Nonlinear mixed-effects models for multivariate longitudinal data with application to HIV disease dynamics

Artz George Luwanda

Submitted in fulfillment of the academic requirements for the degree of

Doctor of Philosophy

in

Statistics

in the School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

2014

Declaration

The research presented in this thesis was conducted in the School of Mathematics, Statistics and Computer Science at the University of KwaZulu-Natal, Pietermaritzburg, from March, 2011 to April, 2014 under the supervision of Prof. Henry G. Mwambi.

I declare that this work is a result of my own effort and that wherever other people's work has been used, reference has been duly made. I further declare that this work has not been presented for any academic award at this university or any other university.

Artz George Luwanda: ..... Date:.....

Prof Henry G. Mwambi:..... Date:.....

Acknowledgements

I would like to thank Professor Henry G Mwambi for the focused guidance and timely response whenever consulted for advice. I have also learned a lot of research techniques through our discussions.

I would like to thank NCST (Malawi) for sponsoring my studies. Sincere appreciations also go to the College of Agriculture, Engineering and Science of the University of KwaZulu-Natal for the financial assistance that facilitated the attendance of conferences and academic workshops and for my general welfare. I would also like to extend my gratitude to Mzuzu University for granting me study leave.

I also thank the Lighthouse Trust (Malawi) for providing me with the data used in illustrating methodologies in this thesis.

Lastly but not least, I am grateful for the support I got from staff and fellow graduate students in the School of Mathematics, Statistics and Computer Science at the Pietermaritzburg Campus.

Dedication

This work is dedicated to my wife Mercent, daughters Loveness and Lincey and my son Wanangwa.

Abstract

The motivation for the study of nonlinear mixed-effects models is due to the growing interest in the estimation of parameters in HIV disease dynamical models using real multivariate longitudinal data with varying degrees of informativeness. Special analytical and approximation techniques are needed to deal with such data because the repeated observations on any experimental unit are likely to be correlated over time while multiple outcomes within the unit will also be correlated. Furthermore, observations may be irregularly made within and between individuals making direct use of standard methods practically impossible.

In this thesis, we consider a nonlinear mixed-effects model for a multivariate response variable that takes into account left-censored observations. Then we study a case where data are unbalanced among subjects and also within a subject because for some reason only a subset of the multiple outcomes of the response variable are observed at any one occasion. Dropout models that take into consideration the partially observed outcomes are proposed. We further derive a joint likelihood function which takes into account the multivariate responses and the unbalancedness in such data as a result of censoring and dropout. We then show how the methodology can be used in the estimation of the parameters that characterise HIV dynamical system in the presence of several covariates. We have also used multiple imputation to compare covariate coefficients in the complete data and

the partially observed data. Through a simulation study, we have also seen that a small limit of quantification provides better parameter estimates in the sense of standard errors and confidence limits of the parameters. Since there are usually no analytic solutions for such complex models, the stochastic approximation Expectation-Maximisation (SAEM) is used as an approximation method. The methodology is illustrated using a routine observational dataset from two HIV clinics in Malawi.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

DO = drop-out.

EM Algorithm = expectation-maximization algorithm.

HIV = Human immunodeficiency virus.

LOQ = Limit of quantification.

MI = Multiple imputation.

PDO = partial drop-out.

SAEM Algorithm = stochastic approximation expectation-maximization algorithm.

Chapter 1

Introduction

1.1 General introduction

Longitudinal data consist of measurements or observations taken repeatedly over a period of time on a particular subject. This usually happens in studies where the objective is to observe change in response of subjects under study and the rate of change of the response for an individual subject relative to others on the same treatment. Longitudinal studies can distinguish changes over time within subjects (*time effects*) from differences among subjects in levels taken at baseline (*population or cohort effects*) (see Diggle et al., 2002). Thus they allow direct study of change in response with time and factors that influence that change (Fitzmaurice et al., 2009, 2011). Subjects in the study act as their own control as response values are compared over the period spanning the study. It could be of importance to determine if these within-subject changes are influenced by a combination of selected predictor variables. This is advantageous because, generally, within-subject variability is less than the between-subject variation. This then results in more reliable parameter estimates.

These are some of the direct advantages of longitudinal studies over cross-sectional-studies where measurements are taken only once. Cross-sectional data do not offer the chance to analyse within individual change due to treatment and the rate of growth of a particular subject relative to others under the same study conditions. Another advantage of longitudinal or repeated measurement studies is that even with a small sample size of individuals additional statistical power can be gained because of pooling information from repeated measurements compared to a cross-sectional study of equal sample size.

This, however, does not mean that longitudinal studies do not have disadvantages. Data in longitudinal studies form natural clusters. The clusters have measures observed or obtained from individual subjects at different occasions during the study for their elements. These elements are usually positively correlated and this correlation has to be accounted for in the analysis. This is even more challenging if the data are to be fit to complex models because in some cases available software may be rendered less helpful in the face of the combination of such models and longitudinal data (Hedeker and Gibbons, 2006).

Since longitudinal data are collected by repeated measurement or observation of the same subject over a period of time, there is a likelihood of subjects missing one or more scheduled observation occasions. The missingness could be *intermittent* or complete withdrawal (or *dropout*) from the study. This results in efficiency of estimates of the population parameters being compromised (Laird, 1988). Dropout can also lead to selection bias when the distribution of covariates and the response variable depend on the subject's continued presence in the study (see Newsom et al., 2012, page 7). We discuss this point in more detail in Chapter 4.

Subjects may be required to report repeatedly to the study centre at appointed dates or times. As a result a longitudinal study is usually more expensive and

time consuming than the cross-sectional study (Twist, 2003). In general, however, the advantages of a longitudinal study outweigh the disadvantages. This makes longitudinal studies popular especially in epidemiological, clinical, psychometric and other social investigations. In this thesis we use such type of data to estimate parameters that characterise the HIV dynamical system.

1.2 Motivation of the study

Within host HIV dynamic models are important because they have led to a better understanding of the disease process and progression (Duffin and Tullis, 2002; Wu, 2005). These dynamic models are based on systems of nonlinear ordinary differential equations where the compartments are defined using HIV disease biomarkers, usually viral load measurements and CD4+ T cell counts (Perelson and Nelson, 1999; Nowak and May, 2000). Several process parameters, including treatment efficacy and clearance rate of virus particles, are estimated using the biomarkers by applying statistical models. Thus more reliable statistical methodologies are required to provide estimates or approximations of the parameters of the dynamic system which in turn would translate into an improved interpretation of the disease progress and treatment intervention.

In the existing literature, system compartments are based on only two observed markers. There is a need to adequately characterise these compartments using the observed and those the missing markers in order to realise more meaningful and less biased parameter estimates.

Moreover, an estimation technique needs to take into account the compartment nature. This is an area that has received much attention recently (Perelson, 2002; Huang et al., 2006; Huang and Lu, 2008; Guedj et al., 2011). However, the inclusion

of a multiplicity of explanatory variables that may latently influence the system parameters and partial dropout of the subjects have not been fully modelled in processes related to estimation of parameters of these biological systems. These are some of the main objectives of this thesis. We, however, acknowledge the deficiencies we may have with our data to be able to fully address the estimation problem effectively because they are observational in nature. A full description of the data for the study is given in Chapter 2.

1.3 Multivariate longitudinal data

The important characteristic of longitudinal data analysis is that a response variable consists of measurements or observations taken repeatedly from an individual over a period of time that spans the study duration. The response variable may be in the form of a single characteristic observed a number of times or multiple characteristics observed at each time point on a particular subject (Ferrer et al., 2005; Weiss, 2005; Verbeke and Molenberghs, 2009; Fitzmaurice et al., 2011). In this thesis, we use the multivariate longitudinal data and we describe them a little more in this section.

Analysis of multiple outcomes from the same experimental unit at a given time has become an increasingly interesting and important concept especially in clinical, psychometric and other social science studies. For instance, to measure the severity of schizophrenic symptoms one comes up with a response variable that has several symptom outcomes (Diggle et al., 2002, page 332). This results in a response vector of high dimension and this clearly introduces the problem of having to jointly model the different outcomes in order to make unbiased inference about the disease.

One of the pioneering works that has motivated studies in multivariate longitudinal modelling is found in a discussion by Potthoff and Roy (1964). This discussion provided a generalised platform for handling estimation and inference procedures with respect to the multivariate longitudinal data setting. Reinsel (1982) considered a multivariate random effects covariance structure in the analysis of multivariate longitudinal data. He derived closed-form expressions for parameter estimates and for testing hypotheses related to model parameters. In his follow-up discussion Reinsel (1984) derived formulas for the prediction of response values and their mean square errors. In both cases, response outcomes are balanced and completely observed.

Lundbye-Christensen (1991) proposed a model for multivariate longitudinal responses under the assumption that these outcomes are proportional to an unknown underlying growth process. The article also discussed the aspects regarding cross-sectional distribution of the growth process. The discussion assumed a complete multivariate outcome setting.

In their discussion Mickey et al. (1994) outlined a procedure for the summarisation of multiple outcomes in terms of linear combination of such outcomes. They observed that maximisation of the variance of the linear combination and that of the fit of the model can be used as tools for determining model coefficients. The study by Nummi and Möttönen (2000) discussed a multivariate regression model for handling growth curve data. They derived closed-form expressions for model parameters under the maximum likelihood and restricted maximum likelihood setting. Furthermore, under a complete data assumption they proposed a procedure for testing linear hypotheses.

Apiolaza and Garrick (2001) used longitudinal data analysis in a different context to explain the relationships among several models represented by an additive generic covariance matrix. According to Oort (2001) multivariate longitudinal data

can be characterised by three elements: (i) response variable, (ii) subject under study and (iii) occasion of observation. He he provided a criterion for measurement invariance and used structural equation modelling techniques to estimate parameters in autoregressive and latent curve models.

In modelling Alzheimer's disease, Beckett et al. (2004) used a random-effects approach for the association between two characteristics of the cognitive domain. Furthermore, they compared the joint results to those obtained by treating them as separate growth curves analysed using ordinary least squares approaches. In their submission, Jia and Weiss (2009) observed that a set of multivariate outcomes may share common explanatory effects such that the response to changes in these effects is similar among such common outcomes. They proceeded to develop an algorithm for fitting the model using standard software for univariate longitudinal data.

In general, both linear and nonlinear models for complete multivariate longitudinal data have been widely discussed in the literature (Gueorguieva, 2001; Ferrer and MacArdle, 2003; Harvey et al., 2003; Zhang, 2004; Dubin and Müller, 2005; Fieuws and Verbeke, 2006; Blozis et al., 2007). In these discussions, estimation and approximation techniques have been illustrated either by way of examples or through simulation studies. Requirements for model selection have also been discussed and applied. There has been a general assumption that the data are completely observed in these discussions. But in most longitudinal studies missing a scheduled visit or indeed premature withdrawal from the study is not uncommon.

1.4 Missing data in longitudinal studies

The design of longitudinal studies is such that complete information is to be collected from a fixed number of subjects for a specified number of occasions. Since such studies require that measurements be taken repeatedly on a particular subject, it is not uncommon for some subjects to miss some occasions or completely drop out of the trials before the end of the prescribed study periods. This results in incomplete data and this consequently renders standard analysis techniques inappropriate because the underlying causes of incompleteness need to be addressed (Gad and Ahmed, 2006; Diggle and Kenward, 1994). The analysis procedures need to account for these missing data so that estimates (or their approximations) are meaningful (Rubin, 2004; Molenberghs and Kenwrad, 2007). There has been considerable discussion on univariate longitudinal data with missingness in different forms (Cnaan et al., 1997; Everitt, 1998; Qu and Song, 2002; Crouchley and Ganjali, 2002; Hu and Sale, 2003; Wu and Wu, 2007).

In their analysis of multivariate longitudinal data with non-ignorable dropout, Roy and Lin (2002) assumed that observed outcomes measure a latent variable with error and modelled the relationship between this latent variable and covariates using a linear mixed model. Pan and Louis (2000) also applied a linear mixed-effects model to multivariate longitudinal data with censored observations. A regression calibration approach for finding a joint model for multiple longitudinal measurements and discrete time-to-event data was proposed by Albert and Shih (2010). They discussed an approach that gives accurate estimates of model parameters in the presence of informative dropout.

A marginalized analysis technique for multivariate longitudinal binary data was developed by Lee et al. (2009) and this procedure takes into account the dropout

process. Sy et al. (1997) presented a bivariate longitudinal data model that incorporates random effects, correlated random processes and measurement errors. The model they developed accommodates unbalanced observations and missing outcomes. More accounts on incomplete multivariate longitudinal data can be found in Jorgensen et al. (1996); Shah et al. (1997); Schafer (1997); Thièbaut et al. (2003); Pantazis et al. (2005) and Deslandes and Chevret (2010).

The missingness mechanisms for multivariate longitudinal data discussed in the literature have assumed that if one characteristic of a response variable is missing, then the others are also missing. In other words, the response variable is missing as an entity for a particular occasion or a set of occasions. However, there are a lot of multivariate outcomes problems in practice where the response variable is partially observed for the occasions the measurements are taken. For instance, in HIV disease modelling, the dynamic system depends on outcomes some of which are not observed. Moreover, there are a lot of values of some markers which are below detectable limits. However, these unobserved outcomes are usually related to observed markers.

1.5 Research objectives

This thesis has been motivated by the need to estimate parameters of the HIV dynamical systems which are based on nonlinear ordinary differential equations of the components of CD4+ T cell counts and viral loads which are important HIV disease markers. This is achieved by using different scenarios of multivariate longitudinal data when there are several covariates and unobserved markers. Specifically, the thesis has the following objectives:

- to review nonlinear mixed-effects models for complete multivariate longitudinal data as applied to estimation of HIV dynamical system while including several covariates;

- to propose subject dropout and partial dropout models for multivariate longitudinal data in which data are unbalanced among subjects and also within a subject because for some reason only a subset of the multiple outcomes of the response variable are observed at any one occasion;

- to derive joint likelihood functions that take into account subject dropout, partial dropout and left-censored data;

- to illustrate the methodology using the stochastic approximation Expectation -Maximisation (SAEM) algorithm as an approximation tool. We use a routine observational dataset from a constrained resource setting in the analyses;

- to compare estimates of parameters found using complete data through multiple imputation and those found using partially missing data;

- to conduct a sensitivity analysis of estimates of the HIV dynamical system for different limits of quantification of the viral load measurements through a simulation study based on data with partial dropout.

1.6 Outline of the thesis

The rest of the thesis is organized as follows. Chapter 2 gives a description of the data to be used in illustrations of methodologies in this thesis. In that chapter, we state that the data are observational in nature where information and laboratory

measurements were taken from HIV patients reporting to clinics under the Lighthouse Trust in Malawi. These data are characterised by dropout and a lot of viral load measurements that are below the limit of quantification which is 400 copies per ml (for the current data). Some results from exploratory analysis of these data are presented in this chapter.

In Chapter 3 we consider a nonlinear mixed-effects model for complete multivariate longitudinal data. Tests have been conducted in this chapter regarding the most appropriate set of covariates that can be included in the analysis involving such type of data. We also introduce the stochastic approximation Expectation Maximisation (SAEM) algorithm which is used throughout the thesis as an approximation procedure because there are no closed-form expressions for the parameter estimates of such models.

Chapter 4 models the dropout mechanism. Specifically we consider a case where dropout requires that all the markers of the response variable are not available after a subject drops out. We have used the approach of Diggle and Kenward (1994) to derive a dropout model for a multivariate response variable and the likelihood function that is used to estimate parameters under this setting. We have also considered a case where data are unbalanced among subjects and also within a subject because for some reason only a subset of the multiple outcomes of the response variable are observed at any one occasion. This is the topic that has been covered in Chapter 5. We have modelled the partial dropout mechanism of the markers of the response variable using the occasions of dropout. A joint likelihood function that incorporates the left-censored response values due to equipment failure to measure some outcomes accurately and this partial dropout mechanism has also been proposed in this chapter. Furthermore, we have shown using the proposed joint likelihood function and the algorithm how one can find approximate

estimates of the standard errors of the parameter estimates.

The sensitivity analysis of the estimates of the HIV dynamical system parameters has been carried out in Chapter 6. Specifically we have looked at the role played by the limits of quantification (LOQ) of the viral loads in determining the size of bias of estimates of the system parameters. It has been observed that, in general, the width of the confidence interval of the parameters is small when the LOQ is also small. In this chapter, we also conducted multiple imputation with the aim of comparing significance of covariate coefficients in the data with partial dropout and those with complete observations for the same number of patients and also the parameter estimates with their standard errors for the two scenarios.

A summary of the findings of the thesis are presented in Chapter 7. Limitations of the current research are discussed in this chapter. We have made some suggestions for future work. It could be possible to use functional data analysis in order to pool the randomness in the observations from the various data sources such as clinics in our case (Martínez-Camblor and Corral, 2011). It has been pointed out that in practice there are cases of viral rebound which are usually associated with resistance to treatment regimen and it would be in order to include parameters that characterise this phenomenon and elements of disease resistance in the dynamical model considered in this thesis.

Chapter 2

Data description and exploratory analysis

In this chapter we describe the data that will be used in illustrating the methodologies which we review and develop in the subsequent chapters.

2.1 Data description

Depletion of the CD4+ T cells is the manifestation and the source of the central immune system defect of HIV disease and the viral load has become an important marker and is also now used in some cases as the primary marker for antiretroviral treatment policy. In Malawi for instance, an adult patient is eligible to start treatment if their CD4+ T count is below 250 cells per mm^3 (Ministry of Health, 2008). These markers are also used in evaluating subjects' response to antiretroviral therapy (Huang and Lu, 2008). Thus viral load measurements and CD4+ T cell counts are the two major markers in the study of the HIV disease.

HIV patients at the Lighthouse clinics (in Malawi) are routinely monitored to check

how they are responding to treatment and those that need to start treatment have their CD4+ T cell counts and viral loads checked to determine if the disease is advanced enough to warrant entry into a treatment course. This means that the collected data are longitudinal in nature as measurements for each patient are taken repeatedly over a period of time. For this study there were two facilities from which the data were collected: Lighthouse clinic (LHC) and Bwaila clinic (MPC). Both of these clinics are under the Lighthouse Trust in Malawi which is within Kamuzu Central Hospital. At the time of acquisition of these data, there were many subjects whose information was mostly available. However, patients included in the analysis were adults (18 years old or more) whose data were captured between 2008 and 2010 and were on a three-course treatment of Stavudine-Lamivudine-Nevirapine or Zidovudine-Lamivudine-Nevirapine.

For most patients the clinical information including CD4+ T cell counts and viral load measurements were collected at the time of treatment initiation. In the course of treatment, however, not all patients have frequent measurements of viral loads and this leaves CD4+ T cell count measurements to be obtained more frequently. Viral load levels are usually not quantifiable below a certain value which is known as the limit of quantification (LOQ). This value depends on the assay used in the measurement process. At the time of acquisition of the current data the lowest quantifiable value for viral loads was 400 copies per ml so that all readings less than this value were quantified as 400 copies per ml. Furthermore, only subjects with a minimum of three (including baseline) measurements of both the CD4+ T cell count and viral load were included in the analyses in the chapters that follow. This would ensure that we had adequate bivariate longitudinal data. Three data scenarios were considered in this thesis depending on whether we have complete and balanced observations or unbalanced among subjects or indeed partially observed

data within and among the subjects.

The first case involved complete and balanced data with each patient having five pairs of the markers observed on five different occasions. There were seventy-eight subjects with such balanced data and these gave a total of 780 laboratory measurements. This type of dataset is used in Chapter 3. About 40% of the subjects had viral load values that were less that the threshold value and this is illustrated by the cumulative distribution function (F_n) of the log10-viral loads in Figure 2.1. The value 2.6021 on the x-axis of this graph corresponds to 400 copies per ml which is the lower limit of quantification of the current data.

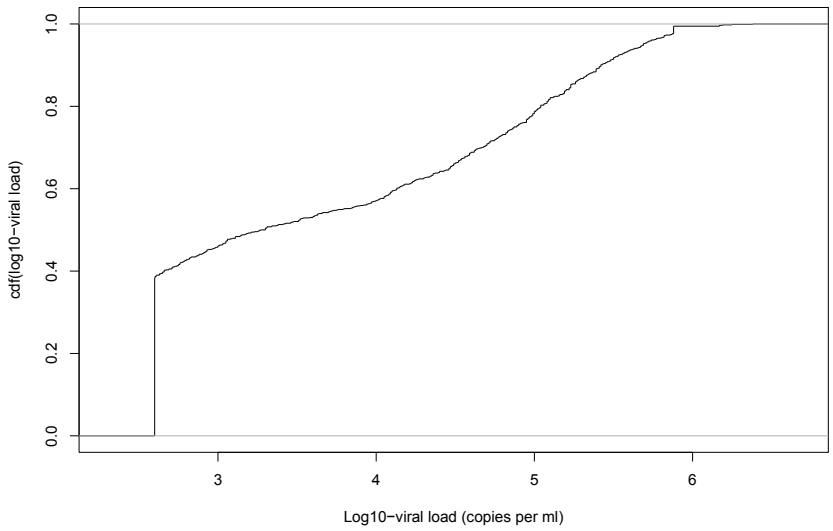


Figure 2.1: The distribution function of the log10-viral load.

The second scenario consisted of unbalanced data due to different number of occasions in which markers were observed for each patient. In this setting all subjects with $3-7$ bivariate measurements (including the baseline values) were included in the analyses. The truncation was done to ensure that we had adequate marker

readings from each patient for the analysis. This truncation is a common practice in problems involving dropout mechanisms (Encrenaz et al., 2005). There were two hundred and fourteen (214) patients with this information representing about 8% of the target population which was all the HIV patients at the Lighthouse Clinics who were on treatment at the time of collection of these data. The total number of clinical measurements in this setting was 1856 for both CD4+ T cell counts and viral load. Thirty eight (37.8%) percent of the viral load measurements were below the threshold value. The percentages of patients with both CD4+ T cell count and viral load measurements on each occasion are displayed in Table 2.1. From this table, we note that only a small proportion of patients had more than five viral load measurements.

Table 2.1: Percentage of patients with both markers by week

Week	0	14	28	42	56	70	84
Percentage observed	100	100	100	82.3	36.4	11.2	3.3

Most patients had CD4+ T cell count observed at all the visits but no viral load readings recorded. That is, for a particular subject it was possible for the number of the viral load measurements and that of the CD4+ T cell counts to be different. This gave us the third scenario which we term the *partially observed* data because the bivariate response variable was partially observed. As noted in the second case above the subjects were not necessarily observed the same number of occasions. For purposes of the analyses in this thesis, we required that at least three pairs of the marker measurements be complete and that eight be the maximum number of occasions. There were only seven patients that had CD4+ T cell counts recorded up to the eighth occasion. A total of two hundred and fourteen (214) subjects were

observed and 2024 marker observations were made, of which 1096 were CD4+ T cell count measurements (representing 54% of the total number of marker observations) and the rest were viral load measurements. Note also that the datasets in the first two scenarios are subsets of this dataset. Thus all exploratory analytical properties of the partially observed dataset are shared by the other two cases and, therefore, data descriptions in the rest of the chapter will be based on data from the third scenario.

2.1.1 Occasions for observation

The observation times (or occasions) were purely for purposes of assessing the progress of treatment for individual patients and as such, their spacing was widely varied. Furthermore, the CD4+ T cell counts and viral load measurements were not observed at the same occasions. This means that the data were unbalanced in terms of number of occasions at which the patients were observed and the intervals between any two occasions for a particular patient. In cases like this, it helps in the analyses to consider using the overall rate of marker observation for all the patients under consideration (Romih et al., 2010). In the current analyses we took the average observation interval between the markers in order to achieve this and it was set at 14 weeks.

For the discussion of mixed-effects models for balanced multivariate longitudinal data in Chapter 3, we assumed that the seventy-eight subjects were each observed on five equally spaced occasions. In Chapters 4 and 5 where we discuss dropout mechanisms, the minimum number of occasions before a withdrawal or drop-out was taken to be three. Apart from proposing models for partially observed data, the other objective was to highlight the contribution of covariates in the estimation

of parameters of the virus dynamical system (Steiner et al., 2010; Rice et al., 1999).

2.1.2 Covariates

There were several covariates collected along with the bivariate response variable. The aim was to study and describe the influence of these covariates on parameters of the statistical model as well as those of the HIV dynamical system which we discuss later. The covariates include: (i) age of the patient at the time of initiation of treatment (coded -1 = age between $18-25$, 0 = between $26-33$, 1 = between $34-40$ and 2 = over 40); (ii) Sex of subject (coded 0 = Female, 1 = Male); (iii) Facility (clinic) from which subject i was getting medication and general health care as part of treatment (0 = LHC, 1 = MPC); (iv) Supplementary treatment (0 = No, 1 = Yes) and (v) compliance to treatment (0 = No, 1 = Yes).

The coding for age-groups was necessitated by the scale of measurement which was higher for this group as compared to the other covariates and this rescaling helps in stabilising model parameters estimates which results in improved interpretation of the parameters (Roy and Lin, 2002). There are two facilities under the Lighthouse Trust from which the data were collected as explained in Section 2.1: LHC and MPC. The last two covariates depended on the patient's adherence to consultation (visiting) schedule and are described as follows. If the patient visited the clinic at least a week before appointment date, we assumed there would have been a problem that needed clinical attention (supplementary treatment). A patient coming for a visit to the clinic a week (or later) after appointment date had missed the recommended treatment plan and as such it indicated lack of compliance to the treatment schedule on the part of the patient. This description was arrived at after consulting the experts who were mandated to capture the patients' information for

this study.

In this thesis, we assume that these covariates are time invariant. In general, however, this may not be the case. For instance, compliance to treatment is determined during follow-up visits.

2.2 Exploratory data analysis

2.2.1 Descriptive statistics of the dataset

Age at treatment commencement was used as a grouping variable and as such several data features were considered. The frequency distribution of the patients among the four age-groups is displayed in Table 2.2. Using the Chi-square test for goodness-of-fit it can be argued that the number of patients in each age-group is independent of the age-group ($p = 0.0617$). Figure 2.2 displays the box plot of

Table 2.2: Distribution of patients for the age-groups.

Age-group	$18-25$	$26-33$	$34-40$	over 40
No. of patients	41	68	49	56
Percentage	19.1	31.8	22.9	26.2

the markers for the ages. We note that for the viral load the 18-25 age-group has a bigger median than the other three groups (top panel) whereas for the CD4+ T counts the older patients (those over forty) show a slightly higher median cell count compared to the other age-groups.

At this point we cannot give reasons for this but further analyses in later chapters will possibly shed light into the phenomenon. However, one could say that since

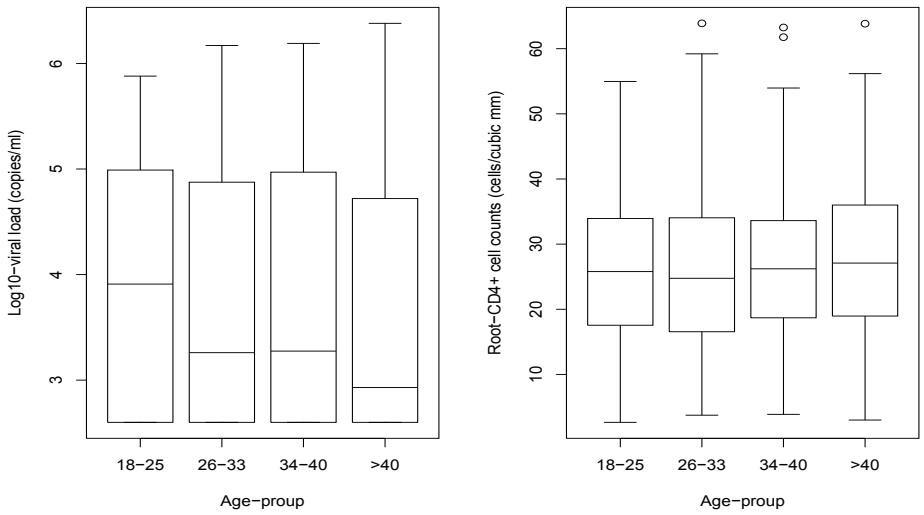


Figure 2.2: Box plot of the markers for the four age-groups.

high viral load levels imply high rate of virus production we could suggest that in the age-group $18 - 25$ viral production is higher than in the other three age-groups. A small percentage of outliers is seen in the second and third age-groups. This proportion (of outliers) is not large enough to suggest the existence of sub-populations within these two age-groups.

