naive method may perform better. When the variations of data are decreased (both between-individual variations and within-individual variations), the difference between subjects and the difference between repeated measurements within each subject are small. The naive method and the TS method may have similar performance, since the impact of ignoring the uncertainty of censored values and separate model fitting is small. A larger sample size may generally result in better

estimations, since more information about the longitudinal covariate process is provided. When the detection limit is increased, the proportion of censored values also increase, and the naive method might perform poorly.

Since the performance of MLEs of mixed effects models may depend on the above conditions, we conduct several simulation studies by considering different settings to compare the performance of the naive method and the proposed TS method. Setting I is the baseline setting, which is closest to the situation of the real data set. We then change some features in Setting I to mimic the foregoing conditions. For example, in Setting II, we consider more frequent repeated measurements. In Setting III, we consider smaller variations of data. In Setting IV, the sample size is increased. In Setting V, the detection limit is increased.

We evaluate the performance of the proposed TS method and the naive method based on bias, mean square error (MSE), and coverage rates of 95% confidence intervals. For a parameter β and its estimate $\hat{\beta}$, the bias and root MSE (rMSE) are defined as $bias = \hat{\beta} - \beta$, $rMSE = \sqrt{MSE}$, and the coverage rate is the proportion of confidence intervals that cover the true value. We will also compare the TS method to a naive method which replaces censored viral loads by half the detection limits without bootstrapping. For the TS method, the number of bootstrap samples is $B = 100$. The simulations are repeated 100 times. While a larger number of repetitions may be desirable, the bootstrap procedure may be computationally expensive (Morris et al., 2019). The simulation results show that 100 repetitions seem sufficient for us to make reasonable conclusions.

In the tables displaying the results, for each parameter,"Estimate" is the average of the 100 estimates from the 100 simulation repetitions. For the TS method, "SE" is the average of the 100 bootstrap standard errors. Note that for each simulation run, we generate 100 bootstrap samples, and the bootstrap SE is the standard deviation of the 100 estimates from the 100 bootstrap samples. For the naive method, "SE" is the average of the 100 standard errors of the estimates from the 100 simulations. "Bias" is the difference between the "Estimate" and the corresponding true value. "rMSE" is square root of MSE, where MSE is calculated from the sum of the corresponding squared "SE" and the squared "Bias", divided by the corresponding true value. "Coverage" is the percentage of the confidence intervals containing the true value among the 100 confidence intervals from the 100 simulation repetitions.

5.1 Setting I: Baseline

In Setting I, most of the true values are chosen to be similar to those in the data analysis section. Specifically, the sample size is set to be $N = 50$ individuals. For the within-individual longitudinal measurements, to mimic the real dataset, for half the sample, we choose 10 repeated measurements during ART and 9 repeated measurements following ART interruption, while for the remaining half the sample, we choose 11 repeated measurements during ART and 11 repeated measurements following ART interruption. The measurement times are chosen to be similar to those in the real dataset. Two sets of measurement times during ART are $t_1 = (0.5, 1.6, 2.3, 3, 4.6, 6.5, 7.6, 11.2, 14.9, 19.1)$ and $t_2 = (0.5, 0.7, 1.7, 3, 4.4, 5.9, 7.9, 9.6, 11.7, 14, 16.5)$. Two sets of measurement times following ART interruption are $t_1^* = (0.1, 1.1, 1.6, 2.4, 2.8, 3.3, 3.7, 4.2, 5.2)$ and $t_2^* = (0.2, 0.6, 1.1, 1.6, 2.1, 2.5, 3, 3.5, 4, 4.4, 4.9)$.

For model (5.1), some true values are $P_1 = 17.0$, $P_2 = 2.6$, $\lambda_1 = 4$, $\lambda_2 = 0.05$, and $\sigma_1 = 0.5$. The detection limit is set to be $d = 1.60$. For model (5.2), some true parameter values are $\beta_1 = 3.2$, $\beta_2 = 5.6$, $\beta_3 = 10$, $\beta_4 = 1$, $\gamma_3 = 1$, $\sigma_3 = 0.5$,

$$D = \begin{bmatrix} 1.7 & -0.4 & 0.06 & -0.003 \\ -0.4 & 1.5 & -0.1 & 0.005 \\ 0.06 & -0.1 & 0.05 & -0.002 \\ -0.003 & 0.005 & -0.002 & 0.0002 \end{bmatrix}, \text{ and } G = \begin{bmatrix} 0.5 & 0.03 & 0.2 & 0.03 \\ 0.03 & 2.3 & -0.3 & -0.04 \\ 0.2 & -0.3 & 11.8 & 0.06 \\ 0.03 & -0.04 & 0.06 & 0.006 \end{bmatrix}.$$