Women seem to have lower viral loads than men in this group of subjects as displayed in Figure 2.3 and a Chi-square test of independence confirms this assertion with a p-values of 0.04381. These results were also noted in a discussion by Govender et al. (2014). It may not imply, however, that this difference has an effect on the rate at which HIV disease progresses in women and men. A suggestion might be advanced that women have lower viral loads due to the way their immune system responds to viral infection or that in general the viral production in women is lower than in male patients. Meier et al. (2009) claim that when women and men with the same viral loads are compared, HIV disease progression is generally

faster in women than men. That is, as seen in this figure one could conclude that female HIV patients may develop AIDS at lower viral load levels than their male counterparts with similar CD4+ T cell counts. As with all other covariates, there were only a few outliers in the CD4+ T cell count plots for both male and female.

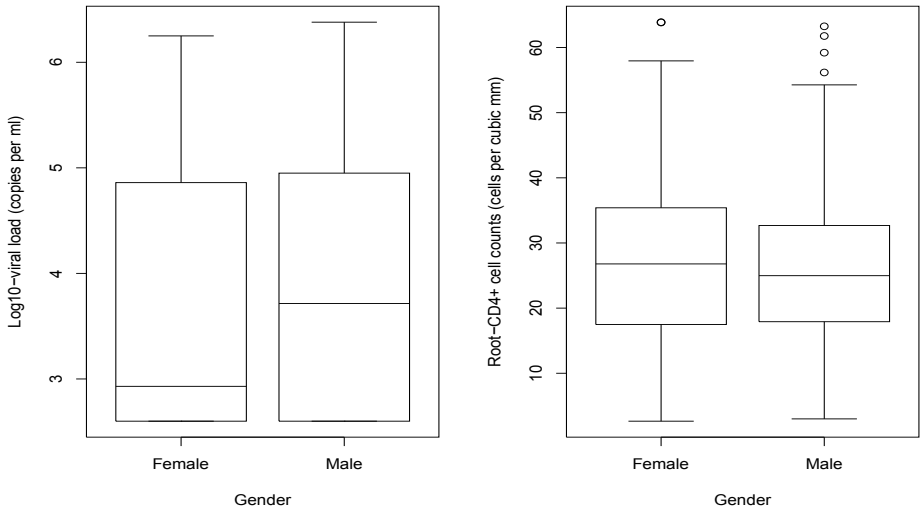


Figure 2.3: Box plot comparing markers for the two sexes.

Tests for goodness-of-fit on the current data show that there were more patients with the required information reporting to Bwaila clinic than Lighthouse clinic. The data, however, do not seem to provide evidence that the choice of facility depended on gender (with a $p-$value of 0.9747). Similarly, the data reveal that at both facilities there were equal proportions of patients that sought supplementary treatment. There is evidence ($p-$value of 0.0016), however, that the choice of facility depended on the patient's age at treatment commencement. For instance, more of those in the $18-25$ age-group preferred to use Lighthouse clinic (see Table 2.3) which represents a proportion of about 2.6 times that of the other clinic.

Table 2.3: Distribution of patients by age-groups reporting to a facility.

		Age-group				
		$18-25$	$26-33$	$34-40$	>40	Total
Facility	Bwaila	17	42	38	41	138 (64.5%)
	Lighthouse	24	26	11	15	76 (35.5%)

The preliminary results also show evidence ($p < 0.001$) that compliance to treatment schedule was associated with the facility to which the patient reported. For example, only 5% of the patients reporting to Lighthouse clinic were non-compliant compared to about 70% at Bwaila clinic. There is no immediate factor that can be attributed to this difference because we did not look at the different characteristics of these clinics since our ultimate objective was the estimation of parameters of the HIV dynamical system. However, this may point to limited financial resources on the part of the patients in order to regularly report to a treatment facility. That is, one facility could be cheaper to access in terms of travel than the other. Another factor for the dependence of compliance on facility could be the overwhelming number of patients getting service from it so that non-compliance could be partly due to administrative reasons by possibly postponing consultation dates (Hogan et al., 2004). For some patients, it could only be the apparent good health they enjoyed as they took ARVs and in others the absence of general impact on anticipation of the future as ARVs may have substantial side effects on them. These, though, cannot be attributed to any one facility.

A Chi-square test on Table 2.4 shows some degree of dependence between supplementary treatment and compliance in these data with a $p-$value of 0.017. This is understood because a subject not complying with the treatment is likely to suf-

Table 2.4: Distribution of patients to compliance-supplementary treatment.

		Compliance		
		No	Yes	Total
Supp. treatment	No	38	63	101
	Yes	62	51	113

fer from opportunistic infections which will require medical advice as a result of such illnesses. This exploratory assessment has also revealed that age-group and compliance to treatment are dependent (p-value of 0.0339). These analyses have also shown that compliance to treatment is not influenced by the patient's gender. Moreover, the proportions of those complying to treatment are equal in two sexes (at about 53% on average in each case).

2.2.2 Bivariate distribution of markers

The response of interest is a bivariate variable with CD4+ T cell count and viral load measurements as characteristic variables. These markers are known to be negatively correlated or are said to have an inverse relationship (Wu and Müller, 2011). More in formation about the relationship between the two markers can also be found in a discussion by Feinberg (1996) where the dynamics of these markers in the course of the HIV disease are highlighted. The overall correlation coefficient between the log10-viral load measurements and the root-CD4+ T cell counts for the current dataset is -0.482 which is not different from values found in the literature. As an example, the value of the correlation coefficient between the two markers for the data used by Mata-Marín et al. (2009) was -0.439.

When we group the data into covariate classes, the values of correlation coefficients obtained show no remarkable difference from one cluster of the category to the other. For instance when we look at the correlation coefficients for the two sex, the values are not different to three places. Both had the same value of -0.48 which is equal to the overall computed correlation coefficient. A substantial difference, however, was noted in the correlations coefficients among the age-groups especially for the groups $34-40$ and >40 with $r=-0.3737$ and $r=-0.5265$ respectively. This is possibly in line with what has been observed in Figure 2.2 where for the same median CD4+ T cell counts for the two groups we note a considerable difference in the median viral load values.

We also checked for normality of the transformed variables: log10-viral load and square-root of the CD4+ T cell count. The results are displayed in Figure 2.4 from which we note that the root-CD4+ T cell counts represent a Gaussian variable (right panel). The log10-transform of the viral load measurements shows a clear departure from the normality assumption and this could, in part, be explained by the large number of viral load values that were below limit of quantification (400 copies per ml). This suggests that a different transformation for the viral load levels would be more appropriate. However, in the analyses we will use the log10-transform for the viral load measurements as is the practice in the literature (Guedj et al., 2007a; Yu and Liang, 2013).

A bivariate linear mixed model for the markers of the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}, \tag{2.1}$$

was considered where $\mathbf{Y}_i$ is the bivariate response variable, $\boldsymbol{\beta}$ is the matrix of the effects and X_i the matrix of the predictor variables. The results of the analysis are displayed in Table 2.5. It can be seen from these results that the intercept,

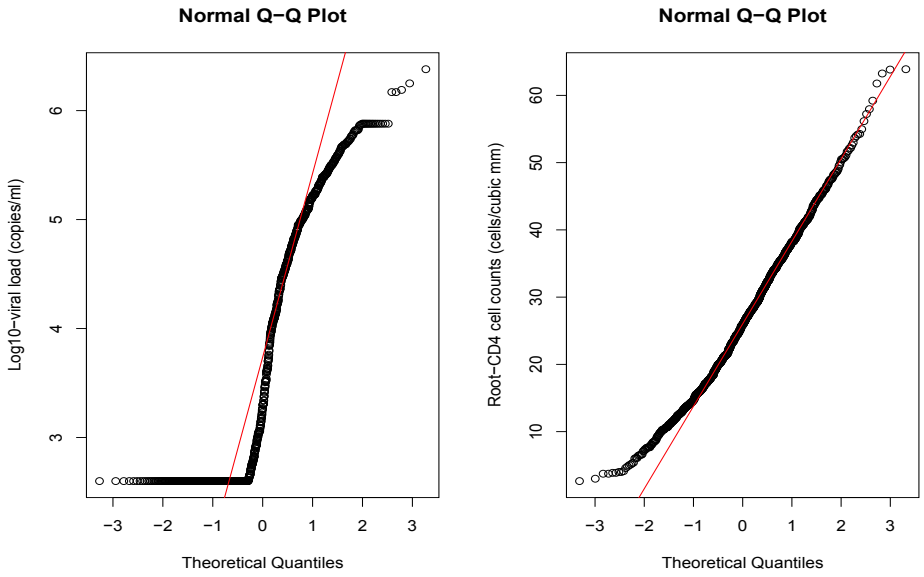


Figure 2.4: Checking for normality of the transformed markers.

coefficient for sex and that of time are significant for both markers. For the viral load measurements we note that age-group is also significant while for CD4+ T cell count we have facility to which the patients presented themselves as significant for the model.

More detailed covariate tests for goodness-of-fit to the models will be presented in the course of discussions in later chapters of this thesis.

Table 2.5: Summary of mixed-effects model for the markers (model (2.1)).

	log10(viral load)		root-CD4+ cell count	
	Value	p-value	Value	p-value
Intercept	4.508	< 0.001	20.52	< 0.001
Supp. treatment	-0.022	0.7471	1.310	0.0739
Compliance	-0.015	0.8658	-1.093	0.2518
Facility	-0.168	0.0729	3.569	< 0.001
Gender	0.243	< 0.001	-1.541	0.0339
Age	-0.069	0.0349	0.605	0.0804
time	-0.029	< 0.001	0.157	< 0.001

Chapter 3

A model for multivariate longitudinal data with complete outcomes with application to HIV disease dynamics

Abstract

Multivariate longitudinal outcomes are increasingly becoming common in many research studies especially in the biomedical and health research areas among others. Special methods are required to analyze such data because repeated observations on any experimental unit are likely to be correlated over time while multiple outcomes within the unit will also be correlated. We consider nonlinear mixed-effects models for complete multivariate longitudinal data. The motivation for the study of nonlinear mixed-effects models arose due to the growing interest in the estimation of parameters governing HIV disease dynamical models using real data. In this discussion these parameters are estimated using the stochastic approximation Expectation Maximisation (SAEM) algorithm in the presence of several covariates. We use routine observational HIV data with multivariate outcomes as an

application example.

3.1 Introduction

Typical longitudinal data from studies such as clinical trials or prospective observational studies consist of measurements or observations taken repeatedly over time on a particular subject. The objective of such studies is to assess changes and trends in response variables of study subjects over time. Longitudinal studies also offer an opportunity to compare the rates of response change of individual subjects on the same (or different) experimental treatment or under the same (or different) conditions. Furthermore, longitudinal analyses describe how these changes and trends in the response variable are related to explanatory variables of interest. This is a direct advantage of longitudinal studies over cross-sectional designs where measurements are taken only once. In the cross-sectional case a dataset does not offer the opportunity to analyse within individual changes due to treatment over time.

In longitudinal data the serial measurements on the same subject are bound to be dependent. With such cases it is reasonable to assume that observations that are closer together are likely to have higher correlations than those observations which are farther apart (Jones, 2000). The overall correlation arises from between-subject heterogeneity which is commonly included in the model in the form of random-effects, within-subject serial correlation and errors in the response variable due to measurement (Fitzmaurice et al., 2011).

In most clinical, epidemiological, psychometric and other social studies there is usually a crucial need to collect information on several outcomes on each occasion for a given subject. This gives rise to *multivariate* longitudinal data. Such data

are correlated over time while multiple outcomes from the same subject on a given occasion also tend to be correlated (Gueorguieva, 2001; Fieuws and Verbeke, 2006). In multivariate longitudinal data, a subset or subsets of the jointly observed outcomes may be similarly affected by the same model covariates. The resulting data, therefore, require special models so that parameters of the model can be estimated as efficiently as possible (Jia and Weiss, 2009). In this application, we assumed that the two markers are independently influenced by the covariates so that any correlation between the two is no caused by a third variable.

Multivariate longitudinal models have been widely discussed in the literature. In his contribution, Reinsel (1982) considered models for complete and balanced multivariate longitudinal data under the multivariate random-effects covariance structure. Using maximum likelihood estimation, he derived closed-form expressions for estimates of fixed-effects and covariance matrices for random-effects and for measurement errors. Jia and Weiss (2009) proposed a clustered common explanatory effects model for multivariate longitudinal data. The aim was to cluster outcomes that respond similarly to changes in values of explanatory variables. They developed an algorithm for fitting the model using standard software for univariate longitudinal data. The clusters of outcomes were selected by data using model selection tools specifically designed for this purpose.

To analyse complete continuous non-normal or categorical multivariate longitudinal data, Chaganty and Naik (2002) proposed a quasi-least squares approach. They obtained a set of objective functions in terms of correlation parameters. Furthermore, they described an algorithm for finding estimates of these parameters. They used several correlation structures to illustrate the method of minimisation described in their algorithm.

In modelling Alzheimer's disease, Beckett et al. (2004) considered a random-effects

approach for the association between the two characteristics of the response variable. Furthermore, they compared these joint results with those obtained by treating the models of individual characteristics as separate growth curves analysed using ordinary least squares approaches. Their study reveals that multivariate models reduce the effect of measurement error on the inference of correlation coefficients between rates of change in the markers of the response variable.

Oort (2001) described multivariate longitudinal data as being characterised by response variables, subjects under study and occasions of observation. He used structural equation modelling techniques to estimate parameters in the auto-regressive and latent curve models because it has been argued therein that the three-mode model is a special case of the linear latent variable model.

In this application we consider maximum likelihood parameter estimation strategy for nonlinear mixed-effects models for complete multivariate longitudinal data. The likelihood for nonlinear mixed effects models for balanced multivariate longitudinal responses is usually not tractable so that no analytic expressions for estimates are readily available and this becomes more complex with the increase in the number of simultaneous outcome measurements for a given subject. We, therefore, use the stochastic approximation Expectation-Maximisation (SAEM) algorithm as an estimation tool for parameters of nonlinear mixed-effects models (Delyon et al., 1999). The estimation method is applied to a routine observational dataset in the context of HIV dynamics and in our analysis we have included several covariates.

The rest of the chapter is organised as follows. Section 3.2 contains the motivation of the discussion through the HIV dynamical system and the parameters that characterise such a system. The nonlinear mixed-effects model for multivariate longitudinal outcomes is described in Section 3.3. The log-likelihood of the parameters given the response and random-effects is also presented in this section.

A procedure for the estimation of the parameters is described in Section 3.4. An analysis of a real bivariate observational dataset is given in Section 3.5. Section 3.6 concludes the chapter with a discussion.

3.2 The dynamical model for HIV

Biological models have considerably assisted in the study of disease dynamics and subsequent issues related to antiviral policy and treatment (Perelson and Nelson, 1999; Nowak and May, 2000). This has been made possible by considering estimates of the parameters that characterise such complex systems (Huang et al., 2006).

The study reported in this thesis has been motivated by the HIV dynamical system of nonlinear ordinary differential equations which models the pathogenesis of the disease in order to assess the effectiveness of antiviral therapies (Wu, 2005). This work aims at developing methodologies that could help in estimation of the parameters that characterise this system. To infer on the underlying HIV dynamics from the data on CD4+ T cell and viral load measurements we consider the latent dynamic model discussed by Lavielle et al. (2011). The model considered here distinguishes between the uninfected (T_N), latently infected (T_L) and activated infected (T_A) CD4+ T cell counts and the infectious virus particles (V_I) and non-infectious ones, (V_N). The T_N cells are continuously produced in the body (e.g. by the thymus) at a rate a and only a proportion π of infected CD4+ T cells are activated cells. Latently infected cells become activated at a rate of d. It is assumed that only activated CD4+ T cells produce virus particles. A description of the HIV dynamical system parameters is given in Table 3.1. This leads to the following 5-dimensional system of nonlinear ordinary differential equations for

Table 3.1: Parameters of the biological model for HIV.

Parameter	Description
a (cells/mm^3/day)	Rate of CD4+ T cell production
c_N (per day)	Death rate of uninfected CD4+ T cells
c_A (per day)	Death rate of actively infected CD4+ T cells
p (per day)	Number of virions produced by a CD4+ T cell
d	Activation rate of T_L cells
π	Proportion of activated infected CD4+ T cells
τ_{RTI}	Efficacy of reverse transcriptase inhibitor
τ_{PI}	Efficacy of protease inhibitor
	Fixed parameters
γ	Infection rate of T_N cells per virus particle (0.0021^*)
c_L (per day)	Death rate of latently infected CD4+ T cells ($0.0092^\dagger$)
c_V (per day)	Death rate of the virus particles ($30^\ddagger$)

* Lavielle et al. (2011); † Guedj et al. (2007b); ‡ Ribeiro et al. (2002)

individuals assumed to be on treatment:

$$\begin{aligned}
\dot{T_N} &= a - (1-\tau_{RTI})\gamma T_N V_I - c_N T_N \\
\dot{T_L} &= (1-\pi)(1-\tau_{RTI})\gamma T_N V_I - dT_L - c_L T_L \\
\dot{T_A} &= \pi(1-\tau_{RTI})\gamma T_N V_I + dT_L - c_A T_A. \\
\dot{V_I} &= (1-\tau_{PI})pT_A - c_V V_I \\
\dot{V_N} &= \tau_{PI} pT_A - c_V V_N
\end{aligned} \tag{3.1}$$

The parameters γ, c_L and c_V are the mass action terms between uninfected CD4+ T cells and the infectious virions, the death rate of unobservable latently infected cells and the death rate of the virus particles respectively. We note that the components $T_I,\ T_L$ and T_N constitute the observed CD4+ T cell counts and V_I and V_N constitute the viral load measurements.

As observed in other discussions, the system in equation (3.1) does not have an analytic solution (Xia and Moog, 2003; Huang et al., 2006; Wu et al., 2008). In dealing with this complex dynamic system the quality and informativeness of the data used to support the estimation of the system parameters is important. The usual practice, however, is to fix some parameters at set values (either from literature or expert opinion, see Table 3.1) if they are not estimable directly from the available information and estimate the rest of the parameters that can be supported by the data at hand.

In most discussions regarding viral dynamics it is assumed that before initiation of treatment the system is at equilibrium meaning that the viral load is in stable state where the viral production and clearance are balanced (Ribeiro, 2007). This means that all components of the HIV disease markers are all in stable state at this point. For the system presented in equation (3.1) the equilibrium values of the compartments are given by

$$T_{N(o)} = \frac{c_N c_V (d + c_L)}{\gamma p (d + \pi c_L)},$$

$$
\begin{aligned}
T_{L(o)} &= \frac{c_V V_{I(o)}}{p}, \\
T_{A(o)} &= \frac{(1-\pi)\gamma T_{N(o)} V_{I(o)}}{d + c_L}, \\
V_{I(o)} &= \frac{a - c_N T_{N(o)}}{\gamma T_{N(o)}}, \\
V_{N(o)} &= 0,
\end{aligned}
$$

where the constants have been described in Table 3.1. The actual measured or observed viral load is the total of the virus particles: infectious and no-infectious and the measured CD4+ T cell count is the total of uninfected, latently infected and actively infected CD4+ T cells. This system will be used in illustrations of the methodology developed in the subsequent chapters where we consider various aspects of the data.

3.3 The nonlinear mixed-effects model for multivariate outcomes

Often in the biomedical and biological studies information is collected from each of the individuals in the study with the objective of describing the response dynamics within the individual that govern the relationship between such response variables and the system covariates. These dynamics include pharmacokinetics, pharmacodynamics and HIV disease process. The last process has been described in the previous section. Making inference about parameters that govern such dynamics from data is a common challenge because of the general behaviour of nonlinear growth and decay curves including random-effects (or subject-specific effects) underlying these data. Furthermore, the data used in such inferences may not be rich enough to allow for the estimation of all parameters in a system. Such processes are best analysed using nonlinear mixed-effects models or hierarchical nonlinear

models (Davidian and Giltinan, 1995).

3.3.1 The model

Let k markers be jointly observed on each of the N individuals at $n_i = n$ occasions spanning the longitudinal study. Let y_{hij} denote the outcome of the hth marker on the ith subject at occasion j, $h = 1, 2, \ldots, k$, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, n$. Then $\mathbf{Y}_{hi}$ is an $n \times 1$ response vector for the hth marker for this subject so that the response is an $n \times k$ matrix given by

$$\mathbf{Y}_i = \begin{pmatrix} y_{1i1} & y_{2i1} & \cdots & y_{ki1} \\ y_{1i2} & y_{2i2} & \cdots & y_{ki2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1in} & y_{2in} & \cdots & y_{kin} \end{pmatrix}. \tag{3.2}$$

In this chapter we are estimating parameters of the disease process in equation (3.1) using balanced data where each subject is assumed to be observed on n occasions. This is in contrast to what we will consider in Chapters 4 and 5 where subject i will be available for observation on n_i ($< n$) occasions only.

This multiple-outcome response matrix can now be expressed in terms of the parameter vector and the explanatory variables so that it takes the form

$$\mathbf{Y}_i = g_i(\boldsymbol{\psi}_i, \mathbf{X}_i) + \boldsymbol{\epsilon}_i,$$

where $g_i(.)$ is an $n \times k$ matrix of functions such that at least one column is non-linear in the parameters $\boldsymbol{\psi}_i$ and the exploratory variables $\mathbf{X}_i$ (Cudeck, 1996). The quantity $\boldsymbol{\epsilon}_i$ is the corresponding $n \times k$ measurement error matrix related linearly to the response matrix $\mathbf{Y}_i$.

The common practice when working with multivariate response matrix in (3.2) is to form a new nk-dimensional vector $\mathbf{y}_i$ through a process called vectorization

(Reinsel, 1982; Shah et al., 1997; Marshall et al., 2006). This is a mathematical operation where columns of a matrix are stacked to form a vector with length corresponding to the size of the matrix so that

$$\mathbf{y}_i = \text{vec}(\mathbf{Y}_i) = (\mathbf{Y}_{1i}^T, \mathbf{Y}_{2i}^T, \ldots, \mathbf{Y}_{ki}^T)^T,$$

where T stands for transpose and each $\{\mathbf{Y}_{hi} : h = 1, 2, \ldots, k\}$ is an n-dimensional vector. Then in the spirit of Lindstrom and Bates (1990) the new response vector for the ith subject assumes a model of the form

$$\mathbf{y}_i = g(\boldsymbol{\psi}_i, \mathbf{X}_i) + \mathbf{e}_i, \tag{3.3}$$

where

$$g(\boldsymbol{\psi}_i, \mathbf{X}_i) = \{g_h(\boldsymbol{\psi}_i, \mathbf{X}_i) : h = 1, 2, \ldots, k\},$$

of which at least one is a nonlinear function of the parameter vector $\boldsymbol{\psi}_i$ and the model covariates $\mathbf{X}_i$ (also see Davidian and Giltinan, 1995). The quantity $\mathbf{e}_i$ (= vec($\boldsymbol{\epsilon}_i$)) is the random error vector reflecting uncertainty in the response vector as a result of measurement or observation. It is assumed that $\mathbf{e}_i \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ and that the measurement errors of different individuals are such that that $\text{cov}(\mathbf{e}_i, \mathbf{e}_{i'}) = 0$ for $i \neq i'$.

The parameter vector $\boldsymbol{\psi_i}$ takes into account the fixed-effects and random-effects. The procedure is to include this vector in model (3.3) through a model of the form

$$\boldsymbol{\psi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \tag{3.4}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of the fixed-effects, $\mathbf{b}_i$ is a $q \times 1$ vector of random-effects. The quantities $\mathbf{A}_i$ and $\mathbf{B}_i$ are known design matrices usually of the baseline covariates and they link the fixed-effects and random-effects, respectively, to the vector $\boldsymbol{\psi_i}$.

As in a discussion by Liu and Wu (2007), we assume that $\mathbf{b}_i$ has a multivariate normal distribution with mean zero and a variance-covariance matrix $\mathbf{G}$ and that the random-effects and the error terms for a particular subject are also not correlated so that $\text{cov}(\mathbf{e}_i, \mathbf{b}_i) = 0$. From model (3.4) we note that

$$\boldsymbol{\psi}_i \sim \text{MVN}(\mathbf{A}_i\boldsymbol{\beta},\ \mathbf{B}_i\mathbf{G}\mathbf{B}_i^T),$$

because it is linear in $\mathbf{b}_i$. This assumption implies that one can find an expression for a generalised least squares estimate of $\boldsymbol{\beta}$ which takes the form

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{A}_i^T \mathbf{V}_{\psi_i}^{-1} \mathbf{A}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{A}_i^T \mathbf{V}_{\psi_i}^{-1} \boldsymbol{\psi}_i, \tag{3.5}$$

where $\mathbf{V}_{\psi_i} = \mathbf{B}_i\mathbf{G}\mathbf{B}_i^T$ and the quantity $\boldsymbol{\psi}_i$ is unobserved just like the random-effects. This complicates the estimation process especially in situations where some of the response values are missing due to dropout or presence of left-censored data. In this chapter we are considering a case where data are balanced so that no response values are missing.

The choice of the structure of the covariance matrix $\boldsymbol{\Sigma}_i$ matters in parameter inference because it determines the actual number of unknown model parameters to be estimated. One of the convenient choices is to use the covariance matrix of the measurement error of the univariate response (Shah et al., 1997; Marshall et al., 2006). For instance, if one considers the jth row of the error matrix $\boldsymbol{\epsilon}_{i(j)}$ and assumes a normal distribution with mean zero and $k \times k$ covariance matrix $\boldsymbol{\Sigma}$, then the overall error covariance matrix of has the form

$$\text{cov}(\mathbf{e}_i) = \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \otimes \mathbf{I}_{n\times n},$$

where $\otimes$ denotes the usual Kronecker product and $\mathbf{I}$ is the identity matrix. The implication of this assumption is that the error terms for the markers of a particular response variable from the same subject have a covariance matrix $\boldsymbol{\Sigma}$ and the rest are assumed to be mutually independent.