The simulation results of Setting I are shown in Table 5.1. We can see that the proposed TS method performs quite well and shows a better performance than the naive method in terms of the coverage probabilities and the biases. Estimates based on the TS method are approximately unbiased as the estimated coverage probabilities are close to the nominal level 0.95. However, estimates based on the naive method may sometimes produce biased results as some of the estimated coverage probabilities are way below the nominal level 0.95. The reason for the biased results based on the naive method may be the under-estimated SE's as well as biased estimates. Note that the naive method ignores the uncertainties of the censored values and the separate NLME model fitting, as the censored values are imputed by half of the detection limit and the random effect $\mathbf{b}_i$ in the second step are substituted by their empirical Bayes estimates $\hat{\mathbf{b}}_i$. As a result, the SE's based on the naive method may be under-estimated, and the naive method may lead to smaller MSE's but low coverage probabilities. On the other hand,

Table 5.1: Simulation results for Setting I.

Parameter	True value	Method	Estimate	SE	Bias	rMSE	Coverage
P_1	17.0	TS	17.097	0.425	0.097	0.435	0.95
		Naive	17.088	0.257	0.088	0.272	0.91
λ_1	4.0	TS	4.092	0.240	0.092	0.257	0.91
		Naive	4.137	0.216	0.137	0.256	0.91
P_2	2.6	TS	2.794	0.398	0.194	0.442	0.95
		Naive	2.222	0.111	-0.378	0.394	0.39
λ_2	0.1	TS	0.040	0.081	-0.010	0.081	0.95
		Naive	0.029	0.009	-0.021	0.023	0.40
β_1	3.2	TS	3.267	0.140	0.067	0.155	0.95
		Naive	3.324	0.130	0.124	0.180	0.86
γ_3	1.0	TS	0.986	0.096	-0.014	0.097	0.94
		Naive	0.999	0.083	-0.001	0.083	0.92
β_2	5.6	TS	5.588	1.187	-0.012	1.187	0.94
		Naive	6.370	1.354	0.770	1.558	0.84
β_3	10.0	TS	10.079	2.062	0.079	2.064	0.94
		Naive	11.170	2.650	1.170	2.897	0.88
β_4	1.0	TS	0.930	0.092	-0.070	0.115	0.85
		Naive	0.861	0.072	-0.139	0.157	0.62

the proposed TS method incorporates the uncertainties of the censored values by the SAEM algorithm and the uncertainties of separate NLME model fitting by bootstrap, so it may lead to larger MSE's but correct coverage probabilities.

5.2 Setting II: More Frequent Repeated Measurements

Since the performance of MLEs of mixed effects models may depend on the number of repeated measurements, we conduct another simulation study by choosing more frequent repeated measurements, with other true parameter values remain the same. The measurement times are chosen to be close to those in the real dataset with additional measurement times in between. Specifically, the new sets of measurement times during ART are chosen to be $t_1 = (0.4, 1.2, 1.6, 2.1, 3.2, 4.6, 5.3, 7.8, 10.4, 13.4, 17)$, and $t_2 = (0.4, 0.5, 1.2, 2.1, 3.1, 4.1, 5.5, 6.7, 8.2, 9.8, 11.6, 13.5, 15.5, 18)$. The new

Table 5.2: Simulation results for more frequent repeated measurements.

Parameter	True value	Method	Estimate	SE	Bias	rMSE	Coverage
P_1	17.0	TS	16.820	0.733	-0.180	0.755	0.93
		Naive	16.981	0.259	-0.019	0.260	0.89
λ_1	4.0	TS	4.011	0.288	0.011	0.288	0.92
		Naive	4.103	0.202	0.103	0.227	0.85
P_2	2.6	TS	2.968	0.705	0.368	0.796	0.95
		Naive	2.273	0.142	-0.327	0.357	0.51
λ_2	0.1	TS	0.096	0.186	0.046	0.191	0.95
		Naive	0.041	0.012	-0.009	0.015	0.34
β_1	3.2	TS	3.223	0.132	0.023	0.134	0.94
		Naive	3.304	0.121	0.104	0.159	0.87
γ_3	1.0	TS	0.986	0.095	-0.014	0.096	0.97
		Naive	0.996	0.089	-0.004	0.089	0.96
β_2	5.6	TS	5.553	0.873	-0.047	0.874	0.95
		Naive	6.258	0.676	0.658	0.943	0.79
β_3	10.0	TS	9.757	1.460	-0.243	1.481	0.96
		Naive	10.715	1.166	0.715	1.368	0.82
β_4	1.0	TS	0.980	0.079	-0.020	0.081	0.88
		Naive	0.886	0.048	-0.114	0.124	0.63

sets of measurement times following ART interruption are chosen to be $t_1^* = (0, 0.6, 0.8, 1.2, 1.4, 1.7, 1.9, 2.1, 2.6, 3.1, 3.8, 4.5, 5.4, 6)$ and $t_2^* = (0, 0.3, 0.6, 0.8, 1.0, 1.3, 1.5, 1.8, 2.0, 2.2, 2.5, 2.9, 3.3, 3.7, 4.1, 4.5, 5.0, 5.5, 6.0, 6.8, 7.6)$. In Setting II, n_i increases to 11 and 14 during ART, and increases to 14 and 21 following ART interruption. The simulation results for more frequent repeated measurements are shown in Table 5.2. We can see that the proposed TS method performs better than the naive method, and the two methods seem to perform roughly the same as in Setting I.