3.3.2 Parameter estimation in the nonlinear mixed-effects model

The estimates of the parameters $\boldsymbol{\beta}$ and the covariance matrices $\mathbf{G}$ and $\boldsymbol{\Sigma}_i$ are obtained using numerical approximation techniques because of the nature of the nonlinear model besides the fact that the random-effects are treated as missing data. There are two classes of parameter estimation methods for nonlinear mixed-effects models. The first and possibly the most popular class involves linearisation of the regression model with respect to the random-effects so that parameter estimation can be implemented in standard software (Lindstrom and Bates, 1990; Vonesh and Carter, 1992; Davidian and Giltinan, 1995; Marshall et al., 2006). The main disadvantage of these methods is that they give inconsistent parameter estimates notably in studies with few occasions. The other class involves approximating the likelihood function by integrating out the unobserved quantities by either using the Gaussian quadrature like in Walker (1996) and Nummi and Möttönen (2000) or stochastic methods (Kuhn and Lavielle, 2005; Samson et al., 2006). In this application we use one of the latter methods called the stochastic approximation version of the Expectation-Maximisation (SAEM) algorithm which we describe in Section 3.4.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{G}, \boldsymbol{\Sigma}_i)$ be the vector of the parameters characterising the two-stage model (3.3) and (3.4) and assuming that the response variable is such $\mathbf{y}_i|\boldsymbol{\psi}_i \sim \mathrm{N}(\boldsymbol{\mu}_i(\boldsymbol{\psi}_i, \mathbf{x}_{ij}), \boldsymbol{\Sigma}_i)$ where

$$\boldsymbol{\mu}_i(\boldsymbol{\psi}_i, \mathbf{x}_{ij}) = g(\boldsymbol{\psi}_i, \mathbf{x}_{ij}),$$

and

$$\boldsymbol{\Sigma}_i = \mathrm{var}(\mathbf{y}_i|\boldsymbol{\psi}_i).$$

Then the joint density of the response and random-effects is given by

$$f(\mathbf{y}_i, \mathbf{b}_i|\boldsymbol{\theta}) = f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\Sigma}_i)p(\mathbf{b}_i|\mathbf{G}).$$

We need the maximum likelihood estimate of $\boldsymbol{\theta}$ and this is obtained by using the likelihood of the observations which is given by

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \int f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\Sigma}_i)p(\mathbf{b}_i|\mathbf{G})d\mathbf{b}_i.$$

Since maximum likelihood estimation is known to underestimate the covariance components, in this chapter we use the restricted maximum likelihood estimation so that the complete sequence log-likelihood for estimation of the parameters takes the form

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}_i) &= \text{constant} - \frac{n}{2}\log|\mathbf{G}| - \frac{N_a}{2}\log|\boldsymbol{\Sigma}_i| \\ &\quad -\frac{1}{2}\sum \mathbf{b}_i^T\mathbf{G}^{-1}\mathbf{b}_i - \frac{1}{2}\sum(\mathbf{y}_i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i), \end{aligned} \tag{3.6}$$

where N_a is the total of the k-tuples observed for all N subjects. To estimate the parameters of the model we need to account for the random-effects which are regarded as missing data. This is achieved by considering the conditional likelihood of $\mathbf{b}_i$ given $\mathbf{y}_i$ and this is given by

$$p(\mathbf{b}_i|\mathbf{y}_i) = \frac{\prod_{i=1}^{n} f(\mathbf{y}_i, \mathbf{b}_i|\boldsymbol{\theta})}{\int \prod_{i=1}^{n} f(\mathbf{y}_i, \mathbf{b}_i|\boldsymbol{\theta})d\mathbf{y}_i}. \tag{3.7}$$

In the algorithm we describe in the next section the procedure is to maximise the expectation of $l(\boldsymbol{\theta}|y)$ with respect to the conditional density of the random-effects.

3.4 Parameter estimation using the SAEM algorithm

There are a number of approximation procedures that are applicable to the estimation of parameters in cases where closed-form expressions are not attainable. One of the most widely used techniques in statistical approximations is the EM (**E**xpectation-**M**aximisation) algorithm which was introduced by Dempster et al. (1977). The EM algorithm is an efficient iterative procedure for computing the maximum likelihood (ML) or restricted maximum likelihood (REML) estimates in the presence of missing or hidden data. Each iteration of the EM algorithm consists of two steps: the E-step, and the M-step. In the E-step, the missing data are estimated given the observed data and current estimates of the model parameters. This is achieved by using the conditional expectation of sufficient statistics for parameters (especially those of the covariance matrices $\mathbf{G}$ and $\mathbf{\Sigma}$) given the observed data and the current values of these parameters. In the rth iteration of the M-step, parameters are found by equating them to expectations of their sufficient statistics. The estimated data from the E-step are used in lieu of the actual unobserved data. There are cases where this approximation technique cannot be applied because the E-step may not give closed form expressions that may be used in the M-step. Thus other methods of estimating the dynamical system parameters have to be considered.

In this thesis we use the stochastic approximation version of the EM algorithm (SAEM). This is a parameter estimation method which was proposed by Delyon et al. (1999). It consists of replacing the E-step of the EM algorithm by two steps:

Step 1 simulation of the missing data using a priori density of the missing values

given the observed values and initial parameter values and;

Step 2 updating the set of sufficient statistics of the unknown parameters like the covariance matrices.

The second step is followed by the maximization step. It has an advantage of converging to maximum likelihood estimates faster that the EM algorithm. In this algorithm, simulated missing values are used in the evaluation of the quantity that is eventually maximised to give parameter estimates and these simulated values are gradually discarded. It can be implemented in `R` and `MONOLIX`. Codes of this algorithm have also been proposed for other statistical software like `SAS`. In this thesis, we have implemented this estimation method in `R` and `MONOLIX`.

Kuhn and Lavielle (2005) applied this algorithm in finding maximum likelihood estimates for nonlinear mixed-effects models for univariate longitudinal data. In their contribution Samson et al. (2006) extended the SAEM algorithm to accommodate left-censored data in a nonlinear mixed-effects model as an exact maximum likelihood estimation method and illustrated this extension with a simulation study in the HIV dynamics context. We extend the application of the SAEM algorithm to multivariate longitudinal outcomes with several covariates where random-effects and measurement errors are treated as unobserved. Then the SAEM algorithm is implemented as follows.

Step 1 Let iterations be indexed by $r = 0, 1, \ldots, \infty$ with $r = 0$ corresponding to initial values assigned to $\boldsymbol{\theta}$, the vector containing components of the covariance matrices $\boldsymbol{\Sigma}_i$, $\mathbf{G}$ and $\boldsymbol{\beta}$. Then $\boldsymbol{\theta}^{(r)}$ denotes the value of $\boldsymbol{\theta}$ at the end of the rth iteration. The complete data for finding the estimates consist of $\mathbf{y}_i$ and $\mathbf{b}_i$ with log-likelihood given in (3.6).

Step 2 Simulation Step : Let $p(.|\mathbf{y}_i, \boldsymbol{\theta}^{(r)})$ be a conditional density of the unobserved random-effects as given in (3.7). Then Simulate m(r) values of the unobserved random-effects $(\mathbf{b}_i^{(r+1)})$ of size $m(r)$ from this distribution.

Step 3 Stochastic Approximation : Define $\mathbf{s}^{(r)}$ by

$$\begin{aligned} \mathbf{s}^{(r)} &= \mathrm{E}[l(\boldsymbol{\theta})|\mathbf{b}_i, \boldsymbol{\theta}^{(r)}] \\ &= \int l(\boldsymbol{\theta}) p(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(r)}) d\mathbf{b}_i. \end{aligned}$$

Then the approximation step involves updating $\mathbf{s}^{(r)}$ to $\mathbf{s}^{(r+1)}$ according to the relation

$$\mathbf{s}^{(r+1)} = \mathbf{s}^{(r)} + \delta_{r+1}\left(\mathbf{S}(\mathbf{y}, \mathbf{b}_i^{(r+1)}) - \mathbf{s}^{(r)}\right), \tag{3.8}$$

where δ_{r+1} is a step-size scalar such that $\delta_{r+1} \to 0$ as $r \to \infty$ and the quantity $\mathbf{S}(\mathbf{y}, \mathbf{b}_i^{(r+1)})$ is given by

$$\mathbf{S}(\mathbf{y}, \mathbf{b}^{(r+1)}) = \frac{\sum_1^{m(r)} l(\mathbf{y}_i, \mathbf{b}_i^{(r)}, \boldsymbol{\theta}^{(r+1)})}{m(r)}.$$

This step reduces to updating sufficient statistics $\sum_{i=1}^N \mathbf{e}_i\mathbf{e}_i^T$ and $\sum_{i=1}^N \mathbf{b}_i\mathbf{b}_i^T$ of the complete model. At iteration $(r+1)$, these quantities are updated as follows

$$\begin{aligned} \mathbf{s}_{1,(r+1)} &= \mathbf{s}_{1,(r)} + \delta_{r+1}\left(\frac{\sum_1^{m(r)}(\mathbf{b}_i^{(r+1)}(\mathbf{b}_i^{(r+1)})^T)}{m(r)} - \mathbf{s}_{1,(r)}\right) \\ \mathbf{s}_{2,(r+1)} &= \mathbf{s}_{2,(r)} + \delta_{r+1}\left(\frac{\sum_1^{m(r)}(\mathbf{e}_i^{(r+1)}(\mathbf{e}_i^{(r+1)})^T)}{m(r)} - \mathbf{s}_{2,(r)}\right). \end{aligned} \tag{3.9}$$

Step 4 M-Step: In this step we find the values of $\hat{\boldsymbol{\beta}}$, $\mathbf{G}^{(r+1)}$ and $\boldsymbol{\Sigma}_i^{(r+1)}$ that maximize the updated quantity in Equation (3.8). Elements of $\mathbf{G}^{(r+1)}$ and $\boldsymbol{\Sigma}_i^{(r+1)}$ are found by using their sufficient statistics. The iterative equation for estimating the covariance of random-effects is given by

$$\hat{\mathbf{G}}^{(r+1)} = \frac{\sum_1^N \left(\mathrm{E}[\mathbf{b}_i\mathbf{b}_i^T|\mathbf{y}_i, \boldsymbol{\theta}^{(r)}]\right)}{N - pk}. \tag{3.10}$$

From our assumption of the covariance matrix $\boldsymbol{\Sigma}_i$ we can find particular elements on the diagonal by using the relation

$$\hat{\boldsymbol{\Sigma}}^{(r+1)} = \frac{\sum_i^N \mathrm{E}[\mathbf{e}_i \mathbf{e}_i^T | \mathbf{y}_i, \boldsymbol{\theta}^{(r)}]}{Nn - pk},$$

so that $\hat{\boldsymbol{\Sigma}}_i^{(r+1)} = \hat{\boldsymbol{\Sigma}}^{(r+1)} \otimes \mathbf{I}_{n \times n}$. The updated value of $\boldsymbol{\beta}$ is then found by using the relation from (3.5) and is given by

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\sum \mathbf{A}_i^T (\mathbf{V}_{\psi_i}^{(r+1)})^{-1} \mathbf{A}_i \right)^{-1} \sum \mathbf{A}_i (\mathbf{V}_{\psi_i}^{(r+1)})^{-1} \boldsymbol{\psi}_i^{(r+1)}.$$

The values found at the M-step are then used in the S-step and this procedure is repeated until a predetermined convergence criterion is satisfied.

Starting values can be obtained in a number of ways mostly dependent on the choice of the model and the nature of the dataset. In some cases identity matrices are used as starting covariance matrices at the first iteration of the algorithm. But these may not be the most appropriate choice especially in terms of the number of iterations needed to attain convergence (Shah et al., 1997). For other forms and techniques for obtaining starting values see Laird et al. (1987) and Harville (1977). In this thesis we have used identity covariance matrices since the SAEM algorithm speeds up convergence to maximum likelihood estimates.

This algorithm does not provide standard errors of the parameter estimates at the end of the iterations (Vermunt, 2004). Thus one would get these quantities by finding the observed information matrix through the matrix of second-order derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$. The evaluation of this matrix is complex because it does not have a closed-form expression. Delyon et al. (1999) proposed a stochastic approximation of the Fisher information matrix as follows. Let $l_o(\boldsymbol{\theta})$ and $l_c(\boldsymbol{\theta})$ be the log-likelihood functions of the observed data and complete data, respectively. Then using the modified approach in Samson

et al. (2007) we have

$$\partial l_o(\boldsymbol{\theta}) = \mathrm{E}[\partial l_c(\boldsymbol{\theta})|\mathbf{y}_{io}, \boldsymbol{\theta}],$$

and the Hessian of $L(\boldsymbol{\theta})$ can be written as

$$\partial^2 l_o(\boldsymbol{\theta}) = \mathrm{E}[\partial^2 l_c(\boldsymbol{\theta})] + \mathrm{var}(\partial l_c(\boldsymbol{\theta})),$$

where the partial differentials are with respect to $\boldsymbol{\theta}$. Thus at the end of the rth step of the algorithm one gets the following approximations

$$\boldsymbol{\Lambda}_{r+1} = \boldsymbol{\Lambda}_r + \delta_{r+1}\left(\frac{1}{m(r)}\sum_{1}^{m(r)} \partial l_c(\mathbf{y}_{io}, \mathbf{w}_i^{(r)}; \boldsymbol{\theta}^{(r)}) - \boldsymbol{\Lambda}_r\right),$$

and

$$\begin{aligned}\boldsymbol{\upsilon}_{r+1} &= \boldsymbol{\upsilon}_r + \delta_{r+1}\left(\frac{1}{m(r)}\sum_{1}^{m(r)} \Big(\partial^2 l_c(\mathbf{y}_{io}, \mathbf{w}_i^{(r)}; \boldsymbol{\theta}^{(r)}) \right.\\ &\quad \left. + \partial l_c(\mathbf{y}_{io}, \mathbf{w}_i^{(r)}; \boldsymbol{\theta}^{(r)}) \partial (l_c(\mathbf{y}_{io}, \mathbf{w}_i^{(r)}; \boldsymbol{\theta}^{(r)}))^T\Big) - \boldsymbol{\upsilon}_r\right),\end{aligned}$$

where

$$\boldsymbol{\Lambda}_{r+1} = \mathrm{E}[\partial l_c(\boldsymbol{\theta})|\mathbf{y}_{io}, \boldsymbol{\theta}],$$

and

$$\boldsymbol{\upsilon}_{r+1} = \mathrm{E}[(\partial^2 l_c(\boldsymbol{\theta}))] + \mathrm{var}(\partial l_c(\boldsymbol{\theta})).$$

Using these expressions we get the quantity

$$H_{r+1} = \boldsymbol{\upsilon}_{r+1} - \boldsymbol{\Lambda}_{r+1}\boldsymbol{\Lambda}_{r+1}^T. \tag{3.11}$$

As the sequence $\{\boldsymbol{\theta}_{r+1} : r \geq 0\}$ converges to maximiser of the complete-data log-likelihood given in equation (3.6), the sequence $\{H_{r+1} : r \geq 0\}$ given by equation (3.11) converges to the Fisher information matrix and the inverse of this matrix provides the estimated variance-covariance matrix for the parameter estimates. It should be noted that the information matrix is also used for checking model identifiability (Vermunt, 2004).

3.5 Application

3.5.1 The Statistical model

The observable markers are resultants of the compartments of the HIV disease dynamical model introduced in Section 3.2. Thus $V = V_I + V_N$ and $T = T_N + T_L + T_A$ are the total observed viral loads and the CD4+ T cell counts for the ith subject at occasion j respectively. These repeated marker data are usually skewed and as such the modelled components $g_1(.)$ and $g_2(.)$ are transforms of the observed total viral load and CD4+ T cell count such that $g_1(V) = \log_{10}(V_I + V_N)$ and $g_2(T) = (T_N + T_L + T_A)$. In this thesis, however, we used an identity transformation for g_2 because the usual transformations of square-root or fourth-root for the CD4+ T cell count provided parameter estimates that are either very large or too small compared to those in the literature. This justifies the choice of a constant error term for the log10-viral load and the proportional error model for CD4+ T cell count.

Thus if we let $\mathbf{y}_i$ denote, as defined in model (3.3), the measurements for subject i then the bivariate response variable can be written as $\mathbf{y}_i = (y_{1i}, y_{2i})$ where

$$\begin{aligned} y_{1i} &= g_1(V(t_{ij}, \boldsymbol{\psi_i})) + e_{Vi},\ j \leq n \ \text{ and} \\ y_{2i} &= g_2(T(t_{ij}, \boldsymbol{\psi}_i)) + g_2(T(t_{ij}, \boldsymbol{\psi}_i))e_{Ti} \end{aligned} \quad , \tag{3.12}$$

where n in the number of occasions at which each subject is observed and e_{i1} and e_{i2} are measurement error vectors such that

$$e_{i1} \sim \mathrm{N}(\mathbf{0}, \sigma_1 \mathbf{I}_{n_{i1}}) \text{ and } e_{i2} \sim \mathrm{N}(\mathbf{0}, \sigma_2 \mathbf{I}_{n_{i2}}).$$

The quantity $\boldsymbol{\psi}_i$ is the vector of parameters of the HIV dynamical system described

in Table 3.1. It is included in model (3.12) through the relation

$$\begin{array}{rcl} \log_{10}(\boldsymbol{\psi_1}) & = & \beta_{01} + \sum_{l=1}^{M} \beta_{l1} X_i + b_{i1},\ \boldsymbol{\psi}_1 = (a, c_N, c_A, p, d) \\ \text{logit}(\boldsymbol{\psi_2}) & = & \beta_{02} + \sum_{l=1}^{M} \beta_{l2} X_i + b_{i2},\ \boldsymbol{\psi}_2 = (\tau_{RTI}, \tau_{PI}, \pi) \end{array} , \tag{3.13}$$

where the elements of $\boldsymbol{\psi}_h$ have been defined in Section 3.2 and M is the number of model covariates. The elements of $\boldsymbol{\psi}_1$ have log-transforms because they have to be positive and those of $\boldsymbol{\psi}_2$ are defined as inverse logistic transformations of a normal random variable because as proportions they assume their values in the interval $(0, 1)$. The quantities b_{i1} and b_{i2} represent random-effects with a normal distribution. It will be noted that models (3.12) and (3.13) define the hierarchical nonlinear mixed-effects model that we described in equations (3.3) and (3.4) in Section 3.3.

There is usually a sharp increase in CD4+ T cell count and rapid decline of the viral load in the first few weeks after treatment initiation and flatten out thereafter (Ma et al., 2008). Thièbaut et al. (2003) observed that this happens in or about the first forty days of treatment commencement. Unlike in planned clinical studies, like in Wu (2005) and Huang et al. (2006), our data do not give us enough ground to consider a piecewise linear formulation because of lack of enough measurements in the first few weeks after initiation of treatment as our data are observational in nature. Moreover, we are mainly concerned with post-infection characteristics of the HIV disease including treatment dynamics.

3.5.2 Data

HIV patients at the Lighthouse Trust Clinics (in Malawi) are routinely monitored to check how they are responding to treatment. In this application, there were many subjects whose information was mostly available at the time of collection.

For most patients the clinical information including CD4+ cell counts and viral load measurements were collected at the treatment initiation. However, not all patients had frequent and exact measurements of viral loads. This is possibly because quantifying viral load is usually influenced by the cost and availability of resources (Romih et al., 2010). This suggests the reason for having more CD4+ cell count measurements for most patients compared to viral load measurements.

Thus in the current application only subjects with five measurements of both the CD4+ cell count and viral load (including baseline values) were included. This ensured that we had balanced bivariate longitudinal data in line with the objective of the current application. There were 78 patients with this information representing about 40% of those that had at least three measurements of both the CD4+ T cell count and viral load. The total number of clinical measurements in this representative sample was 780 for both CD4+ cell counts and viral load. The threshold value for viral measurements for the assay used in the quantification was 400 copies per ml. About forty (39.5%) percent of the viral load measurements were below the threshold. Potentially these are left-censored observations and a model to address this problem can be proposed. In the current discussion we took 400 copies per ml as the true observation. This problem will be given due attention in Chapters 4 and 5.

3.5.3 Implementation

We implemented the algorithm in `MONOLIX` and `R`. Both are free statistical software. The former was developed by INRIA (Institut National de la Recherche en Informatique et Automatique) and is obtainable from `www.lixoft.com`. It handles a number of algorithms that are used in the estimation of parameters in nonlinear

mixed-effects models including MCMC. In this application we used `MONOLIX` version 4.1. The latter software was developed by Ross Ihaka and Robert Gentleman and is currently maintained by the `R` Development Core Team.

3.5.4 Results

Tests were conducted on the data in order to identify covariates that are significant to the bivariate response described in the previous section. We did this by comparing models (in particular the model in equation 3.12) with and without covariates. There were eleven sets of covariates to be tested and the results are displayed in Table 3.2. The table only shows results for significant covariates based on the likelihood ratio tests (LRT). From the significant covariate sets shown in the table, we also considered the Akaike information criterion (AIC) as a criterion for choice of the best set. The CT-A combination was selected on account of the values of its AIC which is 6785.48 and this was the smallest among the covariate sets that were tested.

Table 3.2: Choice of model covariates based on model (3.12): A = age-group; C = compliance; F = facility; S = sex; T = treatment.

Covariate set	FS-A	CT-A	CS-A	CFT-A	CST-A
$p-$value (LRT)	0.0456	0.0139	0.0002	0.0122	0.0279
AIC	6809.74	6785.48	6793.14	6815.20	6819.45

We estimated the parameters of model (3.13) using the HIV observational bivariate longitudinal data described in Section 3.5.2. The effects of covariates on the response variable are presented in Table 3.3 and these results are based on model

(3.13) for the two parameters on treatment efficacy. Coefficients of the other parameters can be determined in similar manner. Based on the values of β, it can be seen that of all the age-groups the > 40 group has a bigger absolute contribution to both the treatment efficacy of the protease inhibitor τ_{PI} and that of the transcriptase inhibitor (τ_{RTI}). But this could be significant probably for transcriptase inhibitor only as seen from the size of standard error. In the HIV dynamic model

Table 3.3: Estimates of regression coefficients of the efficacy model (3.13).

Covariate		β_{PI}	(s.e.)	β_{RTI}	(s.e.)
Age-group	$26-33$	1.06	(0.983)	-4.64	(5.251)
	$34-40$	-4.16	(2.065)	-0.693	(0.110)
	over 40	-7.78	(9.1)	6.3	(3.472)
Compliance to treatment		-2.07	(1.086)	0.912	(0.070)
Supplementary treatment		3.04	(3.56)	1.71	(1.411)

the death rate of the virus particles, c_V, was fixed at 30 copies per day based on the literature (Ribeiro et al., 2002). There were two other sets of parameters that did not permit simultaneous estimation and as a result, two more parameters (c_L and γ) were fixed (see Table 3.1). With these fixed values, the other dynamic system parameters were estimated and are presented in Table 3.4.

The results indicate that for the current data, CD4+ T cells were produced at a rate of 11.4 cells/mm^3 per week or 1.63 cells/mm^3 per day which is within the range of estimated values in the literature (Guedj et al. (2007a) estimated the value as 1.67 cells/mm^3 per day and Lavielle et al. (2011) got 2.62 cells/mm^3 per day). These minor differences could be as a result of differences in marker transformations, models used in the estimations and possibly differences in threshold values

for the viral load. Another reason for these variations could result from differences in HIV subtypes with same overall dynamics but differing disease progression rates. Furthermore, datasets used in the estimation of the parameters are obtained under different conditions and assumptions and the resulting estimates are bound to display such differences. For instance in our case, we had included several covariates which was not the case in some discussions like the one by Lavielle et al. (2011). Of particular interest were the values of the treatment efficacy τ_{PI} and τ_{RTI} (0.998 and 0.972 respectively) which compared well with those in Huang and Lu (2008). Figure 3.1 demonstrates individual fits for log10-viral loads for four subjects.

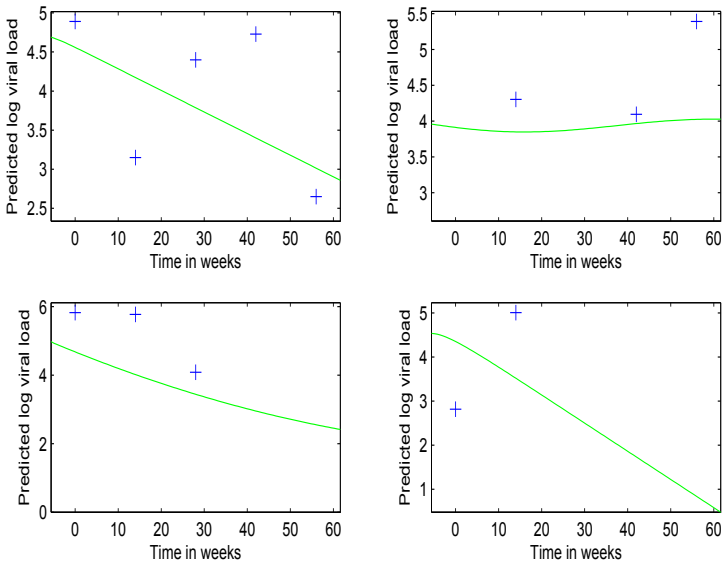


Figure 3.1: Individual fits of the log10-viral load.

We also computed parameter estimates for each age-group and very interesting results were obtained. For efficacy of treatment, it was observed that the group $26-33$ gave estimates that did not deviate significantly (with $\tau_{PI} = 0.99$ and $\tau_{RTI} = 0.796$) from the global values as compared to the other two age-groups.

The estimates for the older group (> 40) showed a wide departure in the value of τ_{PI} compared to the overall value of 0.998 obtained in Table 3.4. This prompted

Table 3.4: Estimates of parameters of the HIV dynamical model (and their std. errors).

Parameter	Estimated value	standard error
a	11.4	5.3
c_N	4.08×10^{-7}	–
c_A	0.272	0.08
p	159	169.801
d	3.02×10^{-7}	–
π	0.779	0.402
τ_{RTI}	0.998	0.17895
τ_{PI}	0.972	0.1145

us to take a closer look at the correlation coefficients between the markers and the individual estimates of the efficacy for each of the three groups. The results of the correlation analysis suggest that lower correlations between τ_{RTI} and CD4+ T cell count values imply lower efficacy for the protease inhibitors and a higher efficacy for the transcriptase inhibitors. This could suggest that lower values of τ_{RTI} result in higher values of the viral load and lower values of the CD4+ T cell count. This is understood from clinical perspective where a poor performing drug results in viral rebound. The markers themselves were negatively correlated, with small variations among age-groups and in the present balanced data the value of the overall correlation coefficient was -0.5204. This can also be looked from the exploratory analysis point of view where it has been shown that the older age-group had higher CD4+ T cell counts on average (see Figure 2.2). It should be

pointed out that τ_{PI} did not provide results that could be directly interpreted and as a result its correlation values with markers for the three age-groups have not been included in Table 3.5.

Table 3.5: Correlation coefficients between efficacy and markers in age-groups.

	Age-group	$26-33$	$34-40$	>40
	Efficacy	τ_{RTI}	τ_{RTI}	τ_{RTI}
		0.796	0.995	1
Correlation	CD4+ T cell count	0.1523	0.0596	-0.3433
	Log10-viral load	-0.1478	0.0281	0.0645

We also looked at the relationship between the two covariates (compliance and supplementary treatment) and $\tau_{eff}(=\tau_{PI}$ or $\tau_{RTI})$. The results indicate that for τ_{PI} both supplementary treatment and compliance to treatment are significant with $p-$values of < 0.001 and 0.0024 respectively. For τ_{RTI}, however, results indicate that only supplementary treatment was significant with a p-value of < 0.001. These results are indicative of the importance of the two covariates to disease treatment efficacy as described by the dynamic system presented here.

3.6 Discussion

In this chapter, we have considered nonlinear models for multivariate longitudinal data as applied to estimation of parameters in a disease dynamic model for HIV. We have studied the efficacy of treatment and other dynamic parameters in the presence of a multiplicity of covariates. In particular, we have shown that fac-

tors like age of a patient at initiation of treatment, compliance to treatment and supplementary treatment have a significant bearing on an individual's response to treatment. We are mindful of the fact that inclusion of many covariates in a complicated nonlinear model for the dynamic system could make computation of parameters more demanding, but inclusion of those significant variables, as demonstrated in this discussion, will make estimation and inference to be more meaningful and practical.

The analyses and approximation procedures presented here can be generalised to find applications in other disease modelling problems. For instance, in modelling other chronic illnesses like cancer one may determine the efficacy of treatment and the impact of treatment compliance on response to treatment regimen (among other system parameters) by applying the procedures discussed in this review.

In the analyses, we assumed that all subjects had their markers measured at the same time. That is, we have assumed a complete and balanced bivariate dataset, where all measurements are taken at the same time for all the subjects which amounts to a complete case analysis. This obviously leads to loss of information from subjects who were partially observed. However, in most practical settings like clinical, epidemiological, psychometric and other social studies it is not uncommon to have unbalanced data where subjects are measured or observed at different time points and with different numbers of occasions.

We have also observed that compliance to treatment is a significant covariate in the bivariate data that we have looked at. As an indicator of treatment compliance, absence at appointment (or scheduled) times could not adequately reflect actual compliance profiles for individual subjects. Absence at time of appointment may not imply that the patient was not complying to treatment. This means that the data quality would affect accuracy of estimation for dynamic system parameters

if the covariates are not properly defined and modelled (Huang and Lu, 2008). It would be worthwhile to re-define this factor (treatment compliance) to include such situations as (i) a patient is checked to be taking prescribed medication at appointed times, (ii) recommended dose and diet is being taken by the patient, (iii) prescribed schedules of patient physical exercises (or related health maintenance or improvement activities) are being followed by the subject under study and (iv) any treatment related advice from the doctor or medical practitioner is being adhered to (Pullar et al., 1989). A proper documentation of getting supplementary treatment would also be a great advantage so that it is more accurately modelled.

For some subjects, the viral load values were frequently below the threshold value and this scenario is known for creating an artificial level-off effect in a dataset. If these censored data are not modelled accordingly, they usually result in some parameters being either underestimated or overestimated (Wu, 2005). We suspect that it was such high proportions of threshold values that led to a few parameters having estimates that departed considerably from those in the literature.

On the whole, however, these shortcomings did not outweigh our observations and results that the parameters in the dynamic systems are better estimated or approximated with reasonable practical reliability if system covariates are included in the analyses. Our results compared very well with other published results. As an example, the number of virus particles produced a CD4+ T cell, p, is within values found in some discussions (104 in Guedj et al. (2007b) and 183 in Yu and Liang (2013)).

Chapter 4

Modelling multivariate longitudinal data with dropout with application to HIV disease dynamics

Abstract

The main challenge in biomedical and clinical studies which involve collection of longitudinal data is the premature withdrawal of the subjects from the study resulting in incomplete data. Standard statistical analysis approaches usually give biased estimates of the model parameters if the mechanisms that led to dropout are ignored. In this chapter we consider nonlinear mixed-effects models for multivariate longitudinal data in the presence of subject dropout. We present techniques for estimation of model parameters. These procedures are applied to estimation of parameters in the HIV dynamical system using routine observational data from an HIV clinic.

4.1 Introduction

Many biomedical and epidemiological studies are designed to collect data that consist of measurements or observations taken repeatedly over time on a particular subject. In clinical trials, for instance, disease progression or effect of treatment are regularly monitored by observing a disease marker over time. The objective of such studies is to assess changes and trends in response variables of study subjects over time.

However, occurrence of incomplete data is not uncommon in these study designs because not all subjects are available for observation or measurement at scheduled appointment times. The incomplete data can arise from intermittent missingness as a result of logistical problems and ill-health or completely dropping out of the study due to attrition or withdrawal because of drug toxicity in case of clinical trials or indeed due to manifestation of little improvement in the presence of treatment (Fitzmaurice, 2003; Liu and Wu, 2007; Ibrahim and Molenberghs, 2009). This results in subject i having only $n_i \leq n$ observations where n is the number of visits intended for each subject in the original study design.

The problem of incomplete data due to intermittent missingness and dropout is also common in observational studies involving a prospective follow-up of patients for some health outcome. The challenge in model formulation for incomplete longitudinal data is the need to take into consideration the latent causes of missing values and the assessment of the resulting biases (Hogan et al., 2004; Gad and Ahmed, 2006; Molenberghs and Kenwrad, 2007). In practice, however, this missing data mechanism is usually not fully specified and this poses statistical challenge.

The dropout mechanism may not be influenced by the values of the response variable (observed or unobserved) on a study subject. For instance, a subject may

drop out of the study because of the occurrence of an unforeseen event like a job transfer to a new work place far from the study location. In this case, the data are said to be missing completely at random (MCAR). When the probability of missing mechanism depends on the observed measures only and not on the missing values, the data are said to be missing at random (MAR). The other scenario is when the probability of non-response is a function of the unobserved data (Little and Rubin, 1987; Rubin, 2004). In this case the data are said to be missing not at random (MNAR). For instance, missing not at random (or non-ignorable missingness) can occur when the subject's outcome trend has a direct bearing on their tendency for early withdrawal from the study (Fitzmaurice and Laird, 2000; Pantazis et al., 2005). In some cases, the absence of a marker reading results from censoring due to lower detection limits of the assays used in quantifying the markers like viral load levels.

The nature of the missingness processes has implications on estimation of parameters in a longitudinal data model. It has been suggested in the literature that likelihood-based estimation techniques can provide accurate inferences by ignoring the missingness process when data are missing completely at random or missing at random (Diggle and Kenward, 1994). However, when missingness mechanism is not missing completely at random, caution must be exercised in building models so that these missingness processes are reflected in the models to reduce estimation biases. There are several procedures that have been discussed in the literature for the estimation of parameters in the models of univariate longitudinal data with missing values (Hu and Sale, 2003; Roy, 2003; Molenberghs et al., 2004; Wu and Wu, 2007).

However, most clinical trials and epidemiological studies with a long time course have designs involving measuring or observing several markers that characterise

the response variable related to progression of the disease or indeed any biological process of interest (Pantazis et al., 2005; Cai et al., 2010). Just like in the univariate case, the occurrence of dropouts is a common challenge in the estimation of parameters of the models for multivariate longitudinal data. In fact this presents an additional complication because the multiple outcomes within the response variable will most likely be correlated in addition to the problem of potentially missing values in all the outcome variables. There are few discussions in the literature on the analysis of multivariate longitudinal data with subject dropout. Roy and Lin (2002) modelled the relationship between a latent variable (a variable that is not directly observed) and covariates using a linear mixed model. They assumed that the dropout process depends on the latent variable and applied the selection model in order to account for non-ignorable missing data. They also used a transition model (a model where previous outcomes are used as predictors) in order to accommodate missing covariates in their analysis.

In this chapter we propose strategies for the estimation of parameters in disease dynamical systems using nonlinear mixed-effects models of multivariate longitudinal data in the presence of dropout or monotone missingness (Wu, 2002). Parameter estimation procedures are proposed for implementation using the stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999). In this chapter, we specifically consider a case where dropout requires that all the markers of the response variable are not available after a subject drops out. We also include several covariates which we assume to be completely observed and time invariant. We then estimate parameters characterising HIV disease dynamics in the presence of subject dropout.

We give a brief statement on the motivation of the study in Section 4.2. A nonlinear mixed-effects model for multivariate longitudinal data is presented in Section 4.3.

In Section 4.4 a dropout model for multivariate longitudinal data is given and a joint distribution of the dropout mechanism and multivariate longitudinal response is proposed. A brief overview of the approximation procedure is also outlined in this section. This methodology is illustrated in Section 4.5 using an observational dataset and we close this chapter with a discussion and conclusion in Section 4.6.

4.2 Motivation of the discussion

The study reported in this chapter has been motivated by the latent HIV dynamic system of nonlinear ordinary differential equations which models the pathogenesis of the disease and this has been described in Section 3.2 of Chapter 3. It has the form

$$\dot{\mathbf{F}}(\mathbf{T}, \tilde{\mathbf{V}}) = \mathbf{f}(\mathbf{T}, \tilde{\mathbf{V}}|\boldsymbol{\psi}), \tag{4.1}$$

where $\mathbf{T} = (T_N, T_L, T_A)$ is the set of the uninfected , latently infected and activated infected CD4+ T cells and $\tilde{\mathbf{V}} = (V_I, V_N)$ are infectious virus particles and non-infectious ones. The quantities $\dot{\mathbf{F}}$ and $\mathbf{f}$ represent the left-hand and right-hand members, respectively, of the HIV dynamic system in equation (3.1). The set $\boldsymbol{\psi} = (a, c_N, c_A, c_L, \gamma, p, \pi, \tau_{RTI}, \tau_{PI})$ of constants characterises this dynamical system and are described in Table 3.1. The interest is in finding suitable models which can be used in the identification and estimation of parameters governing this dynamical system while taking into account dropout processes . As can be seen from Figure 4.1, most subjects included in the analysis did not have all their markers observed at scheduled consultation times. Most of them dropped out before the designated observation period and this could be attributed to a number of reasons which may include lack of favourable response to treatment, change of location or indeed loss due to death (a full account of reasons for dropping out is given in the previous

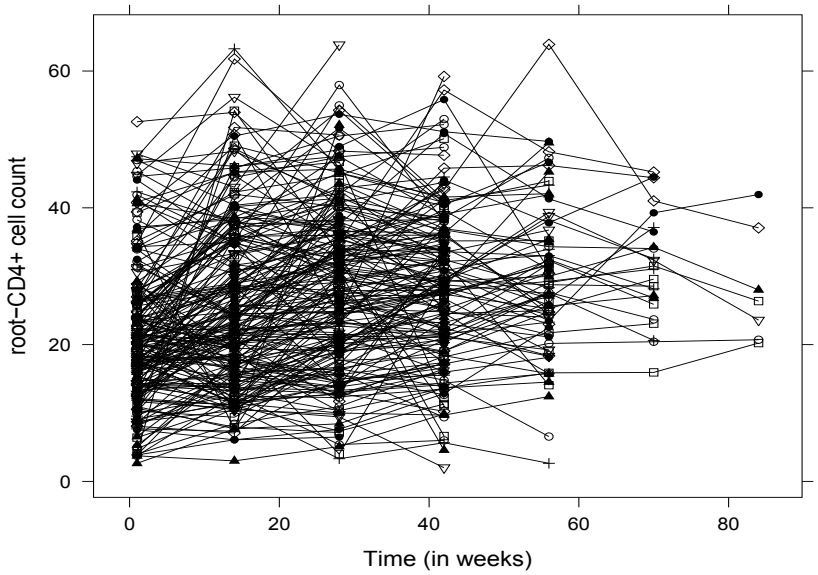


Figure 4.1: Time-plot for root-CD4+ T cell counts: illustrating subject dropout.

section). As pointed out in the introduction one cannot fully discern all the reasons for dropout a prior, hence the need for a modelling approach.

4.3 The nonlinear mixed-effects model for multivariate longitudinal data

Inference about parameters that govern biomedical and biological processes like pharmacokinetics, pharmacodynamics and viral disease process from data is a common challenge because of the general behaviour of nonlinear growth and decay curves underlying these data. Such processes are best analysed using nonlinear mixed-effects models or hierarchical nonlinear models (Davidian and Giltinan, 1995). The general nonlinear model has been described in Section 3.3 of the previous chapter.

4.4 Modelling the dropout process

In many biomedical, epidemiological and biological studies and particularly those studies involving disease history which span long periods of time the aim is to collect repeated measurements of the outcomes that characterise the disease process. The main challenge in such designs is the occurrence of missing data which results from missing visits or prematurely withdrawing from the study. In this section we propose a model for the dropout process for multivariate longitudinal data based on the work of Diggle and Kenward (1994).

Subjects drop out of the study due to a multiplicity of reasons like death or toxicity in the case of clinical trials (Liu and Wu, 2007). The assumption is that for a particular subject all the markers are either available or are not available after dropout. Let n be the number of occasions marked for observation (measurement) and $n_i \leq n$ be the number actually observed. Thus $n_i = n$ means a complete scenario and $n_i < n$ corresponds to a dropout for the ith subject.

4.4.1 The dropout model

Let $C_i = n_i + 1$ be the indicator of the occasion of dropout where $C_i \leq n$ and $C_i = n + 1$ means a complete case. We assume that measurements have been obtained at baseline on all subjects before they drop out of the study since a unit without an observation does not have any contribution to the model analysis and that the k response characteristics are measured at the same time. Suppose $\mathbf{Y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in})^T$ is an $n \times k$ matrix of complete k-outcome measurements of the ith subject and $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in})^T$ be defined as the corresponding vector of measurement time points. Let the matrix of the observed multiple outcomes be given by $\mathbf{Y}_{i,o} = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in_i})^T$ and the corresponding missing data matrix be

denoted by $\mathbf{Y}_{i,m} = \left(\mathbf{y}_{i(n_i+1)}, \mathbf{y}_{i(n_1+2)}, \ldots, \mathbf{y}_{in}\right)^T$. Define j' as an occasion such that $2 \leq j' < n$. Then models for $\mathbf{Y}_i$ and $\mathbf{Y}_{i,o}$ that satisfy the condition

$$\mathbf{y}_{ij} = \begin{cases} \mathbf{y}_{ij,o} & j = 1, 2, \ldots, C_i - 1, \\ \mathbf{y}_{ij,m} & j \geq C_i, \end{cases}$$

can be formulated where C_i has been defined above (Diggle and Kenward, 1994).

Recall that the $\mathbf{y}_i$ has three sets of parameters associated with it as given in the previous chapter: $\boldsymbol{\beta}$ which represents the fixed-effects, $\boldsymbol{\Sigma}_i$ the covariance matrix of the measurement error and $\mathbf{G}$ the covariance matrix of the random-effects (see Section 3.3). The inferences about model parameters are done based on the density of complete data. Let $f(\mathbf{y}; \boldsymbol{\theta})$ denote this vector-valued density function under the distribution of model (3.3), where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \mathbf{G})$. Let

$$\mathbf{H}_{ij'} = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{i(j'-1)})$$

be the observed history of the response variable for subject i up to occasion j' (or up to time $t_{j'-1}$) and $\mathbf{y}_{ij'}$ be the unobserved k-dimensional response vector at time $t_{j'}$. Then a model can be proposed so that the monotone missingness process is such that the conditional probability of dropout depends on the observed outcomes up to time point t_{c_i}. Thus, for $c_i \leq n$ we have that

$$\Pr(C_i = c_i | \mathbf{H}_{c_i}) = p_{c_i}(\mathbf{H}_{c_i}, \mathbf{y}_{c_i}; \boldsymbol{\lambda}), \tag{4.2}$$

where $\boldsymbol{\lambda}$ is a vector of the unknown constants that characterise the dropout process and p_{c_i} is the value of the probability mass function at c_i.

Let $f_{j'}(\mathbf{y}|\mathbf{H}_{j'})$ be the conditional density function of the overall complete response vector given history

$$\mathbf{H}_{j'} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{j'-1}),$$

and also suppose $f_o(\mathbf{y}|\mathbf{H}_{j',o})$ is the conditional density of $\mathbf{Y}_o$ given the corresponding history $\mathbf{H}_{j',o}$. By definition of dropout of subject on occasion j' we have that

$$\Pr(\mathbf{Y}_{j'} = \mathbf{Y}_m|\mathbf{H}_{j'}, \mathbf{Y}_{j'-1} = \mathbf{Y}_m) = 1,$$

where $\{\mathbf{Y}_{j'} = \mathbf{Y}_m|\mathbf{H}_{j'}, \mathbf{Y}_{j'-1} = \mathbf{Y}_m\}$ is the event that a subject is not observed on occasion j' given that it is not observed on occasion $j'-1$. And the conditional probability of the missing sequence on occasion j' given that it was not missing on the previous occasion has the form

$$\Pr(\mathbf{Y}_{j'} = \mathbf{Y}_m|\mathbf{H}_{j'}, \mathbf{Y}_{j'-1} \neq \mathbf{Y}_m) = \int p_{j'}(\mathbf{H}_{j'}, \mathbf{y}; \boldsymbol{\lambda}) f_{j'}(\mathbf{y}|\mathbf{H}_{j'}; \boldsymbol{\theta}) \mathrm{d}\mathbf{y}_o. \tag{4.3}$$

The relation in model (4.3) suggests that in the formulation $\mathbf{Y}_c$ and Y_o are taken as processes. By extension, we find an expression for the conditional density function of the observed sequence, $\mathbf{Y}_o$, given the history of the response variable and dropout model parameters and it is expressed as

$$f_o(\mathbf{y}|\mathbf{H}_{j'}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = (1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}; \boldsymbol{\lambda}))\, f_{j'}(\mathbf{y}|\mathbf{H}_{j'}; \boldsymbol{\theta}).$$

Let $\mathbf{w}_i$ be a set of unobserved data which includes the random-effects and the unobserved response values. Then the joint distribution of the complete data under this dropout setting is given by

$$f(\mathbf{y}, \mathbf{b}_i) = \prod_{j'=2}^{n} (1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}_{j'}, \boldsymbol{\lambda})) f_{j'}(\mathbf{y}_o, \mathbf{w}_i|\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{4.4}$$

where $\boldsymbol{\theta}$ has been defined above. Following the arguments in Hu and Sale (2003), the model for the incomplete set of outcomes with dropout occurring at occasion c_i,

$$\mathbf{Y} = (\mathbf{Y}_{1,o}, \mathbf{Y}_{2,o}, \ldots, \mathbf{Y}_{c_i-1,o}, \mathbf{Y}_{c_i,m}, \mathbf{Y}_{c_i+1,m}, \ldots, \mathbf{Y}_{n,m})$$

is given by

$$f(\mathbf{y}_i) \;=\; f(\mathbf{y}_1) \left(\prod_{j'=2}^{c_i-1} f_{j'}(\mathbf{y}_{j'}|\mathbf{H}_{j'}) \right) \Pr(\mathbf{Y}_{j'} = \mathbf{Y}_m|\mathbf{H}_{c_i})$$

$$= f_{c_i-1}((\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{c_i-1})^T) \prod_{j'=2}^{c_i-1} [1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}_{j'})] \quad (4.5)$$
$$\times \Pr(\mathbf{Y}_{j'} = \mathbf{Y}_m | \mathbf{H}_{c_i}),$$

with f_{c_i-1} denoting the density function of the observed data. We note that the relation in model (4.5) includes subject-specific effects through $f_{c_i-1}(.)$. Therefore, the conditional density of the missing sequences given the observed response values is given by

$$p(\mathbf{w}|\mathbf{y}_o) = \frac{\prod_{j'=2}^{n}(1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}_{j'}, \boldsymbol{\lambda}))f_{j'}(\mathbf{y}_o, \mathbf{w}_i|\boldsymbol{\theta}, \boldsymbol{\lambda})}{\int \prod_{j'=2}^{n}(1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}_{j'}, \boldsymbol{\lambda}))f_{j'}(\mathbf{y}_o, \mathbf{w}_i|\boldsymbol{\theta}, \boldsymbol{\lambda})d\mathbf{w}}. \quad (4.6)$$

This function is used in the algorithm which we have described earlier in Chapter 3 as a priori density for the unobserved values. The difference in the setting with that scenario is that in the present chapter $\mathbf{w} = (\mathbf{y}_m, \mathbf{b}_i)$ while in that case $\mathbf{w}$ was equal to the vector of random-effects $\mathbf{b}_i$ since the data were balanced. These expressions provide a basis for the estimation of the response and dropout model parameters. However, modelling dropouts of multivariate longitudinal data poses challenges because of the increased number of parameters characterising the dropout mechanism as a result of the possibility of correlations among the k outcomes.

4.4.2 The likelihood function

The previous section gives us the tools for the formulation of the likelihood function that takes into account the response profile and the dropout mechanism. In reality, however, the dropout mechanism is not fully identified because the underlying causes may not be known with certainty. This likelihood function will be used in the estimation of model parameters. In particular we need expressions for the components of the right hand of equation (4.5) that may constitute the joint

likelihood of the multiple-outcome response variable and the dropout parameter vector $\boldsymbol{\lambda}$.

For the dropout probability given in equation (4.2) one considers a logistic linear model of the form

$$\text{logit}(p_{j'}(\mathbf{H}_{j'}, \mathbf{y}; \boldsymbol{\lambda})) = \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_1 \mathbf{y} + \sum_{j=2}^{j'} \boldsymbol{\lambda}_j \mathbf{y}_{j'+1-j}, \tag{4.7}$$

where $\mathbf{H}(.)$ and $\boldsymbol{\lambda}$ are the response history and dropout parameter vector respectively as defined above. The parameter vector $\boldsymbol{\lambda}$ can also be chosen such that it is a function of the covariates that may be time variant. This would result in probability of dropout increasing with time. For ease of exposition and parameter identification, we assume $\boldsymbol{\lambda}$ depends on covariates only through the response variable. When $\boldsymbol{\lambda}_1 \neq \mathbf{0}$ the dropout process is said to be informative because the link function in equation (4.7) depends on the set of missing values in the form of a single unobserved value.

Assume that the conditional density of response vector (see model (3.3)) given random-effects is Gaussian with mean $\boldsymbol{\mu}_i = g(\boldsymbol{\psi}_i, \mathbf{X}_i)$ and corresponding covariance matrix $\boldsymbol{\Sigma}_i$. Then letting $f(\mathbf{y}_i)$ to be the joint probability density of the $c_i - 1$ available measurements for subject i and using the normality assumption we have the modified log-likelihood given by

$$\begin{aligned} \log(f(\mathbf{y}_i)) &= \text{constant} - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\log|\mathbf{G}| - \frac{1}{2}\mathbf{b}_i^T \mathbf{G}^{-1}\mathbf{b}_i \\ &\quad -\frac{1}{2}\Big(\mathbf{y}_i - g(\boldsymbol{\psi}_i, \mathbf{X}_i))^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - g(\boldsymbol{\psi}_i, \mathbf{X}_i)\Big), \end{aligned} \tag{4.8}$$

where $\boldsymbol{\psi}_i$ and $\mathbf{G}$ are as defined in Section 3.3 in the previous chapter. It will also be noted that for computational purposes, it is easier to work with the conditional density of the response variable given random-effects than to work with its marginal density (Lee and Nelder, 2004). The latter case would have a covariance matrix of

the form $\mathbf{V}_i = \mathbf{R}_i + \mathbf{\Sigma}_i$ where $\mathbf{R}_i = \text{var}[g(\boldsymbol{\psi}, \mathbf{X}_i)]$. From model (4.7) we have

$$\log(1 - p_{j'}(\mathbf{H}_{j'}, \mathbf{y}_{j'})) = -\log(1 + \exp(\boldsymbol{\lambda}_0 + \sum_{j=1}^{j'} \boldsymbol{\lambda}_j \mathbf{y}_{j'+1-j})). \tag{4.9}$$

We can now use the relation in equation (4.5) to get the likelihood function for all subjects in terms of the parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$). This function is given by

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \sum_{i=1}^{N} \log f(\mathbf{y}_i) + \sum_{i=1}^{N} \sum_{j'=2}^{c_i - 1} \log(1 - p_{j'}(\mathbf{H}_{ij'}, \mathbf{y}_{i'})) \\ &\quad + \sum_{i}^{N} \log \Pr(C_i = c_i | \mathbf{y}_i) \\ &= l_1(\boldsymbol{\theta}) + l_2(\boldsymbol{\lambda}) + l_3(\boldsymbol{\theta}, \boldsymbol{\lambda}). \end{aligned} \tag{4.10}$$

The terms $l_1(\boldsymbol{\theta})$ and $l_2(\boldsymbol{\lambda})$ can be explicitly expressed using equations (4.8) and (4.9) respectively. In similar manner expression for l_3 can be derived and simplified from equations (4.2) and (4.3). This means that separate maximisation of l_1 and $l_2 + l_3$ can be carried out. Usually, there is no closed-form expression for such integrals and thus one applies approximation procedures in the analyses. We use the stochastic approximation EM (SAEM) algorithm for maximisation of the overall log-likelihood.

4.4.3 Estimation using the SAEM algorithm

The aim of this subsection is to consider a procedure that can be used in finding estimates of the parameters that characterise the nonlinear model for the multiple-outcome response and the dropout process using the log-likelihood given in equation 4.10. In this approximation the idea is to use the expectation of $l(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to the conditional density given in equation 4.6.

In the present discussion we use the stochastic approximation expectation maximisation algorithm (SAEM) which was proposed by Delyon et al. (1999). It consists

of replacing the E-step of the EM (Expectation-Maximisation) algorithm by two steps: simulation of the missing data using a priori density and updating the set of sufficient statistics of components of $\mathbf{V}_i$. This step is followed by the maximization step.

We are going to use the same notation as in Section 3.4 of the previous chapter. However, the expectation of the complete log-likelihood is now conditioned on $\mathbf{w} = (\mathbf{b}_i, \mathbf{y}_m)$ instead of being conditioned on $\mathbf{b}_i$ alone. It will be noted that the conditional density of the missing sequence depends on $\boldsymbol{\psi}_i$ through the quantity $\mathbf{b}_i$. Thus in this algorithm, we use the quantity

$$\begin{aligned} s(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \mathrm{E}[l(\boldsymbol{\theta}, \boldsymbol{\lambda})|\mathbf{y}_o, \boldsymbol{\lambda}', \boldsymbol{\theta}'] \\ &= \int \left[\sum_{i=1}^{N} \log f(\mathbf{y}_i) + \sum_{i=1}^{N} \sum_{j'=2}^{c_i - 1} \log(1 - p_{j'}(\mathbf{H}_{ij'}, \mathbf{y}_{i'})) \right. \\ &\quad \left. + \sum_{i}^{N} \log \Pr(C_i = c_i|\mathbf{y}_i) \right] p(\mathbf{w}|\mathbf{y}_o) d\mathbf{w}, \end{aligned} \tag{4.11}$$

where $\boldsymbol{\lambda}'$ and $\boldsymbol{\theta}'$ denote the current values of the parameters and $p(\mathbf{w}|\mathbf{y}_o)$ is given by equation (4.6).

Index iterations by $r = 0, 1, \ldots, \infty$ and let $\boldsymbol{\theta}^{(r)}$, $\boldsymbol{\lambda}^{(r)}$ be estimates of the parameters at the end of the rth iteration. The algorithm proceeds as follows:

Step 1 Initialise $r = 0$**:** $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$. Find estimates of $\boldsymbol{\theta} = (\boldsymbol{\Sigma}, \mathbf{G}, \boldsymbol{\beta})$ and $\boldsymbol{\lambda}$ that maximise the quantity in equation (4.10).

Step 2 Simulation: Draw sequence $(\mathbf{w}^{(r+1)})$ of size $m(r)$ from a conditional density $p(.|\mathbf{y}_o, \boldsymbol{\theta}^{(r)})$ of the unobserved quantity, $\mathbf{w}_i = (\mathbf{b}_i, \mathbf{y}_m)$.

Step 3 Stochastic approximation: Update the expectation of quantity in equation (4.11) with respect to $p(.|\mathbf{y}_{i,o}, \boldsymbol{\theta}^{(r)}, \boldsymbol{\lambda}^{r})$ using

$$\mathbf{s}^{(r+1)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{s}^{(r)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \delta_{r+1} \left(\mathbf{S}(\mathbf{y}_o, \mathbf{w}^{(r+1)}; \boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathbf{s}^{(r)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \right),$$

where $\delta_{r+1} \in (0,1)$ is a smoothing parameter and decreases as $r \to \infty$. It helps to accelerate convergence to maximum likelihood estimates (Meza et al., 2007). The term $\mathbf{S}(\mathbf{y}_o, \mathbf{w}^{(r+1)}; \boldsymbol{\theta}, \boldsymbol{\lambda})$ is given by

$$\mathbf{S}(\mathbf{y}_o, \mathbf{w}^{(r+1)}; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{\sum_1^{m(r)} l(\mathbf{y}_o, \mathbf{w}^{(r)}; \boldsymbol{\theta}^r, \boldsymbol{\lambda}^r)}{m(r)}.$$

The quantity $m(r)$ is usually picked such that its value is between 5 and 30 (Meza et al., 2007).