Since both the TS method and the naive method require large within-individual repeated measurements to perform well, and more frequent repeated measurements provide more information about the longitudinal covariate process, so these two methods may perform well when there are more frequent repeated measurements. Compared to Table 5.1, the two methods do not seem to perform better when there are more frequent repeated measurements, probably because in Setting I, the number of repeated measurements is large enough to produce good results.

5.3 Setting III: Smaller Variations of Data

We also consider another different setting to compare the performance of the naive method and the proposed TS method by reducing variations of data, including both within-individual variations and between-individual variations. In Setting III, σ_1 in model (5.1) decreases from 0.5 to 0.2, and σ_3 in model (5.2) decreases from 0.5 to 0.2. The covariance matrices D and G are simplified to diagonal matrices

$$D = \begin{bmatrix} 1.7 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0.0001 \end{bmatrix}, \text{ and } G = \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0.0003 \end{bmatrix}.$$

Table 5.3 shows the simulation results by choosing smaller variations of data, with other true parameter values remain the same. The naive method and the proposed TS method do not show a better performance than Setting I in terms of the coverage probabilities. In general, when the between-individual variations and within-individual variations are decreased, the difference between subjects and the difference between repeated measurements within each subject is small. A key disadvantage of the naive method is that it does not incorporate the uncertainties in the estimation in the first step. If the variations of data are small, ignoring the uncertainties in the estimation may have less impact on the results, so the performance of the naive method and the TS method may be close to each other. In Table 5.3, the performances of the two methods are somewhat different, and the proposed TS method has a better performance than the naive method.

5.4 Setting IV: Increased Sample Size

In Setting IV, we simulate data with a larger number of individuals $N = 200$ to check how sample size affects parameter estimations in the naive method and the proposed TS method. Table 5.4 shows the simulation results for Setting IV with other parameter values remain the same as in Setting I. Compared to the simulation results in Table 5.1, both the naive method and the proposed TS method produce slightly lower coverage probabilities. In general, a larger sample size leads to better estimations. However, a larger sample size may lead to more accurate estimations of parameters and standard errors, making the differences between the methods more obvious. Overall, the TS method outperforms the naive method.

Table 5.3: Simulation results for smaller variations of data.

Parameter	True value	Method	Estimate	SE	Bias	rMSE	Coverage
P_1	17.0	TS	16.955	0.393	-0.045	0.396	0.94
		Naive	17.061	0.207	0.061	0.216	0.96
λ_1	4.0	TS	3.990	0.147	-0.010	0.148	0.96
		Naive	4.095	0.115	0.095	0.149	0.97
P_2	2.6	TS	2.327	0.587	-0.273	0.647	0.88
		Naive	1.830	0.040	-0.770	0.772	0.10
λ_2	0.1	TS	0.043	0.095	-0.007	0.095	0.94
		Naive	0.011	0.004	-0.039	0.040	0.09
β_1	3.2	TS	3.453	0.131	0.253	0.285	0.55
		Naive	3.409	0.075	0.209	0.222	0.22
γ_3	1.0	TS	1.007	0.116	0.007	0.116	0.91
		Naive	1.009	0.095	0.009	0.096	0.91
β_2	5.6	TS	3.928	0.607	-1.672	1.779	0.23
		Naive	5.341	0.459	-0.259	0.527	0.13
β_3	10.0	TS	7.704	0.945	-2.296	2.483	0.30
		Naive	9.560	0.689	-0.440	0.818	0.21
β_4	1.0	TS	0.754	0.110	-0.246	0.269	0.38
		Naive	0.795	0.021	-0.205	0.206	0.03

Table 5.4: Simulation results for increased sample size.

Parameter	True value	Method	Estimate	SE	Bias	rMSE	Coverage
P_1	17.0	TS	16.981	0.465	-0.019	0.465	0.94
		Naive	17.011	0.130	0.011	0.130	0.89
λ_1	4.0	TS	4.033	0.172	0.033	0.176	0.97
		Naive	4.092	0.105	0.092	0.140	0.94
P_2	2.6	TS	2.776	0.447	0.176	0.480	0.95
		Naive	2.218	0.039	-0.382	0.384	0.24
λ_2	0.1	TS	0.048	0.112	-0.002	0.112	0.84
		Naive	0.033	0.005	-0.017	0.017	0.08
β_1	3.2	TS	3.259	0.072	0.059	0.093	0.86
		Naive	3.331	0.060	0.131	0.144	0.81
γ_3	1.0	TS	0.989	0.047	-0.011	0.048	0.97
		Naive	0.998	0.042	-0.002	0.042	0.94
β_2	5.6	TS	4.843	0.558	-0.757	0.940	0.70
		Naive	5.791	0.442	0.191	0.481	0.60
β_3	10.0	TS	8.772	0.874	-1.228	1.507	0.67
		Naive	10.090	0.701	0.090	0.707	0.54
β_4	1.0	TS	0.960	0.048	-0.040	0.063	0.86
		Naive	0.875	0.027	-0.125	0.127	0.54

Table 5.5: Simulation results for increased detection limit.

Parameter	True value	Method	Estimate	SE	Bias	rMSE	Coverage
P_1	17.0	TS	17.056	0.426	0.056	0.430	1.00
		Naive	17.076	0.266	0.076	0.277	0.86
λ_1	4.0	TS	4.028	0.237	0.028	0.239	0.94
		Naive	4.152	0.217	0.152	0.264	0.83
P_2	2.6	TS	2.880	0.447	0.280	0.527	0.96
		Naive	2.391	0.096	-0.209	0.229	0.05
λ_2	0.1	TS	0.037	0.083	-0.013	0.084	1.00
		Naive	0.037	0.010	-0.013	0.016	0.00
β_1	3.2	TS	3.252	0.145	0.052	0.154	0.90
		Naive	3.267	0.128	0.067	0.144	0.91
γ_3	1.0	TS	0.995	0.097	-0.005	0.097	0.97
		Naive	1.129	12.969	0.129	12.970	0.86
β_2	5.6	TS	5.641	1.190	0.041	1.190	0.97
		Naive	6.940	1.546	1.340	2.046	0.91
β_3	10.0	TS	10.023	2.024	0.023	2.025	0.95
		Naive	11.750	2.904	1.750	3.391	0.92
β_4	1.0	TS	0.964	0.099	-0.036	0.105	0.94
		Naive	0.931	0.061	-0.069	0.092	0.91

5.5 Setting V: Increased Detection Limit

In Setting V, we compare the simulation results between the naive method and the proposed TS method by increasing the detection limit from 1.6 to 1.9. Since the censored values are imputed by half of the detection limit in the naive method, increasing the detection limit may affect the simulation results because the proportion of censored values increases. Table 5.5 shows the simulation results for Setting V, with other parameter values remain the same as in Setting I.

In Table 5.5, we can see that the proposed TS method still performs better than the naive method in terms of the coverage probabilities. Note that the censored values are imputed by half of the detection limit in the naive method, and the uncertainties of the censored values are ignored. When the detection limit is increased, the proportion of censored values may increase, and more uncertainties in data may be ignored. Thus, the naive method may perform much worse than the proposed TS method.

5.6 Conclusions

From the simulation results, the proposed TS method clearly performs well and outperforms the naive method, as the proposed TS method has smaller bias and the coverage probabilities mostly close to the nominal level 0.95. These results confirm that the proposed TS method produces less biased estimates and more reliable standard errors, by adjusting the standard errors through bootstrap. The naive method has relatively large bias and low coverage probabilities because the naive method ignores the uncertainties of the censored values and the separate NLME model fitting.

The performances of the methods depend on the variations of data, sample size, and detection limit. Based on the simulation results in Table 5.2, more frequent repeated measurements do not seem to affect the simulation results, which may be because the number of repeated measurements in Setting I is large enough to produce good results. Choosing even more frequent repeated measurements does not have a significant impact on the simulation results. Smaller variations of data seem to somewhat lower the coverage probabilities of both the naive method and the proposed TS method. Increasing the detection limit leads to worse performance of the naive method, but not the proposed TS method. Overall, the proposed TS method has a better performance than the naive method, because it takes both the uncertainties of the censored values and the separate NLME model fitting into consideration. Note that the simulation is repeated 100 times. A larger number of repetitions may be considered for more reliable results in the future with more computing power.

Chapter 6

Conclusions and Future Research

6.1 Conclusions and Discussions

In this thesis, we have reviewed different models for fitting longitudinal data, including linear mixed effects (LME) models and nonlinear mixed effects (NLME) models. We have also reviewed several parameter estimation methods for NLME models and NLME models with censoring, including Monte Carlo EM (MCEM) algorithm, linearization method, and Stochastic Approximation EM (SAEM) algorithm.