Step 4 M-Step: Find

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}) = \arg\max \mathbf{s}^{(r+1)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$$

Step 5 Repeat steps 1-4 until

$$\mathbf{s}^{(r+1)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathbf{s}^{(r)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) < \boldsymbol{\eta},$$

for some $\boldsymbol{\eta} > \mathbf{0}$.

The estimate for $\boldsymbol{\beta}$, the vector of the fixed-effects, is obtainable from the likelihood function in (4.8) such that at the $(r+1)$th iteration

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\sum_{i=1}^{N} \mathbf{A}_i^T (\mathbf{V}_{\psi_i}^{-1})^{(r+1)} \mathbf{A}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{A}_i^T (\mathbf{V}_{\psi_i}^{-1})^{(r+1)} \boldsymbol{\psi}_i^{(r+1)},$$

where $\mathbf{V}_{\psi_i} = \text{var}(\boldsymbol{\psi}_i)$ (see equation 3.5 in previous chapter).

As advanced by Meza et al. (2007), under the modified maximum likelihood estimation the quantities $\mathbf{G}$ and $\boldsymbol{\Sigma}_i$ are estimated using their respective statistics, $\sum_1^N \mathbf{b}_i \mathbf{b}_i^T$ and $\sum_i^N \mathbf{e}_i \mathbf{e}_i^T$, so that at the $(r+1)$th iteration their estimates are given by

$$\hat{\mathbf{G}}^{(r+1)} = \frac{\sum_1^N \left(\text{E}[\mathbf{b}_i^{(r+1)} (\mathbf{b}_i^{r+1})^T | \mathbf{y}_i, \boldsymbol{\theta}^{(r)}] \right)}{N - pk} \quad \text{and}$$

$$\hat{\boldsymbol{\Sigma}}_i^{(r+1)} = \frac{\sum_i^N \text{E}[\mathbf{e}_i^{(r+1)} (\mathbf{e}_i^{(r+1)})^T | \mathbf{y}_i, \boldsymbol{\theta}^{(r)}]}{N n_i - pk},$$

respectively, where p is the length of the vector of fixed-effects in equation (3.4). The estimates of the error terms $\mathbf{e}_i$ are found by the predicted values of the response variables so that $\mathbf{e}_i = \mathbf{y}_i - \hat{g}(\boldsymbol{\psi}_i, \mathbf{X}_i)$ where $\hat{g}(\boldsymbol{\psi}_i, \mathbf{X}_i)$ is given by

$$\hat{g}(\boldsymbol{\psi}_i, \mathbf{X}_i) = \mathrm{E}(g(\boldsymbol{\psi}_i, \mathbf{X}_i)|\mathbf{y}_i, \boldsymbol{\theta}^{(r+1)}),$$

when we use equation 3.3 in Chapter 3 (for example, see Meza et al., 2012). Starting values can be obtained in a number of ways mostly dependent on the choice of the model and the nature of the dataset (Dempster et al., 1977; Laird et al., 1987). In this application, we used the identity matrices for as starting values for covariance matrices as done in a discussion by Shah et al. (1997).

4.5 Application

The objective of the study was to estimate the parameters in the HIV dynamic system of nonlinear ordinary differential equations described in Section 4.2. We were also interested in analysing the influence of different explanatory variables on the dynamic system parameters through the significance of their coefficients.

4.5.1 Data

We used routine observational data collected from two HIV/AIDS clinics under the Lighthouse Trust of Kamuzu Central Hospital in Malawi. In this application we included all subjects with three up to seven bivariate measurements (including the baseline values). There were two hundred and fourteen (214) patients with this information and between them they accounted for a total of 1856 clinical measurements of the markers. This represents an average dropout proportion of 38%. The severity of dropout is presented in Table 4.1 where the numbers of

subjects still present on each occasion are recorded (also see Figure 4.1 in Section 4.2).

Table 4.1: Number of subjects by week

Week	0	14	28	42	56	70	84
No. patients	214	214	214	177	78	24	7
Percentage	100	100	100	82.7	36.4	11.2	3.3

The limit of quantification (LOQ) for the viral loads for this set of data was 400 copies per ml. About 39% of viral load measurements were below this threshold value. The scenario is illustrated in the distribution function of the log10-viral loads shown in Figure 4.2 where the height of the vertical line at 2.6021 ($\log_{10}(400)$) corresponds to the cumulative probability (of 0.39) for the threshold values.

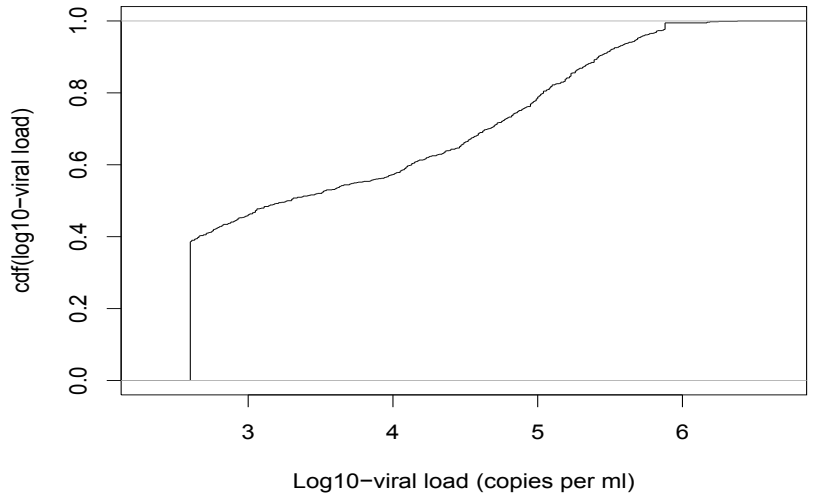


Figure 4.2: Distribution function of log10-viral load showing proportion of values below LOQ.

There were five covariates collected along with the disease markers and tests were carried out in order to chose the most appropriate set of covariates to be included in the analysis. Five such covariate sets were tested for their significance in the response model (3.13) of Chapter 3 (see also the next subsection) and parameter model (4.12). These were TCFS-A, TFS-A, TCS-A, FCS-A and TFC-A, where T = supplementary treatment, C = compliance to treatment, F = facility to which the patient reported, S = gender of the patient and A = age-group. This differs from the covariate sets in Chapter 3 because in the current chapter we have sets of up to four covariates included in the model as guided by the outcomes of the tests. The only significant combinations were TFS-A and TCS-A with p-values of < 0.001 and 0.0018 respectively. From exploratory cross-tabulation analysis of the covariates, it was noted that there was high dependence between compliance to treatment and facility to which the patients presented themselves with a p-value of $p < 0.001$ and a correlation coefficient of -0.6221. The exploratory results also revealed significant evidence (with a p-value 0.0016) of association between the facility of choice and age at treatment initiation. On the basis of this inter-dependence, therefore, we decided to include age-group, supplementary treatment, compliance to treatment and gender as explanatory variables in the analyses that follow.

4.5.2 Models for response variable and system parameters

The occasion spacing was about fourteen weeks on average and this implies that most of the first viral load measurements and CD4+ T cell counts were obtained several weeks after treatment initiation. Thus using the current data, it would not be possible to estimate the entire set of model parameters in the absence of such values (Putter et al., 2002; Guedj et al., 2011). In this analysis, therefore,

three parameters were fixed in order to facilitate identification and estimation of the other parameters (see Table 3.1 in Chapter 3).

The modelling for the system parameters proceeds as in model (3.13) of the previous chapter. The difference is that in this setting we have introduced four explanatory variables (of age-group, supplementary treatment, compliance to treatment and gender). Thus the effect of the covariates on the set of parameters could be assessed through the model given in the following equation

$$\begin{aligned} \log_{10}(\boldsymbol{\psi_1}) &= \beta_{01} + \sum_{l=1}^{M} \beta_{l1} X_i + b_{i1}, \ \boldsymbol{\psi}_1 = (a, c_N, c_A, p, d) \\ \text{logit}(\boldsymbol{\psi_2}) &= \beta_{02} + \sum_{l=1}^{M} \beta_{l2} X_i + b_{2i}, \ \boldsymbol{\psi}_2 = (\tau_{RTI}, \tau_{PI}, \pi) \end{aligned}$$

where X_i represents the covariates and M is the number of covariates. Other quantities remain as defined earlier (see model (3.13)).

The observed marker values are totals of the components of the dynamic system introduced in Section 3.2 of the previous chapter. That is, the observed CD4+ T cell count is the total $T = T_A + T_N + T_L$ and the observed viral load is the total viral load $V = V_I + V_N$. Thus $\mathbf{Y}_{ih} = g(\boldsymbol{\psi}_i, \mathbf{X}_i) + \mathbf{e}_{ih}, \ h = 1, 2$ where

$$g(\boldsymbol{\psi}_i, \mathbf{X}_i) = (\log_{10}(V(t_i, \boldsymbol{\psi}_i)), T(t_i, \boldsymbol{\psi}_i))^T, \tag{4.12}$$

the term $\mathbf{e}_{ih} = (e_{i1}, T(t_i, \boldsymbol{\psi}_i) e_{i2})$ is the measurement error with mean-zero normal distribution and the quantity $\boldsymbol{\psi}_i$ is described in model (3.13).

4.5.3 Results

We estimated the parameters of the biological system presented in Section 3.2 (see model (3.13)). Because the current data could not adequately support estimability of all the parameters of this complex dynamic system three of these parameters were fixed at values from the literature as shown in the following Table 4.2.

Table 4.2: Fixed system parameters.

Par.	Description	Value	Source of publication
γ	Infection rate of T_N cells	0.0021 per virion	Lavielle et al. (2011)
c_L	Death rate of T_L cells	0.0092 per day	Guedj et al. (2007b)
c_V	Death rate of virions	30 per day	Ribeiro et al. (2002)

The estimation procedure was applied to an observational bivariate longitudinal dataset which was characterised by high proportions of dropouts and viral load values that were below the lower limit value of 400 copies per ml. The subject dropouts compounded by these viral load values presented challenges in estimation of covariate coefficients because standard errors of some of the coefficients could not be directly produced when the SAEM algorithm was applied. This was also the reason why some parameters of the HIV dynamical system were fixed (see Table 4.2). The other reason could be the misspecification of the parameter link functions. To circumvent this estimation problem for the coefficients, we used individual parameter estimates so that the effects of the covariates on the model parameters could be estimated with their standard errors. These results are presented in Table 4.3. The reference age-group for this analysis was $18-25$. From the results in this table (top part) it can be observed that compared to other age-groups, $34-40$ class had significant contribution to the coefficients of almost all parameters. Older age (over forty years) also tended to have insignificant influence on τ_{RTI} (with a p-value of 0.9818) relative to the other age-groups each of which had p-value smaller that 0.001. There was no immediate explanation for this result.

It can also be observed from this table that the covariates, supplementary treat-

Table 4.3: Estimates of the covariate coefficients (p-values).

Variable		a	c_N	d	τ_{RTI}	τ_{PI}
	Inter.	1.116	-4.566	-0.577	0.737	0.172
		(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(0.001)
	$26-33$	-0.084	-0.294	0.027	-1.026	0.439
		(0.07)	(< 0.001)	(0.74)	(< 0.001)	(< 0.001)
Age-group	$34-40$	-0.123	-0.482	-0.167	-0.430	0.788
		(0.015)	(< 0.001)	(0.06)	(< 0.001)	(< 0.001)
	over 40	-0.021	0.276	0.249	-0.001	1.003
		(0.67)	(< 0.001)	(0.004)	(0.9818)	(< 0.001)
Supplementary		0.113	0.035	0.107	0.158	-0.681
		(0.001)	(0.023)	(0.064)	(< 0.001)	(< 0.001)
Compliance		-0.137	0.923	-0.176	-0.174	0.414
		(< 0.001)	(< 0.001)	(0.003)	(< 0.001)	(< 0.001)
Sex of patient		-0.037	0.076	-0.108	0.254	-0.888
		(0.24)	(< 0.001)	(0.056)	(< 0.001)	(< 0.001)

ment, compliance and gender, were significant for τ_{RTI} and τ_{PI}. Of particular note is compliance to treatment which had significant coefficients for all the parameters. Thus it could be argued that compliance to treatment reflects that HIV therapy is effective and also reduces the risk of developing drug resistance (van der Eijk et al., 2005).

We also computed system parameters and their 95% confidence limits based on the age-groups with 18-25 as the reference group. These results reveal an improvement in the standard errors of the parameter estimates and their corresponding confidence limits compared to the complete case analysis we had in the previous chapter. This has probably come about as a result of the inclusion of a number of explanatory variables and accounting for missing data and this has improved the analysis as compared to results in the discussions where a complete-case-analysis was assumed. Table 4.4 summarises the results for four selected parameters as grouped by age at treatment initiation. We noted, however, that there were wider

Table 4.4: System parameter estimates (95% confidence limits).

Age	a	d	τ_{RTI}	τ_{PI}
$26-33$	11.49	0.27	0.55	0.4
	$(8.59, 14.39)$	$(0.20, 0.34)$	$(0.42, 0.67)$	$(0.31, 0.48)$
$34-40$	11.62	0.26	0.59	0.73
	$(7.93, 15.31)$	$(0.16, 0.35)$	$(0.43, 0.75)$	$(0.55, 0.92)$
Over 40	14.25	0.69	0.69	0.87
	$(10.25, 18.25)$	$(0.47, 0.91)$	$(0.52, 0.87)$	$(0.67, 1)$

confidence intervals for the true parameter values for patients that start medication at older age like in the over 40 group. This could be attributed to a stronger

negative correlation coefficient (of -0.4784) between the biomarkers for this age-group compared to the other groups. The other cause for the wide intervals could be the relatively high rate of dropout in this age-group (at 40%) as the size of bias is known to increase with the proportion of dropout (Lipsitz et al., 2009).

For each age-group we also considered inter-subject variability which measures the extent to which random-effects for individual patients influence their system parameter estimates. As displayed in Table 4.5, we noted that for the rate of CD4+ T cell production, the average variability was consistently small. That

Table 4.5: Inter-individual variability for the five parameters (s.e.).

Age-group	a	d	τ_{RTI}	τ_{PI}
$18-25$	0.76 (0.13)	1.22 (0.46)	1.53 (0.41)	1.48 (0.65)
$26-33$	0.69 (0.15)	0.98 (0.29)	1.97 (0.83)	2.37 (1.50)
$34-40$	0.71 (0.13)	1.41 (0.34)	1.37 (0.55)	1.63 (0.79)
Over 40	0.77 (0.15)	1.51 (0.32)	0.40 (0.58)	0.49 (0.37)

aside however, under a Gaussian assumption all these values were significant (p-value < 0.001). For τ_{PI} it was also observed that there was weak inter-individual variability for the $26-33$ and over 40 age-groups while the other two ($18-25$ and $34-40$) had p-values of 0.0114 and 0.0195 respectively. The results could, however, point to the nature of the data used in this analysis. Furthermore, as has been pointed out earlier, this could also result from mis-specified link functions for some of the system parameters.

4.6 Discussion

There will usually be situations in which nonignorable nonresponse is an important issue so that modelling the missing-data mechanism cannot be overlooked. Thus in this chapter we studied and illustrated the estimation procedure for parameters of the disease dynamic systems while taking into consideration the dropout mechanism using models for multivariate longitudinal data. Specifically, we have used the joint likelihood of the observed multiple outcome response variable, the random-effects, measurement error and the dropout mechanism in finding estimates of parameters of the HIV dynamical system described in equation 4.1

We also showed that the use of explanatory variables cannot be ignored in estimating the parameters and interpretation of the results. In the application of the methodology, we used observational datasets from two HIV clinics where along with the biomarkers (viral load measurements and CD4+ T cell counts), several explanatory values were also obtained. Inclusion of covariates in the estimation of parameters provides an insight into the characteristics of these parameters. For instance, we observed that gender of the patient influences the rate at which the actively infected CD4+ T cells die with a coefficient that is significant (< 0.001). This influence in the estimation is not unexpected because both the multiple response variable and the dropout mechanism are functions of these covariates as described in Section 4.4.

In the analysis, it was assumed that all subjects had their markers measured at the same time and dropout has been defined as all the response components not being observed once the subject withdraws from the study. However, in most longitudinal studies that involve collection of response variables that are characterised by multiple outcomes, it is common to have the response variable for a particular

subject to be partially observed so that only a subset of the k markers dropout on anyone occasion in the course of the study. This means that the remaining subset needs to be included in the models so that the information possessed by the elements thereof can contribute to the analysis. Moreover, there are other more challenging forms of missing data that are encountered in biomedical and other health related areas of research like informative intermittent missingness. Thus further studies are needed so that the estimation biases that arise as a result of missingness in these cases can be determined and remedial models be proposed. This is the topic of Chapter 5.

Moreover, some of the covariates needed to be redefined so that more practical elements are included in their definition. For instance, reporting to a clinic for supplementary treatment can be defined in such a way that more explicit information should describe the nature and severity of the illness that led to demand for supplementary treatment. Similarly, it would be useful to describe compliance to treatment as checking the patient if they are taking prescribed medication at appointed times and following any treatment related advice from the doctor or medical practitioner.

In conclusion, however, the methods and estimation procedures described here would also assist in finding effect-estimates of other grouping covariates such as treatment regimen and food supplementation on the parameters of the viral dynamical system. Moreover, the parameter estimates are within the range of those found in the literature. The estimate of the death rate of uninfected CD4+ T cells, for instance, agrees with those in other discussions ($C_N = 0.00013$ (ours) and $C_N = 0.00033,\ 0.0085$ in Guedj et al. (2011); Lavielle et al. (2011) respectively).

Chapter 5

A nonlinear mixed-effects model for multivariate longitudinal data with partially observed outcomes with application to HIV disease dynamics

Abstract

The several bio-markers may be measured in studying the response variable of interest in order to monitor and model disease progression. For instance, in studying HIV infection two markers (viral load levels and CD4 T cell counts) are used to monitor progression of the disease. In this chapter we consider a case where data are unbalanced among subjects and a situation where, for some reason, only a subset of the multiple outcomes of the response variable are observed at any one occasion for a particular subject. We propose a nonlinear mixed-effects model for the multivariate response variable data and derive a joint likelihood function that takes into account the partial dropout of the outcomes of the response variable.

We further show how the methodology can be used in the estimation of the parameters that characterise HIV disease dynamics. An approximation technique for estimating the parameters is also given and illustrated using routine observational HIV data.

5.1 Introduction

Longitudinal studies are frequently conducted to generate data for clinical trials in order to understand and model disease progression and the effect of treatment. In most of these studies several outcomes are measured or observed at each time point in order to study a response variable which cannot be measured directly. It is assumed that the biomarkers have a strong correlation with the unmeasured outcome. For instance, in the study of cardiac function (response), several outcomes are jointly measured or observed on each of the occasions spanning the study period so that a more complete picture of the response variable is presented (Lipsitz et al., 2009). Moreover, to have a better evaluation and interpretation regarding the process of interest the markers are jointly modelled instead of being analysed independently. However, multivariate longitudinal data have some complexities that pose challenges to the analysis process and these include correlation within a particular marker and also correlation among markers of the same subject and unbalancedness resulting from variation in measurement schedules for the subjects and missing scheduled visits.

In their study, Lipsitz et al. (2009) considered a response variable with four dichotomous markers and these were either measured or not measured on each occasion so that the whole set of outcomes was either observed at any time point or completely missing. They proposed a joint estimation of the marginal models for the binary

markers using a single modified generalised estimating equation where they also specified the pairwise association parameters among the different markers in a bid to get estimates of the main model.

Shah et al. (1997) considered a setting where the markers were observed on each occasion with a possibility of differences in the number of observations for the N subjects in the study. They used the expectation-maximisation (EM) algorithm of a linear mixed-effects model in order to estimate regression parameters under the multivariate response setting (Laird and Ware, 1982). Using a bivariate response variable they also considered the case where only one of the markers is observed on any one observation time.

A non-linear random-effects model for the analysis of multivariate longitudinal data with missing data was studied by Marshall et al. (2006). In their discussion they extended the non-linear random-effects model for a single response to the multiple responses situation with intermittent missing data. In finding estimates of the parameters, they used the first-order linearisation process of Lindstrom and Bates (1990) and then introduced a matrix of complete-case observations by deleting rows corresponding to missing data (Shah et al., 1997). Marshall et al. (2006) used the EM algorithm to estimate the parameters of the model. In this chapter we consider a nonlinear mixed-effects model for multivariate longitudinal data with dropouts such that the data are unbalanced among subjects and also within a subject because only a subset of the multiple outcomes of the response variable may be observed at any one occasion. We have modelled the partial dropout mechanism of the markers of the response variable through their occasions of dropout. We have estimated the parameters that characterise such a mechanism.

A joint likelihood function that incorporates the left-censored response values due to equipment failure to measure some outcomes accurately and this partial dropout

mechanism has also been proposed. We further show how the methodology can be used in the estimation of the parameters that characterise HIV disease dynamics in the presence of a multiplicity of covariates. The stochastic approximation expectation-maximisation (SAEM) algorithm for the estimation of parameters in nonlinear models is also described and illustrated using routine observational data from two HIV clinics in Malawi.

The rest of the chapter is organised as follows. In the next section we describe the disease dynamical model whose parameters we would like to estimate using observed biomarkers. The nonlinear mixed-effects model for multivariate longitudinal data is presented in Section 5.3 in which we also describe some of the general parameter estimation procedures. Our proposed methodology for handling partially observed multiple outcomes longitudinal data is given in Section 5.4. A joint likelihood function that incorporate both the dropout mechanism and the left-censored data is also derived in Section 5.5. Since maximum likelihood estimation can be computationally intensive especially with a large number of markers at each of the several occasions we use the SAEM algorithm in estimating system parameters which we describe in Section 5.6. The methodology is illustrated in Section 5.7 where we use routine observational HIV data to find estimates of parameters that characterise the disease dynamical model within the host. Section 5.8 is devoted to some discussions and conclusions.

5.2 The biological HIV disease model

This chapter aims at developing a methodology that could help in estimation of the parameters that characterise the latent HIV disease dynamical system of nonlinear ordinary differential equations in the presence of treatment. This 5-dimensional

system models the pathogenesis of the disease in order to assess the effectiveness of treatment and general disease progression (Wu, 2005). The dynamical system under consideration has been fully described in Section 3.2 of Chapter 3.

Ideally complex disease dynamical systems like this one require a rich and informative dataset about the process and properly specified statistical models when estimating the parameters. As we will see later in this chapter, the data used for illustration is characterised by unbalancedness both among and within the subjects.

5.3 The nonlinear model for multivariate longitudinal data

Let the multiple-outcome response matrix for subject i be denoted by

$$\mathbf{Y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{ik})$$

where each component $\mathbf{y}_{ih}$ $(h = 1, 2, \ldots, k)$ is an n_i-dimensional vector and k is the number of columns representing the markers which have been jointly observed or measured on each subject at the n_i occasions. Under this setting, we are implying that at the $(n_i + 1)$st occasion, the subject is not available for observation. We assume a model for the ith subject to take the form

$$\mathbf{y}_i = g(\boldsymbol{\psi}_i, \mathbf{X}_i) + \mathbf{e}_i, \;\; i = 1, 2, \ldots, N, \tag{5.1}$$

where $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i)$ and

$$g(\boldsymbol{\psi}_i, \mathbf{X}_i) = \{g_h(\boldsymbol{\psi}_i, \mathbf{X}_i) : h = 1, 2, \ldots, k\}$$

of which at least one is non-linear in $\boldsymbol{\psi}_i$ and the model covariates $\mathbf{X}_i$. The quantity $\mathbf{e}_i$ denotes the measurement error and $\boldsymbol{\psi}_i$ is usually called the parameter vector

possibly because it is a function of the fixed and random-effects. The nonlinear model and the parameter vector, π, are given in Section 3.3 of Chapter 3. The difference arises when we consider the conditional density of the unobserved sequence because in this chapter this sequence has more than one element.

5.4 The partial dropout model

This chapter aims at developing a methodology that can be used in estimation of the parameters that characterise the biological model presented in Chapter 3. Ideally complex disease dynamical systems like this one require a rich and informative dataset about the process and properly specified statistical models when estimating the parameters. As we will see later in this chapter, the data used for illustration is characterised by unbalancedness both among and within the subjects. Thus in this chapter, we propose models for this partial dropout process.

Consider a complete multivariate response variable $\mathbf{Y}_i$. Since we allow data to be partially observed for each occasion we can partition $\mathbf{Y}_i$ into three components: where the response variable is completely observed on all markers, $\mathbf{Y}_{i,o}$; partly observed $\mathbf{Y}_{i,p}$ where some markers are not observed and total dropout where the subject drops out, $\mathbf{Y}_{i,m}$. Thus the response matrix can be expressed as $\mathbf{Y}_i = (\mathbf{Y}_{i,o}, \mathbf{Y}_{i,p}, \mathbf{Y}_{i,m})$.

Suppose marker h is not observable after occasion n_{ih}, where $2 \leq n_{ih} \leq n_i$, so that for the n_i-dimensional 'complete vector', $\mathbf{Y}_{ih}$, the subvector $\mathbf{Y}_{ih,m}$ of the missing part for this marker has length $(n_i - n_{ih})$ where n_i is the occasion at which the last marker(s) for subject i is (are) observed. We wish to model occasion n_{ih} at which marker h ceases to be observed. It is assumed that after this occasion a total of

$k_i \leq k$ markers are no longer observable. It will be noted from this setting that $n_{ih} = n_i,$ for all h means the case where after the dropout occasion all the markers are no longer observed or measured. This case has been modelled in Chapter 4.

Let $\mathbf{H}_{ih} = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{n_{ih}-1})$ be the observed history for marker for the ith subject up to occasion n_{ih}. We may regard the dropout occasion for this marker, N_{ih}, as a categorical random variable taking n possible values. In this application we propose to use the approach of Chen et al. (2013) who modelled a categorical response variable in a different setting. Then the dropout occasion can be modelled using a multivariate logistic regression model of the form

$$\begin{aligned} f(N_{ih}|k_i, \boldsymbol{\lambda}, \mathbf{H}_{ih}) &= \Pr(N_{ih} = n_{ih}|k_i, \boldsymbol{\lambda}, \mathbf{H}_{ih}) \\ &= \begin{cases} \prod_{h=1}^{k_i} \frac{\exp(\lambda_{0n_{ih}} + \boldsymbol{\lambda}_{1n_{ih}}^T \mathbf{x}_i + \boldsymbol{\lambda}_2^T \mathbf{H}_{ih})}{1+\sum_{j=2}^{n} \exp(\lambda_{0j} + \boldsymbol{\lambda}_{1j}^T \mathbf{x}_i + \boldsymbol{\lambda}_{2j}^T \mathbf{H}_{ih})} & \text{for } 2 \leq n_{ih} < n+1, \\ \prod_{h=1}^{k_i} \frac{1}{1+\sum_{j=2}^{n} \exp(\boldsymbol{\lambda}_{0j} + \boldsymbol{\lambda}_{1j}^T \mathbf{x}_i + \boldsymbol{\lambda}_2^T \mathbf{H}_{ih})} & \text{for } n_{ih} = n+1 \end{cases} \end{aligned} \tag{5.2}$$

where j indexes the occasion and $\mathbf{x}_{ih}$ is the vector of covariates for the hth marker which is assumed to be time-invariant and is usually common to all markers. The quantity $\boldsymbol{\lambda}$ is a set of parameters that characterise the model. It is important to assume that each of the k markers has been observed at least once so that they contribute information to the analysis. Furthermore, the covariates are assumed to be completely observed so that the modelling does not have to account for missing covariates as well.