In addition, we have considered joint NLME models for two longitudinal processes which are related to each other. The two longitudinal processes are linked through shared random effects, since these random effects reflect individual characteristics of the longitudinal processes. We have discussed joint inference methods, including the naive two-step method and the joint likelihood method. For parameter estimation in the joint models of two NLME models with censoring, the standard joint likelihood method may be computationally too intensive due to high-dimensional and unobservable random effects. Thus, we have proposed a *three-step (TS) method* based on the SAEM algorithm to reduce the computation burden, and also to produce more reliable results by incorporating the uncertainties in the censored values and separate model fitting using bootstrap. Another advantage of the TS method is easy to implement in standard software, such as *R*.

A real dataset from an HIV study has been analyzed by using a naive method and the proposed TS method. The naive method still uses the SAEM algorithm, but the censored values are substituted by half the detection limit and without bootstrap. We have found that the estimates from the two methods are close to each other, but the naive method usually has under-estimated standard errors (SE's). The naive method may lead to biased estimations, since the uncertainty of censored values and the uncertainty of estimation in the first step are not incorporated in the second

step. The goal of data analysis for the motivating HIV dataset is to study if key features of viral decay during ART may be associated with important features of viral rebound following ART interruption. The main findings are:

- the *initial and second-phase viral decay rates* during ART are negatively associated with the *viral setpoints* following ART interruption;
- the *initial viral load* values and *initial viral decay rate* during ART are negatively correlated with the *rate of rise* during viral rebound following ART interruption;
- the higher the *initial viral load* values during ART, the lower the *initial value* and more *delay in rise* following ART interruption; and
- the faster the *viral decay rate* after the start of ART, the lower the *initial viral load* following ART interruption.

These findings may provide insights into HIV cure research. Recent findings suggest that HIV-1 latent reservoir is primarily established near the time of ART initiation (Abrahams et al., 2019). Interventions in addition to ART to inhibit the formation of the latent reservoir may subsequently lead to a lower viral set point – a key goal of the HIV functional cure.

We have used simulation studies to compare the performances of the `nlme` package and `saemix` package for parameter estimation of NLME models using linearization and SAEM in standard software *R* based on the computational time, accuracy, and number of convergence problems. The results have shown that the `nlme` package produces more accurate estimates than the `saemix` package, but with longer computational time and more frequent convergence problems. The performances of the two packages may depend on the variations of data, sample size, and number of repeated measurements.

We have used another simulation study to compare the performances of the naive method and the proposed TS method based on the bias, mean square error (MSE), and coverage rate of 95% confidence intervals. We have found that the proposed TS method outperforms the naive method. Thus, by adjusting the SE's through bootstrapping, the proposed TS produces more reliable results than the naive method. From the simulation studies, we have also found that the variations of data, sample size, and detection limit may influence the performances of the two methods.

6.2 Future Research

In this thesis, we have proposed a computationally efficient TS method to study the associations among the individual viral dynamic characteristics during ART and following ART interruption, such as the individual viral decay rates and setpoints. The proposed TS method has also been proved to perform reasonably well by simulation studies. However, there are also some limitations in our research:

- The sample size $n = 75$ is relatively small.
- The parameter estimates may not be most efficient if the assumed models hold, since the TS method fit joint longitudinal processes separately rather than simultaneously.
- Many other possible covariates, such as the time from viral suppression to ART interruption and the possible association between the random effects in the CD4 model and the viral load model, are not considered in the model fitting for simplicity.
- We have assumed that the left-censored viral loads follow the same distribution as the observed viral loads, but this assumption is not testable based on the observed data.
- Due to large variations of viral loads following ART interruption, the viral rebound trajectories after reaching peak points may not be easily modelled parametrically.

In the following, we briefly describe some possible improvements in our research and some possible topics for future research. Firstly, we may consider the linearization method and the SAEM algorithm for the joint NLME models, and compare the performance of different joint methods through simulation studies. Basic ideas are as follows.

As discussed in Section 3.1, we may consider the linearization method based on Lindstrom and Bates (1990) for the two joint NLME models, with some modification. For simplicity, we first ignore the censoring. Let's rewrite NLME models (3.1)- (3.6) as a single equation

$$\begin{aligned} z_{ij} &= u_{ij}(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{b}_i) + u^*_{ij}(\mathbf{x}^*_i, \boldsymbol{\beta^*}, \boldsymbol{\tau}_i) + \xi_{ij}, \\ i &= 1, \cdots, n, \quad j = 1, \cdots, m_i \end{aligned}$$

where z_{ij} is the observed response for individual i at time t_{ij} for the entire study period, $u_{ij}(\cdot)$ and $u^*_{ij}(\cdot)$ are two nonlinear functions, $m_i = n_i + n^*_i$