We now consider the distribution of the occasion when the subject completely withdraws from the study. This is equivalent to having the last marker(s) being observed. Given $N_{ih} = n_{ih}$, consider a random variable N_i that describes the occasion at which the remaining $s_i = k - k_i$ markers are observed or measured. This is the occasion at which the subject completely leaves the study and it is such that $n_{ih} \leq n_i \leq n$. This is like in the previous formulation, one would consider a

multivariate logistic regression model for n_i of the form

$$f(N_i|N_{ih},\boldsymbol{\zeta},\mathbf{H}_{ih}) = \Pr(N_i = n_i|N_{ih} = n_{ih},\boldsymbol{\zeta},\mathbf{H}_{ih}) \tag{5.3}$$
$$= \begin{cases} \prod_{h=1}^{s_i} \frac{\exp(\zeta_0+\boldsymbol{\zeta}_1^T\mathbf{x}_i+\boldsymbol{\zeta}_2^T\mathbf{H}_{n_i-1})}{1+\sum_{j=n_{ih}+1}^{n}\exp(\zeta_0+\boldsymbol{\zeta}_1\mathbf{x}_i+\boldsymbol{\zeta}_2^T\mathbf{H}_{ih,n_i-1})} & \text{for } n_{ih} < n_i < n+1 \\ \prod_{h=1}^{s_i} \frac{1}{1+\sum_{j=d_{ih}+1}^{n}\exp(\zeta_0+\boldsymbol{\zeta}_1\mathbf{x}_i+\boldsymbol{\zeta}_2^T\mathbf{H}_{n_{ih},j-1})} & \text{for } n_{ih} < n_i = n+1, \\ 1 & \text{for } n_{ih} = n_i = n+1 \end{cases}$$

where $\boldsymbol{\zeta}$ are the regression parameters of the model. The multivariate response variable comprises of the observed values, the partially observed markers and completely missing values as described at the beginning of this section.

Apart from the complications of partial dropout and serial correlations, the response variable may also be characterised by values that are below the limit of quantification resulting from the limitations of the detection of the assays used in quantifying the markers. This means that the values of some of the markers on that occasion are not known with certainty. For instance, in studies involving HIV disease modelling and monitoring viral load measurements have threshold values below which some readings may not be quantified.

Suppose $\mathbf{q}$ is the vector of lower limits of detection of the marker values $\mathbf{Y}_{ic}$ and let $\mathbf{Y}_{i(k-c)}$ be the set of markers whose measurements are do not have limits of quantification. In our setting we consider a distribution for the multivariate response variable that will accommodate the two types of markers. Using the approach of Guedj et al. (2007a), the multiple response variable is either observed or some components are less than their detection limits. If we let $\kappa = I_{\{\mathbf{Y}_{ic}>\mathbf{q}\}}$ be indicator of event $\{\mathbf{Y}_{ic} > \mathbf{q}\}$ then the density for the response variable for the ith subject is given by

$$f(\mathbf{y}_i|\mathbf{b}_i,\boldsymbol{\beta},\boldsymbol{\Sigma}_i,\mathbf{G}) = \text{constant}$$
$$\times \left\{|\boldsymbol{\Sigma}_{ic}|^{-1/2}\exp[-\frac{1}{2}(\mathbf{y}_{ic}-\boldsymbol{\mu}_{ic})^T\boldsymbol{\Sigma}_{ic}^{-1}(\mathbf{y}_{ic}-\boldsymbol{\mu}_{ic})]\right\}^{\kappa}$$

$$\times \left\{\Phi\left((\mathbf{q}-\boldsymbol{\mu}_{ic})^T\boldsymbol{\Sigma}_{ic}^{-1}(\mathbf{q}-\boldsymbol{\mu}_{ic})\right)\right\}^{1-\kappa} \tag{5.4}$$
$$\times|\boldsymbol{\Sigma}_{i(k-c)}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}_{i(k-c)}-\boldsymbol{\mu}_{i(k-c)})^T\boldsymbol{\Sigma}_{i(k-c)}^{-1}(\mathbf{y}_{i(k-c)}-\boldsymbol{\mu}_{i(k-c)})\right),$$

where $\boldsymbol{\mu}_{ic}$ and $\boldsymbol{\mu}_{i(k-c)}$ are subvectors of $\boldsymbol{\mu}_i = g(\boldsymbol{\psi}_i, \mathbf{X}_i|n_i, n_{ih})$ and in similar manner the quantities $\boldsymbol{\Sigma}_{ic}$ and $\boldsymbol{\Sigma}_{i(k-c)}$ represent submatrices of the covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}(\boldsymbol{\psi}_i)$ of the response variable given the vector $\boldsymbol{\psi}_i$. The quantity Φ denotes the distribution function of the multivariate normal random variable. The term $\left\{\Phi\left((\mathbf{q}-\boldsymbol{\mu}_{ic})^T\boldsymbol{\Sigma}_{ic}^{-1}(\mathbf{q}-\boldsymbol{\mu}_{ic})\right)\right\}^{1-\kappa}$ accounts for the marker values that are below the limit of quantification. In this formulation we have assumed that between them the vectors $\mathbf{y}_{ic}$ and $\mathbf{y}_{i(k-c)}$ account for the observed and missing values described at the beginning of this section. Thus the function given in equation (5.4) represents the likelihood of the response variable given the random-effects.

5.5 The likelihood function

Parameters of the nonlinear mixed-effects models can be estimated by using the maximum likelihood estimation procedure or its variant, the restricted maximum likelihood approach. The best such estimates are those that maximise the likelihood function of the observed data. In our application we require a function that incorporates the missing values of the response variable, the random-effects and the dropout occasions described in the models above. Let the complete sequence in this case be denoted by $\mathbf{D}_{\text{all}} = (\mathbf{y}_i, \mathbf{b}_i, n_i, n_{ih})$ with $\mathbf{y}_i = (\mathbf{y}_{io}, \mathbf{y}_{ip}, \mathbf{y}_{im})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{G}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \zeta)$ as the set of all parameters in the models described above. Then we have the complete-data likelihood function for the sequence $\mathbf{D}_{\text{all}}$ which has the form

$$L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}}) \;=\; \prod_{i=1}^{N}\left\{\prod_{h=1}^{k}[f(n_{ih}|\boldsymbol{\lambda}_i, k_i, \mathbf{H}_{n_{ih}})f(n_i|n_{hi}, \boldsymbol{\zeta})\right.$$

$$\times\ f(\mathbf{y}_i|n_{ih},\mathbf{b}_i,\boldsymbol{\beta}_i,\boldsymbol{\Sigma})]\, f(\mathbf{b}_i|\mathbf{G})\}\,, \tag{5.5}$$

where all the quantities have been defined and described above. To estimate the model parameters one would require the use of the likelihood function of the parameters given the observed sequence of measurements. This function is obtainable from equation (5.5) by integrating out the unobserved quantities, $\mathbf{w}_i = (\mathbf{y}_{i,p}, \mathbf{y}_{i,m}, \mathbf{b}_i)$, and is given by

$$L(\boldsymbol{\theta}|D_o) = \int_{\mathbf{w}_i} L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}})d\mathbf{w}_i.$$

The elements of $\boldsymbol{\theta}$ would then be found by maximising this function or equivalently by maximising

$$l_o(\boldsymbol{\theta}) = \log\left(\int_{\mathbf{w}_i} L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}})d\mathbf{w}_i\right).$$

But this is not possible under the circumstances because of the unknown quantity $\mathbf{w}_i$ and also because the nonlinear models may not have well-defined probability densities for the unobserved values. This suggests a need for a numerical approximation procedure to estimate the unknown model parameters. In this application, we use the SAEM algorithm which is described in the next section. The aim is to find elements of $\boldsymbol{\theta}$ that maximise the complete-data likelihood function presented in equation (5.5).

Let us denote $\log L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}})$ by $l_c(\boldsymbol{\theta})$. Then the approach under this algorithm is to replace $L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}})$ with the expectation of $l_c(\boldsymbol{\theta})$ with respect to the density of unobserved data at the current value of the parameter set. This quantity is expressed as

$$\begin{aligned} s(\boldsymbol{\theta}|\boldsymbol{\theta}^0) &= \mathrm{E}[l_c(\boldsymbol{\theta})|\mathbf{y}_{io}, n_{ih}, n_i, \boldsymbol{\theta}^0] \\ &= \int_{\mathbf{w}_i}\Bigg\{\sum_{i=1}^{N}\sum_{h=1}^{k}[\log f(n_{ih}|\boldsymbol{\lambda}_i, \mathbf{H}_{n_{ih}})] \\ &\quad + \log f(n_i|n_{hi}, \boldsymbol{\zeta}, \mathbf{H}_{n_i}) + \log f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &\quad + \log f(\mathbf{b}_i|\mathbf{G})]\Bigg\}\, p(\mathbf{w}_i|\mathbf{y}_{io}, \boldsymbol{\theta}^0)d\mathbf{w}_i, \end{aligned} \tag{5.6}$$

where $\boldsymbol{\theta}^0$ is the current value of the parameter set at which the expectation is evaluated and $p(\mathbf{w}_i|\mathbf{y}_{io}, \boldsymbol{\theta}^0)$ is the conditional density of $\mathbf{w}_i$ given the observed data and is written as

$$p(\mathbf{w}_i|\mathbf{y}_{io}) = \frac{L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}})}{\int_{\mathbf{w}_i} L(\boldsymbol{\theta}|\mathbf{D}_{\text{all}}) d\mathbf{w}_i}, \tag{5.7}$$

from model (5.5). This approximation approach is largely more efficient than the conventional EM algorithm in handling nonlinear models with missing data because it uses all of the simulated quantities through the iterations. This algorithm also converges to a maximum of the likelihood function of the observed sequence under a wide-range of assumptions and conditions in fewer iterations than the standard EM algorithm (Delyon et al., 1999; Meza et al., 2007).

5.6 Estimation using the SAEM Algorithm

The complete data for estimation of parameters of the hierarchical model given by (5.1) and (3.4) are $\mathbf{y}_i$, n_{ih}, n_i and $\mathbf{b}_i$. Since the quantities $\mathbf{y}_{i,m}, \mathbf{y}_{i,p}$ and $\mathbf{b}_i$ are unobserved we use the SAEM algorithm which we described in Chapter 3. The current setting, however, has additional elements constituting the unobserved sequence $\mathbf{w}_i$. Since we also wish to describe a procedure for the estimation of standard errors of the parameter estimates, we are going to present the algorithm again with appropriate modifications. It has the following steps:

Step 1: Let the iterations be indexed by $r = 0, 1, \ldots, \infty$ so that $\boldsymbol{\theta}^{(0)}$ is the initial set of values assigned to $\boldsymbol{\theta}$. Thus $\boldsymbol{\theta}^{(r)}$ denotes the values at the end of the rth iteration.

Step 2: Draw sequence $(\mathbf{w}_i^{(r+1)})$ of size $m(r)$ from the conditional density

$$p(.|\mathbf{y}_{io}, n_i, \boldsymbol{\theta}^{(r)})$$

given by equation (5.7), where now $\mathbf{w}_i = (\mathbf{y}_{i,m}, \mathbf{y}_{i,p}, \mathbf{b}_i)$.

Step 3 Update $s(\boldsymbol{\theta})$ using

$$\mathbf{s}(\boldsymbol{\theta}^{(\mathbf{r+1})}) = \mathbf{s}(\boldsymbol{\theta}^{(r)}) + \delta_{r+1}\left(\mathbf{S}(\mathbf{y}_{io}, \mathbf{w}_i^{(r+1)}; \boldsymbol{\theta}^{(r)}) - \mathbf{s}(\boldsymbol{\theta}^{(r)})\right),$$

where $s(\boldsymbol{\theta})$ is given by equation (5.6) and $\delta_{r+1}(> 0)$ is a fixed smoothing parameter that helps to accelerate convergence and is such that $\delta_{r+1} \to 0$ as the number of iterations increases. The term $\mathbf{S}(\mathbf{y}_i, \mathbf{w}_i^{(r+1)}; \boldsymbol{\theta}^{(r)})$ is given by

$$\mathbf{S}(\mathbf{y}_i, \mathbf{w}_i^{(r+1)}; \boldsymbol{\theta}^{(r)}) = \frac{\sum_1^{m(r)} l_c(\mathbf{y}_{io}, \mathbf{w}_i^{(r)}; \boldsymbol{\theta}^{(r)})}{m(r)}.$$

Step 4: Find $\hat{\boldsymbol{\theta}}^{(r+1)}$ that maximises $\mathbf{s}(\boldsymbol{\theta}^{(r+1)})$. When one uses REML to estimate the covariance matrices $\mathbf{G}$ and $\boldsymbol{\Sigma_i}$, the sufficient statistics $\sum_1^N \mathbf{b}_i\mathbf{b}_i^T$ and $\sum_{i=1}^N \mathbf{e}_i\mathbf{e}_i^T$ respectively are used, so that

$$\mathbf{G}^{(r+1)} = \frac{\sum_1^N (\mathrm{E}[\mathbf{b}_i\mathbf{b}_i^T|\mathbf{y}_{io}, \boldsymbol{\theta}^{(r)}])}{N - pk},$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{i}}^{(r+1)} = \frac{\sum_1^N \mathrm{E}[\mathbf{e}_i\mathbf{e}_i^T|\mathbf{y}_{io}, \boldsymbol{\theta}^{(r)}]}{Nn_i - pk},$$

where N, p and k denote the number of subjects, length of vector $\boldsymbol{\beta}$ and the number of markers jointly observed per subject respectively. Then we can find the updated value of $\boldsymbol{\beta}$ as follows (from 3.5)

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\sum_{i=1}^N \mathbf{A}_i^T\mathbf{W}_i^{(r+1)}\mathbf{A}_i\right)^{-1} \sum_{i=1}^N \mathbf{A}_i^T\mathbf{W}_i^{(r+1)}\boldsymbol{\psi}_i^{(r+1)},$$

where, for purposes of notation, we have set $\mathbf{W}_i = (\mathbf{B}_i\mathbf{G}\mathbf{B}_i^T)^{-1}$.

This procedure is repeated until

$$\mathbf{s}(\boldsymbol{\theta}^{(r+1)}) - \mathbf{s}(\boldsymbol{\theta}^{(r)}) < \boldsymbol{\eta},$$

for some convergence criterion $\boldsymbol{\eta} > \mathbf{0}$.

This algorithm does not provide standard errors of the parameter estimates at the end of the iterations. These quantities are found by using the observed information matrix through the matrix of second-order derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$. The evaluation of this matrix is complex because it does not have a closed-form expression. Delyon et al. (1999) proposed a stochastic approximation of the Fisher information matrix as follows. This has bee described in Section 3.4, Chapter 3.

5.7 Illustration with bivariate HIV data

5.7.1 Data description and implementation

In this illustration we have included all subjects with a minimum of three bivariate measurements of CD4+ T cell count and viral load and the maximum number of eight occasions. The need for complete first three occasions was necessitated by the type of data which were collected through routine observation of patients and this choice would provide an averaging effect for the baseline measurements as there would be enough information on the markers. Furthermore, the first measurements were taken several weeks after commencement of treatment.

Two hundred and fourteen subjects had this information and between them they shared a total of 2024 clinical measurements of the markers. These measurements suggest an average missing rate of 41%. In the current chapter, however, we allowed the unbalancedness to occur both among subjects and also among the markers within subjects. That is, the number of markers was allowed to be different for each occasion and between patients. The number of viral load measurements was in general less than that of the CD4+ T cell count. Table 5.1 represents the number

of measurements for each occasion and marker and it can be observed that there is a remarkable imbalance in the number of markers after the third occasion (28 weeks). About 39% of viral load measures were below the limit of quantification of

Table 5.1: Number of markers by occasion.

Week	0	14	28	42	56	70	84	98
CD4+ T cell count	214	214	214	183	141	70	37	23
Viral load	214	214	214	177	78	24	7	0

400 copies per ml. Thus in general, our dataset was characterised by high rate of unbalancedness within the subject as well as among subjects and a large proportion of left censored values. The average inter-visit interval was 14 weeks.

5.7.2 Results

We estimated the parameters of the HIV disease dynamical system of nonlinear ordinary differential equations presented in Section 3.2 (equation (3.1)). For reasons described in the previous subsection the data used in this analysis could not support estimability of all the parameters associated with the latent disease process. This resulted to having three of the parameters being set as fixed. These were (i) infection rate of T_N cells per virus particle ($\gamma = 0.0021$, from Lavielle et al. (2011)), (ii) death rate of latently infected CD4+ T cells per day ($c_L = 0.0092$, from Guedj et al. (2007b)) and (iii) death rate of the virus particles per day ($c_V = 30$, from Ribeiro et al. (2002)).

We first needed to justify the inclusion of the covariates in the analysis. Goodness-of-fit tests were conducted on the data in `MONOLIX` in order to choose the significant

Table 5.2: Test for model covariates: A = age-group; C = compliance; F = facility; S = sex; T = treatment.

Covariate set	CFS-A	TFS-A	TCS-A	TCF-A
$p-$value (LRT)	0.2423	0.041	< 0.001	< 0.001
AIC	18551.6	18569.9	18587.86	18535.25

covariate set for inclusion in the current analysis. The tests were done in such a manner that a model with a set of three or four covariates was compared with the one without covariates. There were ten such covariate combinations and the results displayed in Table 5.2 only include sets that were significant and also included four covariates. We note from these results that TCF-A has the smallest AIC (Akaike information criterion) at 18535.25. Thus in our analysis we included supplementary, compliance, facility and age-group as covariates. It is noted that with our modelling approach despite the unbalanced nature of the dataset due to missing values we were able to get standard errors for most of the covariate coefficients directly from the output. This appealing result could be attributed to the use of all available information on the patients even when only one marker was observed for a given occasion. We used individual parameter estimates to obtain standard errors of those coefficients where such values could not be obtained directly. These results are presented in Table 5.3.

Intercepts were significant as can be seen from the second and third columns of this table. For instance, for the death rate of the actively infected cell, (c_A), we have a negative intercept. We may suggest that in interpreting this, the value of c_A for a noncomplying subject who visits Lighthouse clinic and does not get supplementary treatment is about 0.1863. Similar interpretations can be said about the other

intercepts. Except for the insignificant coefficients of the activation rate of the

Table 5.3: Estimates of the covariate coefficients (and their p-values).

	Intercept	Supplementary	Compliance	Facility
a	2.28	0.598	0.293	0.36
	(< 0.001)	(< 0.001)	(0.022)	(0.0069)
c_N	-5.02	1.21	0.525	-0.107
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
d	-3.06	0.348	1.18	-0.837
	(< 0.001)	(0.46)	(0.046)	(0.14)
c_A	-1.68	0.44	1.13	1.15
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
p	3.36	0.87	0.716	-1.45
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
π	-3.97	-1.3	0.749	3.58
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
τ_{PI}	4.49	-2.38	-9.48	-10.1
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
τ_{RTI}	-1.78	3.67	0.189	5.63
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)

latently infected cells (d), the results indicate that the covariates have a high influence on the dynamics of the disease and this provides a basis for the inclusion of these covariates when estimating parameters of the HIV disease dynamics. This becomes even more crucial when data are highly unbalanced like ours because such covariates provide additional information needed for the analysis. Thus it

would be contended that, for example, compliance to treatment guarantees that HIV therapy is effective and also reduces the risk of developing drug resistance and that the facility that offers user-friendly services contributes to the patients general response to medication (van der Eijk et al., 2005). It can also be concluded on the basis of these results that the covariates do not have direct bearing on d largely because, according to Marini et al. (2008), latency of the infected CD4+ T cells is known to persist even during HIV treatment.

For these data we also noted that there was a strong negative correlation between the τ_{RTI} and τ_{PI} with a value of -0.574. We could not immediately have backing from the literature but this could be apparently linked to the association between the HIV disease markers.

Based on the age-groups with $18-25$ as the reference group we also computed the estimates of the eight parameters and their central 95% confidence intervals as shown in Table 5.4. The objective was to check the degree of influence of age at commencement of treatment on the various coefficients associated with the explanatory variables.

For some parameters, it was also observed that the value of the estimate was associated with age at commencement of treatment. For instance from Table 5.4 it can be noted that values of a, c_N and p vary over age-groups as can be seen from non-overlapping pattern in confidence intervals between any two groups or among all the groups. These results were also supported by ANOVA tests which showed there was significant difference in parameter values between the age-groups (we used individual estimates of these parameters).

On average there were wider confidence intervals for the true parameter values for the $34-40$ age-group in this analysis. This could be attributed to the relatively

Table 5.4: Parameter estimates for the within age-groups (and their confidence limits).

	Age-group		
Parameter	$26-33$	$34-40$	>40
a	16.03	21.19	49.12
	$(13.9, 18.16)$	$(17.84, 24.54)$	$(40.15, 58.1)$
c_N	0.0048	0.0107	0.0341
	$(0.0041, 0.0055)$	$(0.0091, 0.012)$	$(0.0284, 0.0399)$
d	0.0604	0.0614	0.6975
	$(0.0417, 0.0792)$	$(0.0345, 0.0884)$	$(0.4627, 0.9123)$
c_A	0.4479	0.8287	3.093
	$(0.3768, 0.519)$	$(0.6881, 0.9692)$	$(2.534, 3.6508)$
p	56.66	101.11	167.71
	$(49.41, 63.9)$	$(88.38, 113.84)$	$(142.64, 192.77)$
π	0.2278	0.2818	0.2177
	$(0.1813, 0.2743)$	$(0.2244, 0.3388)$	$(0.1741, 0.2613)$
τ_{PI}	0.5118	0.7895	0.3106
	$(0.4368, 0.5867)$	$(0.6868, 0.8922)$	$(0.2405, 0.3806)$
τ_{RTI}	0.44	0.1452	0.5294
	$(0.3490.5368)$	$(0.1061, 0.1842)$	$(0.4326, 0.6262)$

high rate of dropout seen in this age-group for both markers (with 42% for the CD4+ T cell count and 48% for the viral load) as the size of bias is usually known to increase with the proportion of dropout (Lipsitz et al., 2009).

For the results of the five parameters that were estimated for both the dropout model (see Table 4.4) and partial dropout model (Table 5.4 above) we noted several interesting results. Firstly, in both cases we included four covariates of which three were common and these are (i) age-group, (ii) supplementary treatment and (iii) compliance to treatment. The only difference was that the dropout model had sex of a patient as the fourth covariate while in the partial dropout case had facility. However, from results of exploratory data analyses in Chapter 2 indicate that the patient's sex does not determine the choice of the facility to which they presented themselves. Thus we are in a better position to assume that the differences in the estimates could be entirely due to the difference in the number of marker values in the two scenarios with 1856 values in the dropout case (Chapter 4) and 2024 values in the current chapter.

Secondly, the results show that in the model with partial dropout (PDO), the estimates from the age-groups have wider ranges than the estimates in the case of ordinary dropout (DO) model. The scenario is illustrated in a box plot in Figure 5.1. This has also been confirmed by tests on differences between the means in the three age-groups for the partial dropout scenario which showed there was significant difference ($p < 0.001$) as discussed above. The large ranges in the partial dropout case can be attributed to the increased variability in the estimates because of the unbalancedness in marker readings within and among the subjects under study, especially across age-groups. For instance in the present case, there were 2014 laboratory measurements of which only 46% were viral load readings and the rest were CD4+ T cell counts. On the other hand, there was an equal number of the

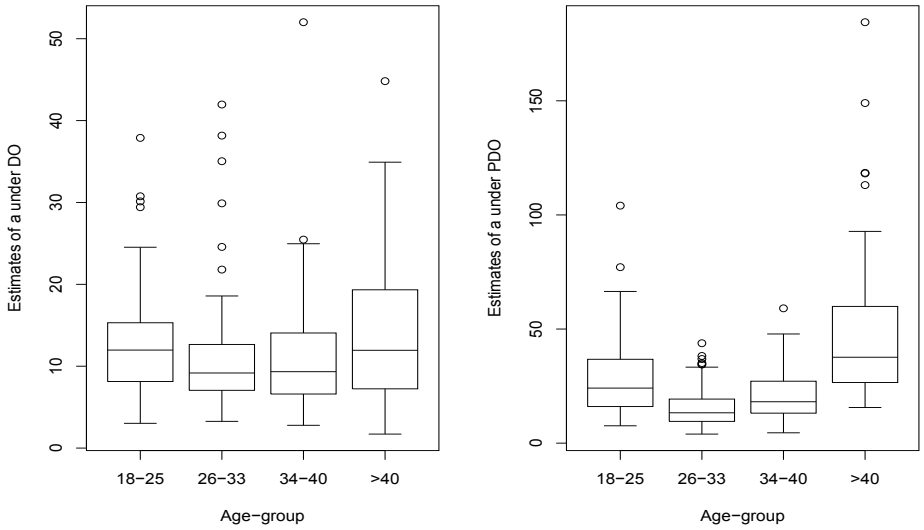


Figure 5.1: Box plot for individual estimates of a under PDO and DO.

markers in the dropout models considered in Chapter 4. The other contributing factor for the differences could be the dependence of these covariates (gender and facility) on the other covariates (see tests in Section 2.2, Chapter 2). Thirdly, the confidence intervals for the parameters estimated from the dropout model were general much wider than in the partial dropout case. This was true in 80% of the comparisons and the remaining values were all connected to c_N, the death rate of the uninfected CD4+ T cells. As already pointed out, this is not surprising because for the same number of covariates, the model developed in this chapter accounts for all the laboratory measurements by accommodating them through the partial dropout mechanism.

In earlier studies, there were varied reasons for studying the HIV disease dynamical system of the nonlinear ordinal differential equations. For instance, in Lavielle et al. (2011) one of the objectives of the discussions was to compare the models under different treatment regimen. Thus without prior discussions related to our

findings, we can only argue that our models present a mechanism for improved inference of the parameters of the HIV disease biological system.

We also computed the inter-subject variability (the results have not been reported) which measures the extent to which random-effects influence the system parameters. Under the normal distribution assumption all the estimates were significantly different from zero ($p < 0.001$).

Parameters characterising the dropout occasions in the models (5.2) and (5.3) were also estimated. Like in the response model (5.1), we included the three covariates (facility, supplementary treatment and compliance to treatment). The results for both scenarios are reported in Table 5.5. The results of the former model indicate that the dropout process could not be ignored in the estimation of the system parameters. It was, however, revealed that the history of the CD4+ T cell did not contribute significantly as did the log10-viral load history. This could result from the fact that the viral load is the marker that was not mostly observed in later occasions for most patients.

We also noted that the dropout process (in model (5.2)) was highly influenced by the covariates, particularly compliance to treatment and facility to which patients presented themselves. These results are in agreement with what was observed in the exploratory cross-tab analyses of the current data where it was noted that for one facility there was high proportion of dropout and small at the other facility. Furthermore, one would argue that a compliant patient would not need further viral load monitoring as they would be deemed responding well to treatment regimen. Little influence on the model process was detected from the supplementary treatment and the negative coefficient is not unexpected since a patient who seeks extra clinical care would not be expected to dropout because they needed close monitoring of the response variable through observation of the biomarkers.

Table 5.5: Estimates for dropout coefficients

	Model (5.2)		Model (5.3)	
Coefficient	λ	p-value	ζ	p-value
Intercept	-0.876	< 0.001	-1.822	< 0.001
Supp. treatment	-0.022	0.736	-0.222	0.035
Compliance	0.173	0.037	0.502	< 0.001
Facility	0.294	< 0.001	0.911	< 0.001
${y_{vl,n_{ih}-1}}^{*}$	0.112	0.001	0.083	0.118
${y_{cd,n_{ih}-1}}^{\dagger}$	7.65×10^{-5}	0.122	$-$	$-$
${y_{cd,n_i-1}}^{\ddagger}$	$-$	$-$	2.21×10^{-4}	0.007

*log10-viral load just before n_{ih}; †CD4+ T cell count just before n_{ih}; ‡CD4+ T cell count just before n_i

There was evidence of nonignorable dropout in the latter model as well (model (5.3)). This dropout model was largely dominated by CD4+ T cell count which had a coefficient that was significant with a p-value of 0.007 and the viral load history was not significant. This was in order since very few values (of the log10-viral load) were observed at the final observation occasion for this model. The results also show that all the covariates generally influenced the dropout mechanism in some way. This could be attributed to the fact that this was what could be termed as *terminal* dropout where all the covariates contribute directly or indirectly to the dropout mechanism. We noted, for instance, that the subject specific covariate compliance to treatment seemed to increase the tendency for a subject to withdraw.

5.8 Discussion

In this chapter, we proposed a joint likelihood function of the multivariate response variable and the dropout occasions in order to facilitate the estimation of the parameters of the latent HIV disease dynamical system of nonlinear ordinary differential equations and those of the dropout mechanisms. We used the SAEM algorithm as an estimation technique for estimating these parameters since no analytic estimation formulas are obtainable for these systems.

This estimation technique was applied to routine observational data from HIV clinics under the Lighthouse Trust at Kamuzu Central Hospital in Malawi. The data were characterised by high proportions of missing observations due to dropout and left-censored viral load measures. In this chapter, we allowed the unbalancedness to occur both among subjects and also among the markers within subjects so that the number of markers was allowed to be different for each occasion and for all the patients. The number of viral load measures was in general less than that of the CD4+ T cell count. Our proposed methodology included this unbalancedness so that all the observed markers contributed to the estimation process. We have compared the results of the illustration of this methodology with the case of dropout mechanism discussed in Chapter 4. It has been established that under that dropout case the confidence intervals of the system parameters are much wider than in the current case. This could point to reduced uncertainty emanating from use of more information in this (partial dropout) scenario.

Unlike in similar studies where only one or no covariate was considered, in this analysis we included several such covariates. Results on coefficients of these covariates indicated that leaving them out when making inferences may result in bias and general loss of valuable information because almost all of them were significant

to the models in which they were used.

We also showed that the dropout occasions play a significant role in the analysis because if included as done here estimation bias is substantially reduced. It has been observed from the results in Table 5.5 that the partial dropout mechanism represented in models (5.2) and (5.3) is dependent on biomarker history and the covariates as manifested by the significant coefficients for the two quantities. The analysis also revealed that the history of the CD4+ T cell count was not significant for model (5.2) and the log10-viral load measurement history was also not significant for model (5.3) of the dropout mechanism.

In the analysis, it was assumed that all subjects had their markers measured at the same time. However, in most longitudinal studies that involve collection of multivariate data it is common to have the outcomes to be measured or observed at different time points. In addition, some of the covariates required to be redefined so that more appropriate features are taken care of. As an example, we could define supplementary treatment in such a way that the nature and severity of the illness that led to demand for supplementary treatment are described. Such definitions would also help in better interpretation of some results.

In the next chapter, we study the use of multiple imputation of the missing marker values on estimation of the parameters of the HIV disease dynamical model. The current results, however, are comparable with those in the literature. For example, the activation rate of latently infected CD4+ T cells d in Guedj et al. (2007b) is 0.042 while our analyses this chapter give an average value of 0.039 on the same unit of measure.

Chapter 6

Multiple imputation and simulation of multivariate longitudinal data with application to HIV dynamical systems

Abstract

In studying the HIV disease pathogenesis it is important to look at characteristics of estimates of the system parameters under incompleteness of the data due to the limit of quantification which results from the methods used to make the measurements. In this chapter we compare the level of variability of the parameter estimates for different simulated partially observed data using different limits of quantification. Secondly, to deal with the missing data problem, we assume the data are at least missing at random so that we use multiple imputation in order to have complete data from the partially observed data which is then used in estimating system parameters. We compare the complete case analysis results with those obtained under the partial dropout scenario.

6.1 Introduction

Mathematical models that describe within host HIV disease dynamics are important in the study of characteristics of HIV infection and progression in the presence of clinical interventions in the form of antiviral treatment. There are many such models that have been proposed in the literature and the problem usually reduces to estimation of parameters that characterise these mathematical models (Huang et al., 2006; Guedj et al., 2011; Perelson and Nelson, 1999; Nowak and May, 2000). To carry out these estimates, bivariate longitudinal data of viral dynamic biomarkers are usually modelled using nonlinear mixed-effects models(Guedj et al., 2007a).

Typical longitudinal data from studies such as clinical trials or prospective observational studies consist of measurements or observations taken repeatedly over time on a particular subject. In such studies missing data occur as a result of the subjects missing scheduled visits or some variables could not be measured or observed at particular visits or simply because subjects drop out of the study. The objective of such studies is to assess changes and trends in the response variables of study subjects. In the presence of missing data, this objective may not be properly achieved.

There are many procedures for handling incomplete datasets and usually these depend on the assumptions placed on the response models and the mechanisms governing the missing data. For instance, one needs to make assumptions about whether the missing data are ignorable or nonignorable and bearing in mind that participating subjects may offer data that is different in important ways from those subjects that have missed some occasions or completely dropped out (Brancato et al., 1997). Approaches can be categorised in line with how the missing values will be treated: either by reweighting of observations or imputation of the

missing values (Liu et al., 2000). The popular choice is the imputation of missing data which involves filling in missing data with values obtained from some specified model of the observed values (Singh, 2009). In most applications, multiple imputation is used because it has advantages of accommodating multiple levels of missing data and also being easily implemented in most statistical software (Kmetic et al., 2002; Schafer and Yucel, 2002). Another important advantage of multiple imputation is that it accounts for uncertainty in the imputed values apart from offering variability in the estimates of the parameters.

One of the objectives of the study was to check the sensitivity of parameters for various levels of limits of quantification. This objective is addressed in this chapter, where we carry out a simulation study with three limits of quantification of the viral load levels and look at the effects on the estimates of the parameters that govern HIV disease dynamics. Furthermore, we consider multiple imputation of the missing data. In particular we look at a case where disease biomarkers are not fully jointly observed (see Chapter 5, Section 5.4). This scenario introduces two practical problems. The first problem is the missing values of the markers and the second problem is the bivariate response variable with one marker having some values that are below the limit of quantification. We compare the results with those found directly by using partially observed data.

The rest of the chapter is organized as follows. Section 6.2 gives a motivation and statistical models of the HIV disease dynamical system. A simulation study is presented in Section 6.3 where we compare parameter estimates for various levels of limits of quantification based on data with partial dropout. Results from analyses based on imputed data and comparisons with results based on partially observed data are discussed in Section 6.4. We conclude the chapter with a discussion in Section 6.5.

6.2 Motivation and Statistical models

There are many biological models that describe the interaction of HIV virus with the immune system while the patient is on a treatment course (Perelson and Nelson, 1999; Nowak and May, 2000; Huang et al., 2006). The discussion of this chapter considers the disease dynamical model that has been described in Section 3.1. In particular, we would like to compare estimates of system parameters obtained under partial dropout and those obtained when multiple imputation of the missing markers is conducted. Furthermore, we would like to assess the sensitivity of parameter estimates to changes in the limit of quantification. The statistical model we consider is a bivariate response variable of the form

$$\mathbf{y}_i = (g_1(V(t_{ij}, \boldsymbol{\psi_i})), g_2(T(t_{ij}, \boldsymbol{\psi}_i)))^T + (e_{i1}, e_{i2})^T, \tag{6.1}$$

which is also described in Chapter 3 (see equation (3.12)). The quantities V and T are the measured viral loads and CD4+ cell counts respectively. It is usually assumed that the parameters ($\boldsymbol{\psi}_i$) of the system in model (6.1) are transformations of a Gaussian random vector. Thus the transforms are modelled as in equation (3.13) in Chapter 3.

In the procedures described in the next sections we use the SAEM algorithm which has been described in Section 5.6 of Chapter 5. The implementation is done in `MONOLIX` and `R`.

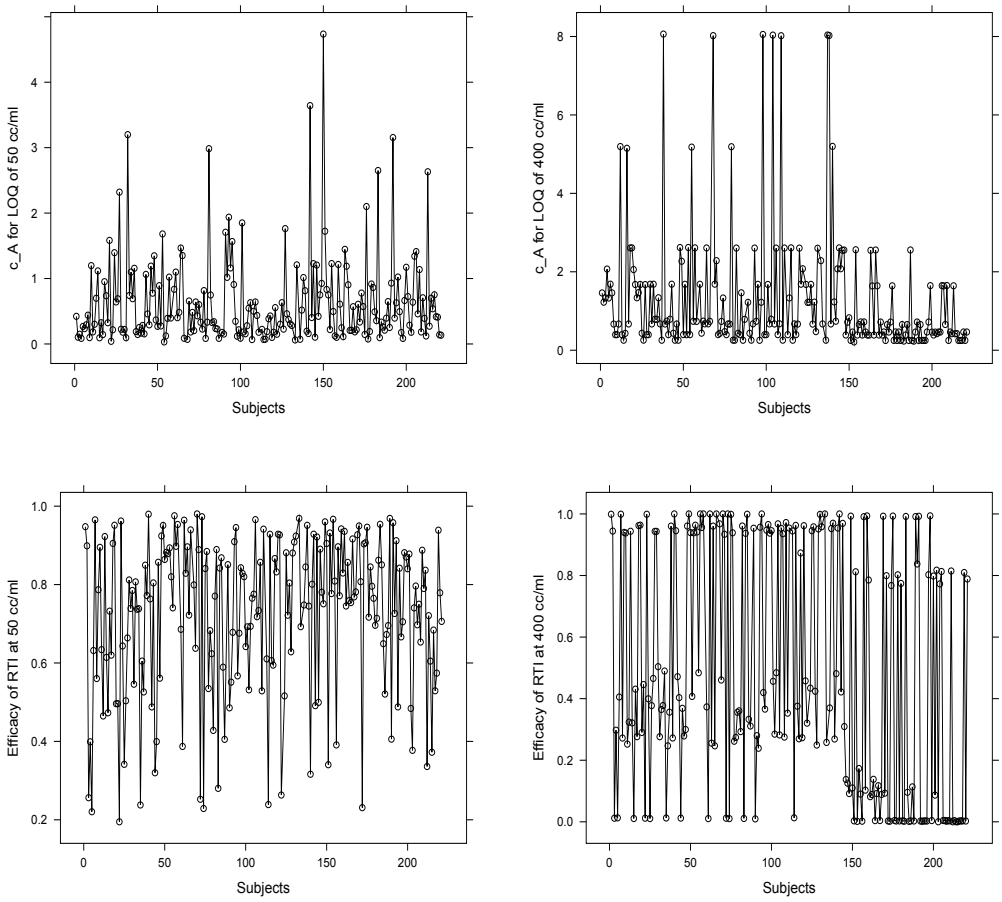
6.3 Parameter estimates under different limits of quantification

In this section, we considered the effect of limits of quantification (LOQ) of the viral load measurements on the estimates of the HIV dynamical system parameters through a simulation study. Estimates and their confidence intervals were compared at three levels of LOQ: 400, 200 and 50 copies per ml. The data with LOQs 200 and 50 copies per ml were simulated based on the partially observed data described in Section 5.4 of Chapter 5. Estimates of the system parameters and their 95% confidence intervals are displayed in Table 6.1 for the different limits of quantification. It will be noted from this table that, in general, the width of the confidence interval is smallest in the data with a LOQ of 50 copies per ml. This is to be expected because lower limits of quantification correspond to increased accuracy of measurement which results in a smaller proportion of unobserved values (due to left-censored values). For instance, the three datasets with LOQs of 400, 200 and 50 copies per ml, had respective proportions of 39%, 25% and 20% of values below their LOQ. It should also be noted that the level-off effect of the data decreases with the size of LOQ.

Figure 6.1 presents plots of individual estimates for the death rate of the activated infected CD4+ T cells (c_A) and the efficacy of the reverse transcriptase inhibitors (τ_{TRI}) using simulated data based on a limit of quantification of 50 copies per ml and those found using the original data with an LOQ of 400 copies per ml. It can be noted from the plots that the individual estimates of the parameters for the simulated data have smaller inter-individual variations for both parameters. This unevenness can be attributed to the reduced level-off effect in the case of simulated data as a result of a smaller LOQ.

Table 6.1: Parameter estimates for the different levels of LOQ (and their confidence limits)

	Limit of quantification		
Parameter	400 copies per ml	200 copies per ml	50 copies per ml
a	28.39	31.23	30.71
	$(25.11, 31.68)$	$(27.84, 34.62)$	$(28.057, 33.36)$
c_N	0.0183	0.028	0.0228
	$(0.016, 0.021)$	$(0.021, 0.0337)$	$(0.019, 0.026)$
d	0.227	0.231	0.131
	$(0.158, 0.297)$	$(0.203, 0.258)$	$(0.116, 0.146)$
c_A	1.29	0.745	0.637
	$(1.08, 1.502)$	$(0.627, 0.863)$	$(0.546, 0.728)$
p	129.64	83.90	91.33
	$(116.24, 143.03)$	$(74.5, 93.31)$	$(80.64, 102.03)$
π	0.207	0.172	0.206
	$(0.182, 0.233)$	$(0.15, 0.194)$	$(0.181, 0.231)$
τ_{PI}	0.488	0.7895	0.653
	$(0.44, 0.536)$	$(0.705, 0.753)$	$(0.626, 0.679)$
τ_{RTI}	0.49	0.645	0.726
	$(0.437, 0.543)$	$(0.62, 0.67)$	$(0.699, 0.753)$



50 copies/ml 400 copies/ml

Figure 6.1: Estimates at LOQ of 50 copies per ml and 400 copies per ml.

The plots reveal that there seems to be sub-populations in the individual estimates. These are evident in the panels of the second column where we see that for the first segment covers the first 138 individuals and the other segment includes the remaining 76 individuals. These numbers correspond to the two facilities from which data were collected. These results could have come about because of the differences in proportions of viral load levels that were below LOQ. For instance, exploratory analyses show that for Lighthouse clinic, only 33% of viral load readings were below LOQ while Bwaila had 42% such values. But this was not the case with simulated data where the proportion of measurements below LOQ was bigger for Lighthouse (25%) clinic than that of Bwaila clinic (17%). The reasons for this phenomenon are not immediately obvious because one would expect that the proportions could have been the same under simulation. The other interesting result is that the individual estimates for the efficacy of reverse transcriptase in original data indicate that the values were mostly on extreme ends, with almost no values between 0.5 and 0.8. There was no immediate explanation for this outcome.

It was further observed that regression coefficients for the covariates in the dataset with an LOQ of 400 copies per ml were significant. For instance, for c_A, death rate of the infected activated CD4+ cells, all the covariates and the intercept were significant when a dataset with an LOQ of 400 copies per ml is used in the analysis. It could, therefore, be argued that in the presence of a higher proportion of values which are below the LOQ, covariates provide important information about subjects that is relevant in the estimation of model parameters.

We also considered inter-subject variability for each dataset. This is a measure of the extent to which subject-to-subject effects influence estimates of parameters that characterise the within-host HIV disease dynamics. The results are displayed in Table 6.2. The variability for all parameters in the two datasets with small

Table 6.2: Inter-individual variability for four system parameters (s.e.) in the three datasets

LOQ	a	d	τ_{RTI}	τ_{PI}
400 copies/ml	0.564 (0.044)	1.21 (0.24)	0.715 (0.31)	1.77 (0.68)
200 copies/ml	0.437 (0.023)	0.697 (0.034)	0.721 (0.037)	0.497 (0.03)
50 copies/ ml	0.45 (0.023)	0.731 (0.036)	0.892 (0.048)	0.564 (0.037)

LOQs (200 and 50 copies per ml) were significant which is expected because for the same number of measurements there is little level-off effect in these sets that comes with data truncation resulting from lower quantification levels.

6.4 Multiple imputation for partially observed multivariate data

Missing data are inevitable in epidemiological and clinical studies and have the potential to compromise the validity of results due to bias. The data used in the current application were observational in nature and were characterised by high proportions of dropout and left-censored values. This information was obtained from two hundred and fourteen subjects and among them they shared 2024 laboratory measurements of the markers. These data had an average dropout rate of 41%. The response variable was such that it was unbalanced both among subjects and also within subjects because on account of resource constraints only one marker could be observed more often. The number of viral load measurements was in general less than that of the CD4+ cell counts (see Table 5.1 in Chapter 5).

6.4.1 A brief description of multiple imputation

We assumed the data were missing at random so that we could use multiple imputation in order to fill in the missing values (Rubin, 2004). Under this missing data method, the missing values of the response variable (or covariate as need be) are predicted by using the observed values (Sterne et al., 2009). Each missing value is replaced by a set of values that are generated using the observed values so that several complete datasets are created depending on the need and computational capacity. This process gives the name multiple imputation.

In order to create imputed values, a set of similar regression models are identified which will allow one to create imputes based on other variables in the dataset. These regression lines present different versions of what the actual equation for the missing data will take. Production of several versions of the data allows one to average over these sets so that better estimates are produced. The number of imputed datasets to create is usually between 3 and 10 data sets. These regression lines will need predictor variables to help preserve relationships in the data. The observed data are used to generate the imputed values because they are assumed to be correlated with these missing variables and the missingness mechanism (Stuart et al., 2009).

Then estimation of the parameters is carried out using standard statistical analyses on each imputed complete dataset, giving multiple estimates for each parameter. These multiple estimates are then combined to provide a grand estimate. Multiple imputation takes into consideration the natural variability in the missing data and also incorporates the uncertainty caused by estimating missing data (Allison, 2000; Lloyd et al., 2013).

Multiple imputation has several advantages which include taking care of multiple

levels of missing data and accounting for uncertainty in the imputed values in addition to creating variability in the estimates of the parameters. This is done by creating imputed values which are based on variables correlated with the missing data and causes of missingness and by creating different versions of the missing data. As a result, it is known to produce estimates that are unbiased and satisfactory even with small sample sizes or with sets of data that have high rates of missingness. Moreover, this missing data method is robust enough to departures form normality assumptions. Additionally, the complete datasets from multiple imputation can be analysed using most known methods and software packages.

Multiple imputation is a preferred solution to missing data problems because it provides quality of results and it is easy to use. It can be performed on a variety of missing data situations and has known to produce unbiased parameter estimates which reflect the uncertainty associated with estimating missing data (Schafer, 1997; Lloyd et al., 2013).

6.4.2 Results from multiple imputation

In this setting, we used multiple imputation on partially observed data (described in Chapter 5). This was implemented in `R` using the package `mi`, Version 0.09-18.03 (Gelman et al., 2013). We generated six complete datasets which were analysed using complete-case methods discussed in Chapter 3. Part of code for multiple imputation of data is given in Appendix A. Then each of the six datasets was used in producing a set of estimates of the parameters that characterise HIV disease dynamic systems (see model (3.1)). The estimation of the parameters was conducted in `MONOLIX`. These parameter estimates and their standard errors were then compared with those obtained by using data with partial dropout. We generated

only six sets because the analysis of any given dataset takes a long time to run in MONOLIX and also this number of imputed datasets is reasonable from practical point of view (Stuart et al., 2009).

The average proportion of values below limit of quantification for the six complete datasets was (38.9%) the same as the proportion of such values (38.8%) in the original data (with partially observed values). Table 6.3 shows estimates and the standard errors of the covariate coefficients for four parameters using imputed data (MI) and those based on data with partial dropout (PDO). It can be seen from this

Table 6.3: Estimates of the covariate coefficients (and their p-values)

Param.		Intercept	Supp.	Compl.	Facility
a	P†	2.28 (< 0.001)	0.60 (< 0.001)	0.29 (0.022)	0.36 (0.007)
	M‡	2.2 (< 0.001)	0.10 (0.0012)	0.003 (0.936)	0.33 (< 0.001)
c_N	P	-5.02 (< 0.001)	1.21 (< 0.001)	0.525 (< 0.001)	-0.11 (< 0.001)
	M	-5.25 (< 0.001)	0.72 (< 0.001)	0.14 (0.41)	-0.38 (0.022)
τ_P	P	4.49 (< 0.001)	-2.38 (< 0.001)	-9.48 (< 0.001)	-10.1 (< 0.001)
	M	4.23 (< 0.001)	-0.57 (< 0.001)	-1.58 (< 0.001)	-0.76 (< 0.001)
τ_T	P	-1.78 (< 0.001)	3.67 (< 0.001)	0.189 (< 0.001)	5.63 (< 0.001)
	M	1.06 (< 0.001)	0.19 (0.017)	0.11 (0.349)	0.84 (< 0.001)

†Partial dropout; ‡Multiple imputation; $\tau_P = \tau_{PI}$; $\tau_T = \tau_{RTI}$

table that the covariate coefficients are smaller in the imputed scenario than in the partial dropout case for most parameters. As an example, we note that compliance to treatment is largely not a significant covariate in the analysis when imputed data are used. This is not unexpected because when multiple imputation is carried out, the resulting data (with 3424 clinical measures) seem to provide adequate

information about the response variable which could result in some covariates being non-significant. We could not immediately explain the difference in signs of the intercepts for the parameter τ_{RTI}.

We also noted that the HIV dynamical system parameters were estimated with smaller standard errors under multiple imputation as compared to the case with partially observed values. These improved results have possibly come about as

Table 6.4: Overall parameter estimates using multiple imputation and partial dropout

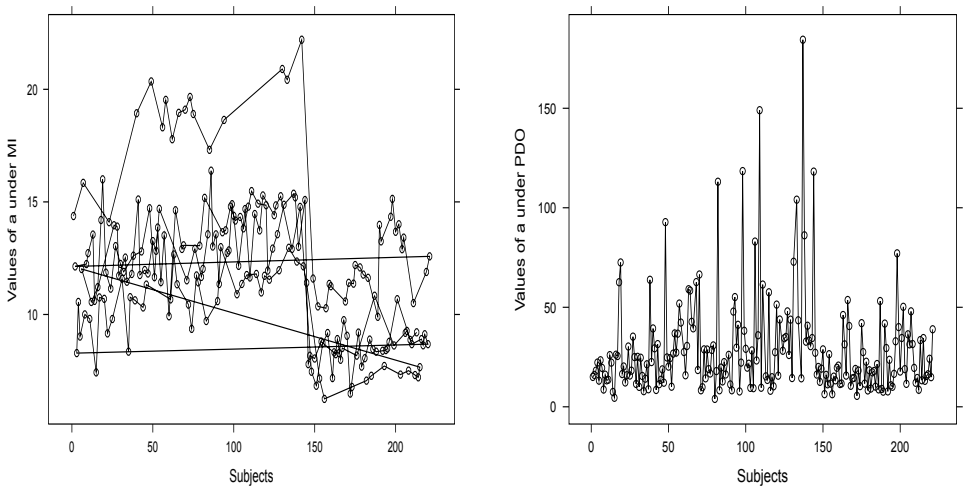
	Multiple imputation		Partial dropout	
Parameter	Value	95% Confidence limits	Value	95% Confidence limits
a	12.64	$(12.28, 13)$	28.39	$(25.11, 31.68)$
C_N	0.009	$(0.008, 0.011)$	0.018	$(0.016, 0.021)$
d	0.008	$(0.007, 0.01)$	0.227	$(0.158, 0.297)$
C_A	0.587	$(0.547, 0.626)$	1.291	$(1.08, 1.502)$
p	80.85	$(76.96, 84.74)$	129.64	$(116.24, 143.03)$
π	0.69	$(0.674, 0.707)$	0.207	$(0.182, 0.233)$
τ_{PI}	0.982	$(0.981, 0.984)$	0.488	$(0.44, 0.536)$
τ_{RTI}	0.742	$(0.706, 0.778)$	0.49	$(0.437, 0.543)$

a result of having more marker measurements in the multiple imputation case (3424) compared to 2024 in the partial dropout case. One would also compare the estimates under multiple imputation with those found under complete case analysis in Chapter 3. As expected, with more marker values and a bigger sample size, the results under multiple imputation analysis are more accurate that in the balanced data case.

Apart from the wide confidence intervals observed in the partial dropout scenario, it was also noted that the estimates assumed values that were mostly larger than in the multiple imputation case. Exploration of the parameter estimates revealed that there were some individuals that had extreme values. For instance, in the multiple imputation analysis, the maximum value of the individual estimate of a had a value of only 18.61 whereas in the partial dropout case the maximum value was as large as 184.44 and with more than half of the estimates taking values greater than 18.61.

The parameter estimates obtained using both multiply imputed and partially observed datasets are, however, within the values found in the literature. As an example consider the value of the activation rate of latently infected CD4+ T cells: the value in other discussions is 0.443 (Lavielle et al., 2011) while the current analyses give us 0.69 (multiply imputed data) and 0.207 (for partially observed data). Thus the only difference in these estimates are purely as a result of the nature of the data used in the estimation process, with the multiply imputed dataset offering smaller standard errors.

The plots in Figure 6.2 illustrate individual estimates under the two cases. We note that the values in the multiple imputation (MI) estimation suggests that there are two strata in the values, with one sub-population having relatively bigger estimate and the other one having smaller value (see the left panel). These results could correspond to facilities to which patients reported. The right panel also suggests the same trend but to a small degree. The other factor could be the differences in the rates of dropout for the two markers at the two facilities. These values are shown in Table 6.5 where it can be seen that the average dropout rate at Bwaila clinic is much higher (44.4%) than at Lighthouse clinic (35%). This suggests that there was a bigger proportion of the marker values to be imputed in the first facility



Multiple imputation Partial dropout

Figure 6.2: Individual estimates of a for the 214 subjects; MI = multiple imputation, PDO = partial dropout.

Table 6.5: Percentage of unobserved marker values

		Facility	
		Lighthouse	Bwaila
Marker	Viral load	42.9	47.4
	CD4+ T cell count	27	40.9

as compared to Lighthouse clinic.

6.5 Discussion

There are a number of strategies for handling missing and left-censored data. In this chapter we used multiple imputation based on a routine observational dataset from two HIV clinics. Since the data were unbalanced both among subjects and also among the markers within subjects, the multiple imputation was carried out for both markers (CD4+ cell counts and the viral load). Using the imputed data we estimated parameters of the HIV dynamical system and compared the results with those obtained using the incomplete data. This was done in the presence of several model covariates. The results, however, showed that estimates of covariate coefficients were not significant in the presence of complete data through multiple imputation compared to those obtained using data with partial dropout. Using the imputed data we also estimated the parameters of the within-host disease dynamics. It was observed that the standard errors of the estimates under imputed data were on average smaller than in the case of incomplete data. To the best of our knowledge, this is the first time multiple imputation has been applied in the estimation of parameters of the HIV dynamical system.

We also considered the effect of the limit of quantification (LOQs) of the assays when quantifying viral load on the estimates of the parameters. This was done by simulating datasets with desired LOQs. As shown in this study, the width of the confidence interval varies with the value of the detection limit. This means that a smaller LOQ results in a smaller standard error due to reduced uncertainty in the data. Like in the case of multiple imputation, the detection limit determines the size of the proportion of observations that are truncated to this limit. Thus as also

noted by Huang (2010) this LOQ influences the level-off which in turn provides the data with variability. The simulation study also offered the opportunity to observe that analyses involving data with smaller LOQ tend to depend less on covariates as shown by the level of significance of the covariate coefficients. This could be interpreted as the data being richer if its limit of quantification is smaller. In general this means that in the presence of a large number of left-censored values, covariates act as a source of additional information about the patients which is important in the estimation of HIV dynamical system parameters.