is the number of repeated measurements both before and after ART interruption, and $\boldsymbol{\xi}_i = (\xi_{i1}, \cdots, \xi_{in_i})$ is a vector of random errors of the repeated measurements within-individual i. Note that u_{ij} are all zeros after the first n_i repeated measurements, and u^*_{ij} are all zeros in the first n_i repeated measurements, which allows different nonlinear functions for the viral decay period and viral rebound period. Let $\mathbf{u}_i = (u_{i1}, \cdots, u_{in_i})^T$ and $\mathbf{u}^*_i = (u^*_{i1}, \cdots, u^*_{in_i})^T$. At k-th iteration, denote the current estimates of $(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \mathbf{b}_i, \boldsymbol{\tau}_i)$ by $(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\beta}}^{*(k)}, \widehat{\mathbf{b}}_i^{(k)}, \widehat{\boldsymbol{\tau}}_i^{(k)})$, where $\widehat{\mathbf{b}}_i^{(k)}$ and $\widehat{\boldsymbol{\tau}}_i^{(k)}$ are the empirical Bayesian estimates of $\mathbf{b}_i$ and $\boldsymbol{\tau}_i$, respectively.

Then the linearization method is used to iteratively solve the following "working" LME model

$$\widetilde{\mathbf{z}}_i = W_i\boldsymbol{\beta} + W^*_i\boldsymbol{\beta}^* + T_i\mathbf{b}_i + V_i\boldsymbol{\tau}_i + \boldsymbol{\xi}_i, \tag{6.1}$$

where

$$\begin{aligned}
\widetilde{\mathbf{z}}_i &= \mathbf{z}_i - \mathbf{u}_i(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i) - \mathbf{u}^*_i(\mathbf{x}^*_i, \widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\tau}}) + W_i\widehat{\boldsymbol{\beta}} + W^*_i\widehat{\boldsymbol{\beta}}^* + T_i\widehat{\mathbf{b}}_i + V_i\widehat{\boldsymbol{\tau}}, \\
W_i &= \frac{\partial \mathbf{u}_i(\mathbf{x}_i, \boldsymbol{\beta}, \widehat{\mathbf{b}}_i)}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \quad T_i = \frac{\partial \mathbf{u}_i(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \mathbf{b}_i)}{\partial \mathbf{b}_i^T}\Big|_{\mathbf{b}=\widehat{\mathbf{b}}_i}, \\
W^*_i &= \frac{\partial \mathbf{u}^*_i(\mathbf{x}^*_i, \boldsymbol{\beta}^*, \widehat{\boldsymbol{\tau}})}{\partial \boldsymbol{\beta}^{*T}}\Big|_{\boldsymbol{\beta}^*=\widehat{\boldsymbol{\beta}}^*}, \quad V_i = \frac{\partial \mathbf{u}^*_i(\mathbf{x}^*_i, \widehat{\boldsymbol{\beta}}^*, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_i^T}\Big|_{\boldsymbol{\tau}=\widehat{\boldsymbol{\tau}}_i}.
\end{aligned}$$

At $(k+1)$-th iteration, the parameters and random effects from the LME model 6.1 are updated by $(\widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\boldsymbol{\beta}}^{*(k+1)}, \widehat{\mathbf{b}}_i^{(k+1)}, \widehat{\boldsymbol{\tau}}^{(k+1)})$ using standard methods described in Section 2.1. As discussed in Section 3.1, the linearization method is computationally more efficient than the MCEM algorithm, but the parameter estimates for joint models based on the linearization method may be less accurate than those based on the MCEM algorithm, and may offer potential convergence problems. We can use the `nlme` package in standard software R for the joint NLME models with linearization, and the performance of the linearization method for the joint NLME models can be studied using simulations in future research.

When some of the response values are censored, we can extend the MCEM algorithm based on Wu (2002). The basic idea is to use a Monte Carlo method in the E-step to approximate the conditional expectations given the observed data and current estimates, incorporating the censoring information. Then, the M-step is to apply a one-step linearization procedure to the nonlinear function and obtain the approximate MLEs by solving the score equations. The E-step and M-step are iterated until convergence. Wu (2002) uses the MCEM algorithm to find the MLEs of the parameters in the

covariate model and the response model simultaneously, and we can extend this method for the joint NLME models with censoring.

In addition, we may also consider the SAEM algorithm for the two joint NLME models with censoring. The SAEM algorithm is computationally more efficient than the MCEM algorithm, since the SAEM algorithm replaces the E-step of the MCEM algorithm by a single draw from the conditional distribution based on an MCMC method. To be more specific, under settings in Section 3.1, at iteration k, the E-step is to generate a sample from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\tau}_i|\mathbf{q}_i, \mathbf{q}_i^*, \mathbf{c}_i, \mathbf{c}_i^*, \mathbf{z}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$ by iteratively sampling from the conditional distributions $f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, $f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)})$, $f(\mathbf{w}_i|\boldsymbol{\tau}_i, \mathbf{b}_i, \mathbf{q}_i^*, \mathbf{c}_i^*, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$, and $f(\boldsymbol{\tau}_i|\mathbf{w}_i, \mathbf{q}_i^*, \mathbf{c}_i^*, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$ based on the Gibbs sampler. Note that

$$\begin{aligned}
f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) &\propto f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) \times f(\mathbf{c}_i|\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}),\\
f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{q}_i, \mathbf{c}_i, \mathbf{z}_i; \boldsymbol{\theta}^{(k)}) &\propto f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)}) \times f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i; \boldsymbol{\theta}^{(k)}),\\
f(\mathbf{w}_i|\boldsymbol{\tau}_i, \mathbf{b}_i, \mathbf{q}_i^*, \mathbf{c}_i^*, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}) &\propto f(\mathbf{w}_i|\boldsymbol{\tau}_i, \mathbf{b}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}) \times f(\mathbf{c}_i^*|\mathbf{w}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}),\\
f(\boldsymbol{\tau}_i|\mathbf{w}_i, \mathbf{q}_i^*, \mathbf{c}_i^*, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)}) &\propto f(\boldsymbol{\tau}_i|\boldsymbol{\theta}^{(k)}) \times f(\mathbf{w}_i|\mathbf{z}_i^*, \boldsymbol{\tau}_i; \boldsymbol{\theta}^{(k)}),
\end{aligned}$$

so we only need to generate samples from the right-hand sides of the above functions, which can be accomplished using rejection sampling methods since the density functions on the right-hand sides are known. The resulting sample roughly follows the distribution $f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\tau}_i|\mathbf{q}_i, \mathbf{q}_i^*, \mathbf{c}_i, \mathbf{c}_i^*, \mathbf{z}_i, \mathbf{z}_i^*; \boldsymbol{\theta}^{(k)})$. Then the conditional expectation is updated based on

$$\begin{aligned}
Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k([&\log f(\mathbf{y}_i|\mathbf{z}_i, \mathbf{b}_i, \boldsymbol{\theta}) + \log f(\mathbf{b}_i|\boldsymbol{\theta}) +\\
&\log f(\mathbf{w}_i|\mathbf{z}_i^*, \mathbf{b}_i, \boldsymbol{\tau}_i, \boldsymbol{\theta}) + \log f(\boldsymbol{\tau}_i|\boldsymbol{\theta})] - Q_{k-1}(\theta)),
\end{aligned}$$

where $\{\gamma_k\}_{k\geq 1}$ is a sequence of positive step size.

Although the SAEM algorithm for joint NLME models based on the joint likelihood is computationally more efficient than the MCEM algorithm, it can still be computationally intensive, since the dimension of the missing data $(\mathbf{y}_{cen,i}, \mathbf{w}_{cen,i}, \mathbf{b}_i, \tau_i)$ is very high. In addition, parameter estimates for the joint NLME models when using SAEM algorithm may be inaccurate. We can use software *Monolix* and *R* for the joint NLME models with SAEM. The performance of different methods for joint NLME models can be compared by simulation studies in future research.

There are some other topics that can be studied in future research:

- We may consider joint likelihood method via MCEM algorithm (Wu, 2009), but this method may be computationally expensive due to high-dimensional and unobserved random effects and censored viral loads.

Other methods, such as approximate joint likelihood inferences based on the so-called h-likelihood (Lee et al., 2017) or based on Laplace approximations (Vonesh et al., 2002), and Bayesian methods (Dey et al., 1997; Huang et al., 2018) may also be considered. However, the accuracy of the estimates based on the approximate joint likelihood methods could be a potential issue. The performances of different methods can be further compared in simulation studies.

- In future research, we may study the association between the individual viral dynamic characteristics and times to viral rebound or times to setpoints after the therapy is stopped. This association can be identified using joint inference of an NLME model for viral dynamics during ART and a time-to-event model such as a Cox proportional hazards model. As discussed in the literature review, there have been extensive studies on joint models for longitudinal and survival data (Yu et al., 2018; Hill et al., 2016; Conway et al., 2019, e.g.).

- We may consider different models in the research. Since we may not able to model the viral rebound trajectories after reaching peak points parametrically, we may consider semi-parametric NLME models for viral rebound. Since the distribution of censored viral loads are not testable, we may consider an approach that does not make such an assumption, e.g., treating the censored values as point masses as in Yu et al. (2018).

- In practice, there may be dropouts in the HIV studies due to some non-ignorable reasons, such as the health conditions of the patients. These dropouts may be associated with the longitudinal patterns of viral loads. Thus, we may consider incorporating missing data mechanisms into the joint models in future research.

In summary, there are still many issues about joint NLME models remain in this thesis. We plan to investigate these issues in our future research.

References

Abrahams, M. R., Joseph, S. B., Garrett, N., Tyers, L., Moeser, M., Archin, N., Council, O. D., Matten, D., Zhou, S., Doolabh, D., et al. (2019). The replication-competent HIV-1 latent reservoir is primarily established near the time of therapy initiation. *Science translational medicine*, 11(513):eeaw5589.

Aceto, L., Karrer, U., Grube, C., Oberholzer, R., Hasse, B., Presterl, E., Böni, J., Kuster, H., Trkola, A., Weber, R., et al. (2005). Primary HIV-1 infection in Zurich: 2002-2004. *Praxis*, 94(32):1199.

Beal, S. L. and Sheiner, L. B. (1982). Estimating population kinetics. *Critical reviews in biomedical engineering*, 8(3):195–222.

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, 80(1):1–41.

Conway, J. M., Perelson, A. S., and Li, J. Z. (2019). Predictions of time to HIV viral rebound following ART suspension that incorporate personal biomarkers. *PLoS computational biology*, 15(7):e1007229.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80(3):475–488.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. CRC press.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, 27(1):94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dey, D. K., Chen, M.-H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 53(4):1239–1252.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Fitzgerald, A. P., DeGruttola, V. G., and Vaida, F. (2002). Modelling HIV viral rebound using non-linear mixed effects models. *Statistics in Medicine*, 21(14):2093–2108.

Geman, S. and Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, 20(5-6):25–62.

Gianella, S., von Wyl, V., Fischer, M., Niederoest, B., Battegay, M., Bernasconi, E., Cavassini, M., Rauch, A., Hirschel, B., Vernazza, P., et al. (2011). Effect of early antiretroviral therapy during primary HIV-1 infection on cell-associated HIV-1 DNA and plasma HIV-1 RNA. *Antiviral therapy*, 16(4):535.

Girard, P. and Mentré, F. (2005). A comparison of estimation methods in nonlinear mixed effects models using a blind analysis. https://www.page-meeting.org/page/page2005/PAGE2005O08.pdf.

Hill, A. L., Rosenbloom, D. I., Goldstein, E., Hanhauser, E., Kuritzkes, D. R., Siliciano, R. F., and Henrich, T. J. (2016). Real-time predictions of reservoir size and rebound time during antiretroviral therapy interruption trials for HIV. *PLoS pathogens*, 12(4):e1005535.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–126.

Huang, Y., Lu, X., Chen, J., Liang, J., and Zangmeister, M. (2018). Joint model-based clustering of nonlinear longitudinal trajectories and associated time-to-event data analysis, linked by latent class membership: with application to AIDS clinical studies. *Lifetime data analysis*, 24(4):699–718.

Hughes, J. P. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, 55(2):625–629.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.

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

Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press.

Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the DIA Bayesian joint modeling working group. *Statistics in medicine*, 34(14):2181–2195.

Lee, Y., Nelder, J., and Pawitan, Y. (2017). *Generalized linear models with random effects: unified analysis via h-likelihood*. CRC Press, second edition.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

Paxton, W. B., Coombs, R. W., McElrath, M. J., Keefer, M. C., Hughes, J., Sinangil, F., Chernoff, D., Demeter, L., Williams, B., Corey, L., et al. (1997). Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with $\geq$ 400 CD4 lymphocytes: implications for applying measurements to individual patients. *Journal of Infectious Diseases*, 175(2):247–254.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2019). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-143.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitu-

dinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.

Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics & Data Analysis*, 51(3):1562–1574.

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, 97(457):271–283.

Wang, R., Bing, A., Wang, C., Hu, Y., Bosch, R. J., and DeGruttola, V. (2020). A flexible nonlinear mixed effects model for HIV viral load rebound after treatment interruption. *Statistics in Medicine*, 39(15):2051–2066.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., et al. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510):117–122.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4):791–795.

Wolfinger, R. D. and Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics & Data Analysis*, 25(4):465–490.

Wu, H. and Ding, A. A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics*, 55(2):410–418.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of the American Statistical association*, 97(460):955–964.

Wu, L. (2009). *Mixed effects models for complex data*. CRC Press.

Wu, L., Hu, X. J., and Wu, H. (2008). Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data. *Biostatistics*, 9(2):308–320.

Yu, T., Wu, L., and Gilbert, P. B. (2018). A joint model for mixed and truncated longitudinal data and survival data, with application to HIV vaccine studies. *Biostatistics*, 19(3):374–390.

Appendix A

Software Codes

Github link for R code is provided below:

`https://github.com/Sihaoyu1220/Thesis_Code`