The inter-patient variability for all parameters in the two datasets with smaller limits of quantification (200 and 50 copies per ml) were significant ($p < 0.001$). Since it seemed like all the datasets provided significant results, in general, an observation using coefficients of variation shows a stark difference between the estimates of the parameters in the data with a LOQ of 400 and the other two. For example, the smallest coefficient of variation in the former set is 8% while the largest among the values of the other two sets is 7%. This is to be expected because for the same number of measurements there is little level-off effect in these sets that comes with data truncation resulting from lower quantification levels. In order to get accurate estimates of the system parameters there is a need for researchers to consider assays that have a much smaller detection limit as suggested by the results in this study.

In the analyses above, we assumed that dropout mechanism was missing at random so that multiple imputed could be possible. This may not have been the case because exploratory results indicate that the tendency to dropout depended on facility to which patient presented themselves. There was a need to use more datasets from multiple imputation instead of only six replications as done in our case. This would ensure parameters were estimated from a wide variety of complete

datasets which in turn would provide a basis for finding more efficient parameter estimates. Furthermore, the input in the simulation could have been enhanced by considering different time points for subjects as compared to the uniform occasions used in this analysis.

In summary this chapter concentrated on improving the richness of the bivariate longitudinal data so that parameters of the HIV dynamics could be estimated with greater accuracy. This has been done through multiple imputation of missing marker values and simulation of different sets based on the routine observational data using selected limits of quantification.

Chapter 7

Conclusion

7.1 Summary and discussion

This thesis was motivated by the need to estimate parameters of the HIV dynamical systems which could improve the knowledge of the disease progression and treatment efficacy. These dynamics involve a system of nonlinear ordinary differential equations in the two main disease biomarkers of CD4+ T cell count and viral load. The objective of such studies is to estimate parameters that characterise these disease dynamics. The most appropriate statistical models for carrying out the inferences of these parameters are nonlinear mixed-effects models for multivariate longitudinal data. These models are also used in various fields where growth and decay are involved like in yield modelling (Hall and Clutter, 2004; Forni et al., 2007).

In estimating the parameters we used two routine observational datasets from two HIV clinics under the Lighthouse Trust at Kamuzu Central Hospital in Malawi. The data were characterised by high proportions of missing observations due to dropout and left-censored viral load measures as a result of the outcomes being

below the limits of quantification. Overall it was observed that 41% of the outcomes were missing as a result of dropout and 39% of the viral load measures were below the limit of quantification. We considered three cases of the statistical model and each one provided estimates of parameters of the HIV disease dynamical system that were comparable with those found in the literature and in each case we included several covariates which was not the case in most discussions (Bortz and Nelson, 2006; Guedj et al., 2011). These three cases differ in the number of marker measurements and the covariates included in the analysis and also the number of occasions.

In the first approach, we considered a nonlinear mixed-effects model for complete multivariate longitudinal data involving 78 subjects observed on five different occasions including the baseline occasion. We included only three covariates in this setting because the others proved to be non-significant upon carrying out tests. Age-group (at commencement of treatment) was used as a grouping variable. The other covariates were compliance to treatment and supplementary treatment.

In the second scenario, we included observations from 214 patients and we assumed that all the subjects had a minimum of three and a maximum of seven bivariate observations. This ensured that we had enough information for the estimation process. There was a high percentage of patient dropout with only seven patients observed on the final occasion. This prompted us to propose a joint likelihood function of the multivariate response variable and the dropout occasions in order to facilitate the estimation of the parameters of the HIV disease dynamical system. Covariate tests provided the inclusion of one more covariate (sex of the patient) on the first case. This could be attributed to the dropouts that occurred in this case so that covariates acted as additional sources of information needed for the analyses.

The third scenario allowed the unbalancedness to occur among subjects such that subjects were observed with different numbers of occasions and also among the markers within subjects so that the number of markers was allowed to be different for each occasion. The number of viral load measurements was in general less than that of the CD4+ T cell counts largely because of resource constraints. Like in the second case, we included four covariates of age-group, compliance to treatment, supplementary treatment and facility to which patients presented themselves.

The differences among the three cases usually affect the degree of inter-individual variability which in turn results in differences in the estimates and their standard errors. However, the results can be of assistance in choosing the best models for a given dataset, while taking into consideration important properties of the data that have been described above. In so doing, the estimation of parameters of the HIV disease dynamics is also improved because we get more efficient estimates.

In this thesis, we have shown that the use of explanatory variables cannot be ignored, as done in similar studies, when making inferences about the system parameters (Audoly et al., 2001; Guedj et al., 2007a). Inclusion of covariates in the estimation of parameters provides an insight into the general properties of these parameters especially when such variables are significant to the analyses and when the data are unbalanced as a result of dropouts. For instance, we observed that gender (in the second scenario described above) of the patient influences the rate at which the actively infected CD4+ T cells die with a coefficient which is very significant ($p < 0.001$). This was also reflected in the third case where coefficients of the covariates for almost all the parameters were highly significant (see Table 5.3).

The nature of the data motivated us into proposing joint likelihood function of the observed multiple outcome response variable with left-censored data, the random-

effects, measurement error and the dropout mechanism in finding estimates in the case where dropout has been defined as all the response components not being observed once the subject withdraws from the study. However, it is common to have the response variable for a particular study subject to be observed partially so that only a subset of markers are not observable as described in case three above where for the same individual, we had more CD4+ T cell count measurements than the viral load measurements. In order to include all observations in the analyses, we have also proposed a joint likelihood of the multivariate response variable, the partial dropout and the subject-specific effects. This function has been used in the estimation of the HIV dynamical system parameters.

Our major contributions are presented in Chapters 4 and 5. Among other things, we showed that the dropout occasions, covariates and marker history play a significant role if included in the analyses because the parameters are estimated with smaller standard errors. We proceeded to find estimates of system parameters through a simulation study of marker data for various levels of limits of quantification (LOQ) of the viral loads. It has been observed that a smaller LOQ reduces uncertainty that comes with left-censored data because a smaller proportion of the values fall below LOQ. We also used multiple imputation to create six complete datasets which were used to find estimates of the system parameters. The estimates from these data had smaller confidence intervals than when partially observed data were used in such analyses.

As mentioned before, there are various possibilities for estimation of model parameters based on maximum likelihood approaches. In the sequel, we have used the stochastic approximation expectation maximisation (SAEM) algorithm for its convergence properties. This algorithm is designed in such a way that information from earlier steps is gradually dropped and more weight is placed on recent

steps with more accurate information relevant to the approximation (Delyon et al., 1999).

7.2 Limitations and future research

7.2.1 Limitations

In order to include practical aspects in the analyses, some of the covariates needed to be redefined. For example, supplementary treatment could be defined in such a way that the nature and severity of the illness that led to demand for supplementary treatment can be explicitly assessed. A proper documentation of getting supplementary treatment would also be a great advantage so that it is more accurately modelled and interpreted. In a similar manner, absence at time of appointment may not imply that the patient was not complying to treatment. It would be worthwhile to re-define this factor (treatment compliance) to include such situations as ensuring that the patient takes prescribed medication at designated times, that recommended dose and diet is being taken by the patient and prescribed schedules of patient physical exercises (or related health maintenance or improvement activities) are being followed by the subject under study among other elements (Pullar et al., 1989). These redefinitions would also ensure that the data is of high quality and rich enough for the estimation and interpretation of parameters of the complex HIV dynamical systems as the one considered in this thesis.

Another challenge is that the data used in our illustrations was obtained for routine check up and general patient monitoring. Moreover, there could also be some challenges with equipment used in the measurements as could be seen from a high

value of limit of quantification (LOQ) of the viral loads. Therefore, this type of data, may not have been collected with a precision needed for the estimation of parameters described in this study and in practice this cannot be avoided because of the design of the equipment used for taking measurements. As observed by Wu (2005), for instance, LOQ creates an artificial level-off effect in a dataset. This was another reason for the need to include several covariates which were deemed to provide additional information. In the illustrations we assumed that the observations were scheduled at equally-spaced occasions for each subject (see Rochon and Helms (1989) or Forcina (1992)). Most applications, however, allow for differences in observation intervals and fortunately, most statistical procedures accommodate such imbalances without overly compromising the consistence and efficiency of model estimators.

On the whole, however, these shortcomings did not outweigh our observations and results that the parameters in the dynamic systems are better estimated or approximated with reasonable practical reliability if system covariates are included in the analyses especially when there is a high proportion of dropout or censored data. Our results compared very well with other published results. For instance, our estimated value of a is in the interval $1.64 - 7.02$ copies per day for the various scenarios looked at in this thesis against the literature values (of 2.61 (Lavielle et al., 2011), 7.05 (Putter et al., 2002), 13.73 (Guedj et al., 2007b), 62 (Yu and Liang, 2013) and 98.1 in a discussion by Huang et al. (2006)). These varied literature values could be a result of a number of factors which may include the informativeness of the data used in the analyses, the assumptions placed on the models and also the subtypes of the HIV disease.

7.2.2 Future work

The analyses and approximation procedures presented here can be generalised to find applications in other disease modelling problems. For instance, in modelling other chronic illnesses like cancer one may determine the efficacy of treatment and the impact of treatment compliance on response to treatment regimen (among other system parameters) by applying the procedures discussed in this thesis. One could also consider using functional data analysis so that samples of random functions of the observations from the clinics are used in estimating the model parameters (Martínez-Camblor and Corral, 2011). Furthermore, treatment could be time-variant and this needs to be factored when developing statistical parameter estimation models (Jones and Boadi-Boateng, 1991).

In our applications, it is to be expected that there is some degree of correlation in the occasions at which the markers stop being observed, random-effects and the measurement errors of the markers. There was a need to explicitly model the serial correlation between these quantities as done in Muñoz et al. (1992) or Lipsitz et al. (2009) especially when setting up partial dropout models. Without such correlation process included, the models of the partial dropout process proposed in Chapter 5 of this thesis may not have provided efficient estimates as would be required.

To get the estimates of the parameters of the HIV disease dynamics, we made several assumptions about the parameters, the models and also the data. However, there is a need to be cautious when making such assumptions. For instance, the CD4+ T cell count could have been transformed usually by taking fourth-root or square-root in order to achieve a normality assumption. As a result of this, in each of the three cases discussed in this thesis we have fitted data by restricted maximum likelihood in order to minimize the unweighted least squares

loss function. It is also well-known that there have been certain cases of viral rebound which may be associated with resistance to treatment regimen (Fitzgerald et al., 2002; de Leenheer and Smith, 2003; Rong et al., 2007; Witten and Perelson, 2004). There is a need, therefore, to include parameters that characterising rate of viral rebound and elements of disease resistance in the estimation of parameters that characterise the HIV dynamic models considered in this thesis.

There are practical situations where we observe change of treatment regimen in the course of a study, especially in long-term studies (Chen et al., 2013). Inclusion of a covariate to this effect could also be appropriate when estimating parameters of the biological process.

When setting up statistical models, we employed hierarchical nonlinear mixed-effects model in order to characterise the population and subject-specific variability (Bortz and Nelson, 2006). One could consider Bayesian approaches in estimating disease parameters using non-informative prior distributions and available data as done in Huang et al. (2006); Huang and Lu (2008); Huang (2010). This would ensure that prior information about the system parameters in used in the estimation process.

References

Albert, P. S. and Shih, J. H. (2010). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4(3):1517–1532.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, 28(3):301–309.

Apiolaza, L. A. and Garrick, D. J. (2001). Analysis of longitudinal data from progeny tests: Some multivariate approaches. *Forest Science*, 47(2):129–140.

Audoly, S., Bellu, G., D'Angiò, L., Saccomani, M. P., and Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering*, 48(1):55–65.

Beckett, L. A., Tancredi, D. J., and Wilson, R. S. (2004). Multivariate longitudinal models for complex change processes. *Statistics in Medicine*, 23(2):231–239.

Blozis, S. A., Conger, K. J., and Harring, J. R. (2007). Nonlinear latent curve models for multivariate longitudinal data. *International Journal of Behavioral Development*, 31(4):340–346.

Bortz, D. M. and Nelson, P. W. (2006). Model selection and mixed-effects modeling of HIV infection dynamics. *Bulletin of Mathematical Biology*, 68(8):2005–2025.

Brancato, G., Pezzotti, P., Rapiti, E., Perucci, C. A., Abeni, D., Babbalacchio,

A., Rezza, G., and the Multicenter Prospective HIV study (1997). Multiple imputation for estimating incidence of HIV infection. *International Journal of Epidemiology*, 26(5):1107–1114.

Cai, B., Dunson, D. B., and Stanford, J. B. (2010). Dynamic model for multivariate markers of fecundability. *Biometrics*, 66(3):905– 913.

Chaganty, N. R. and Naik, D. N. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, 103(1-2):421–436.

Chen, Q., Chen, M. H., Ohlssen, D., and Ibrahim, J. G. (2013). Bayesian modeling and inference for clinical trials with partial retrieved data following dropout. *Statistics in Medicine*, 32:4180–4195.

Cnaan, A., Laird, N. M., and Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16:2349–2380.

Crouchley, R. and Ganjali, M. (2002). The common structure of several models for nonignorable dropout. *Statistical Modeling*, 2:39–62.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, 31(3):371–403.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. Monographs on Statistics and Applied Probability 62. Chapman and Hall.

de Leenheer, P. and Smith, H. L. (2003). Virus dynamics: Global analysis. *SIAM Journal of Applied Mathematics*, 63(4):1313–1327.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of the stochastic approximation version of the EM Algorithm. *The 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):1–38.

Deslandes, E. and Chevret, S. (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: Application to ICU data. *BMC Medical Research Methodology*, 69:1–13.

Diggle, P. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics*, 43(1):49–93.

Diggle, P. J., Heagrty, P., Liang, K., and Zeger., S. L. (2002). *Analysis of longitudinal data*. Oxford Statisticl Science Series 25. Oxford University Press, second edition.

Dubin, J. A. and Müller, H. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881.

Duffin, R. P. and Tullis, R. H. (2002). Mathematical models for the complete course of HIV infection and AIDS. *Journal of Theoretical Medicine*, 4(4):215–221.

Encrenaz, G., Rondeau, V., Messiah, A., and Auriacombe, M. (2005). Examining the influence of dropouts in a follow-up of maintained opiate users. *Drug and Alcohol Dependence*, 79:303–310.

Everitt, B. S. (1998). Analysis of longitudinal data: Beyond MANOVA. *The British Journal of Psychiatry*, 172(1):7–10.

Feinberg, M. B. (1996). Changing the natural history of HIV disease. *Lancet*, 348:239–246.

Ferrer, E. and MacArdle, J. J. (2003). Alternative structural models for multivari-

ate longitudinal data analysis. *Structural Equation Modelling*, 10(4):493–524.

Ferrer, E., Salthouse, T. A., and McArdle, J. J. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging*, 20(3):412–422.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal data. *Biometrics*, 62(2):424–431.

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

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009). *Handbook of modern statistical methods: Longitudinal data analysis*. Chapman and Hall, New York.

Fitzmaurice, G. M. (2003). Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica*, 57(1):75–99.

Fitzmaurice, G. M. and Laird, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, 1(2):141–156.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied longitudinal analysis*. John Wiley and Sons Inc., New Jersey.

Forcina, A. (1992). Modelling balanced longitudinal data: Maximum likelihood estimation and analysis of variance. *Biometrics*, 48(3):743–750.

Forni, S., Piles, M., Blasco, A., Varona, L., Oliveira, H. N., Lôbo, R. B., and Albuquerque, L. G. (2007). Analysis of beef cattle longitudinal data applying a nonlinear model. *Journal of Animal Science*, 85:3189–9197.

Gad, A. M. and Ahmed, A. S. (2006). Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. *Computational Statistics and Data Analysis*, 50:2702–2714.

Gelman, A., Hill, J., Su, Y., Yajima, M., and Pittau, M. G. (2013). *Package 'mi': Missing data imputation and model Checking; Version 0.09-18.03*. R Foundation for Statistical Computing.

Govender, S., Otwombe, K., Essien, T., Panchia, P., de Bruyn, G., Mohapi, L., Gray, G., and Martinson, N. (2014). CD4 counts and viral loads of newly diagnosed HIV infected individuals: Implications for treatment as prevention. *PLoS One*, 9(3):1–7.

Guedj, J., Thièbaut, R., and Commenges, D. (2007a). Maximum likelihood estimation in dynamic models of HIV. *Biometrics*, 63(4):1198–1206.

Guedj, J., Thièbaut, R., and Commenges, D. (2007b). Practical identifiability of HIV dynamic models. *Bulletin of Mathematical Biology*, 69(8):2493–2513.

Guedj, J., Thièbaut, R., and Commenges, D. (2011). Joint modeling of the clinical progression and of the biomarkers' dynamics using a mechanistic model. *Biometrics*, 67(1):59–66.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, 1(3):177–193.

Hall, D. B. and Clutter, M. (2004). Multivariate multilevel nonlinear mixed-effects models for timber yield predictions. *Biometrics*, 60:16–24.

Harvey, D., Beckett, L. A., and Mungas, D. M. (2003). Multivariate modeling of two associated cognitive outcomes in a longitudinal study. *Journal of Alzheimer's Disease*, 5(5):357–365.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley and Sons Inc., New Jersey.

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Biostatistics tutorial: Handling dropout in longitudinal studies. *Statistics in Medicine*, 3:1455–1497.

Hu, C. and Sale, M. E. (2003). A joint model for nonlinear longitudinal data with informative dropout. *Journal of Pharmacokinetics and Pharmacodynamics*, 30(1):83–103.

Huang, Y. (2010). A Bayesian approach in differential equation dynamic models incorporating clinical factors and covariates. *Journal of Applied Statistics*, 37(2):181–199.

Huang, Y., Liu, D., and Wu, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, 62:413–423.

Huang, Y. and Lu, T. (2008). Modeling long-term longitudinal HIV dynamics with application to an AIDS clinical study. *The Annals of Applied Statistics*, 2(4):1384–1408.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *TEST*, 18:1–43.

Jia, J. and Weiss, R. (2009). Common predictor effects for multivariate longitudinal data. *Statistics in Medicine*, 28(13):1793–1804.

Jones, R. H. (2000). *Longitudinal data with serial correlation: A state-space approach*. Monographs on Statistics and Applied Probability 47. Chapman and

Hall/CRC.

Jones, R. H. and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics*, 47(1):161–175.

Jorgensen, B., Lundbye-Christensen, S., Song, P. X., and Sun, L. (1996). State-space models for multivariate longitudinal data of mixed types. *The Canadian Journal of Statistics*, 24(3):385–402.

Kmetic, A., Joseph, L., Berger, C., and Tenenhouse, A. (2002). Multiple imputation to account for missing data in a survey: Estimating the prevalence of osteoporosis. *Epidemiology*, 13(4):437–444.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed-effects models. *Computational Statistics and Data Analysis*, 49(4):1020–1038.

Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, 82(397):97–105.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7(1-2):305–315.

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

Lavielle, M., Samson, A., Fermin, A. K., and Mentrè, F. (2011). Maximum likelihood estimation of long-term HIV dynamic models and antiviral response. *Biometrics*, 67:250–259.

Lee, K., Joo, Y., Yoo, J. K., and Lee, J. (2009). Marginalized random-effects models for multivariate longitudinal binary data. *Statistics in Medicine*, 28(8):1284–1300.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2):219–238.

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

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., Parzen, M., and Lipshultz, S. (2009). Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to AIDS data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):3–20.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley and Sons Inc., New York.

Liu, M., Taylor, J. M. G., and Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, 56(4):1157–1163.

Liu, W. and Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-Effect models with covariate measurement errors and missing responses. *Biometrics*, 63(2):342–350.

Lloyd, J. E., Obradović, J., Carpiano, R. M., and Motti-Stefanidi, F. (2013). Multiple imputation of missing multilevel, longitudinal data: A case when practical considerations trump best practices? *Journal of Modern Applied Statistical Methods*, 12(1):261–275.

Lundbye-Christensen, S. (1991). A multivariate growth curve model for pregnancy. *Biometrics*, 47(2):637–657.

Ma, C. X., Li, Y., and Wu, R. (2008). Modeling the genetic control of HIV-1 dynamics after highly active antiretroviral therapy. *Current Genomics*, 9:208–211.

Marini, A., Harper, J. M., and Romerio, F. (2008). An *In Vitro* system to model the establishment and reactivation of HIV-1 latency. *The Journal of Immunology*, 181(11):7713–7720.

Marshall, G., de la Cruz-Mesía, R., Barón, A. E., Rutledge, J. H., and Zerbe, G. O. (2006). Nonlinear random-effects model for multivariate responses with missing data. *Statistics in Medicine*, 25(16):2817–2830.

Martínez-Camblor, P. and Corral, N. (2011). Repeated measures analysis for functional data. *Compuational Statistics and Data Analysis*, 55:3244–3256.

Mata-Marín, J. A., Gaytán-Martínez, J., Grados-Chavarría, B. H., Fuentes-Allen, J. L., Arroyo-Anduiza, C. I., and Alfaro-Mejía, A. (2009). Correlation between HIV viral load and aminotransferases as liver damage markers in HIV infected naive patients: A concordance cross-sectional study . *Virology Journal*, 6:181–184.

Meier, A., Chang, J. J., Chan, E. S., Pollard, R. B., Sidhu, H. K., Kulkarni, S., Wen, T. F., Lindsay, R. J., Orellana, L., Mildvan, D., Bazner, S., Streeck, H., Alter, G., Lifson, J. D., Carrington, M., Bosch, R. J., Robbins, G. K., and Altfeld, A. (2009). Sex differences in the toll-like receptor-mediated response of plasmacytoid dendritic cells to HIV-1. *Nature Medicine*, 15(8):955–959.

Meza, C., Jaffrézic, F., and Foulley, J. L. (2007). REML estimation of variance parameters in nonlinear mixed-effects models using the SAEM algorithm. *Biometrical Journal*, 49(6):876–888.

Meza, C., Osorio, F., and De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, 12:121–139.

Mickey, R. M., Shema, S. J., Vacek, P. M., and Bell, D. Y. (1994). Analaysis of

multiple outcome variables measured longitudinally. *Computational Statistics and Data Analysis*, 17(1):17–33.

Ministry of Health (2008). *Treatment of AIDS: Guidelines for the use of antiretroviral therapy in Malawi*. Ministry of Health, Secretary for Health, P.O. Box 30377, Lilongwe 3, Malawi, 3rd edition.

Molenberghs, G. and Kenwrad, M. G. (2007). *Missing data in clinical studies*. Statistics in Practice. John Wiley and Sons Ltd., West Sussex.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M., and Rosner, B. (1992). A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48:733–742.

Newsom, J. T., Jones, R. N., and Hofer, S. M., editors (2012). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science*. Taylo and Francis Group.

Nowak, M. A. and May, R. (2000). *Virus dynamics: Mathematical principles of immunology and virology*. Oxford University Press.

Nummi, T. and Möttönen, J. (2000). On the analysis of mulivariate growth curves. *Metrika*, 52:77–89.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 54(1):49–78.

Pan, W. and Louis, T. A. (2000). A linear mixed-effects model for multivariate censored data. *Biometrics*, 56:160–166.

Pantazis, N., Touloumi, G., Walker, A. S., and Babiker, A. G. (2005). Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative dropout. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54(2):405–423.

Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nature Reviews: Immunology*, 2:28–36.

Perelson, A. S. and Nelson, P. W. (1999). Mathematical analysis of HIV-1: Dynamics *In Vivo*. *SIAM Review*, 41(1):3–44.

Potthoff, R. F. and Roy, S. N. (1964). A generalised multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3 and 4):313–326.

Pullar, T., Kumar, S., and Feely, M. (1989). Compliance in clinical trials. *Annals of Rheumatic Diseases*, 48:871–875.

Putter, H., Heisterkamp, S. H., Lange, J. M., and de Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine*, 21(15):2199–2214.

Qu, A. and Song, P. X. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, 89(4):841–850.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77(377):190–195.

Reinsel, G. (1984). Estimation and prediction in a multivariate random-effects generalized linear model. *Journal of the American Statistical Association*, 79(386):406–414.

Ribeiro, R. M. (2007). Dynamics CD4+ T cells in HIV-1 infection (Review).

Immunology and Cell Biology, 85:287–294.

Ribeiro, R. M., Mohri, H., Ho, D. D., and Perelson, A. S. (2002). *In Vivo* dynamics of T cell activation, proliferation, and death in HIV-1 infection: Why are CD4+ but not CD8+ cells depleted? *Proceedings of the National Academy of Sciences*, 99(24):15572–15577.

Rice, J. P., Rochberg, N., Neuman, R. J., Saccone, N. L., Liu, K. Y., Zhang, X., and Culverhouse, R. (1999). Covariates in linkage anaysis. *Genetic Epidemiology*, 17:S691–S695.

Rochon, J. and Helms, R. W. (1989). Maximum likelihood estimation for incomplete repeated-measures experiments under an ARMA covariance structure. *Biometrics*, 45(1):207–218.

Romih, V., Lepej, S. Z., Gedike, K., Lukas, D., and Begovac, J. (2010). Frequency of HIV-1 viral load monitoring of patients initially successfully treated with combination antiretroviral therapy. *PLoS One*, 5(11):1–7.

Rong, L., Feng, Z., and Perelson, A. S. (2007). Mathematical analysis of age-structured HIV-1 dynamics with combination antiretroviral therapy. *SIAM Journal of Applied Mathematics*, 67(3):731–756.

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

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonnignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association*, 97(457):40–52.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley and Sons Inc.

Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the SAEM al-

gorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics and Data Analysis*, 51(3):1562 – 1574.

Samson, A., Lavielle, M., and Mentré, F. (2007). The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model. *Statistics in Medicine*, 26(27):4860–4875.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Monographs on Statistics and Applied Probability 72. Chapman and Hall, first edition.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing data. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

Shah, A., Laird, N., and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, 92(438):775–779.

Singh, S. (2009). A new method of imputation in survey sampling. *Statistics*, 43(5):499–511.

Steiner, P. M., Cook, T. D., Shadish, W. R., and Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3):250–267.

Sterne, J. A., White, I. R., Carlin, J., Spratt, M., Royston, P., Kenward, M., and Wood, A. M. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 339:157–160.

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, 169(9):1133–1139.

Sy, J. P., Taylor, J. M. G., and Cumberland, W. G. (1997). A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, 53(2):542–555.

Thièbaut, R., Jacqmin-Gadda, H., Leport, C., Katlama, C., Costagliola, D., Le Moing, V., Morlat, P., Chêne, G., and the APROCO Study Group (2003). Bivariate longitudinal model for the analysis of the evolution of HIV RNA and CD 4 cell count in HIV infection taking into account left censoring of HIV RNA measures. *Journal of Biopharmaceutical Statistics*, 13(2):271–282.

Twist, J. W. R. (2003). *Applied longitudinal data analysis for epidemiology: A practical guide*. Cambridge University Press, New York, first edition.

van der Eijk, A. A., Hansen, B. E., Niesters, H. G., Janssen, H. L., van de Ende, M., Schalm, S. W., and de Man, R. A. (2005). Viral dynamics during tenofovir therapy in patients infected with lamivudine-resistant hepatitis B virus mutants. *Journal of Viral Hepatitis*, 12(4):364–372.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Series in Statistics. Springer Verlag, New York.

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, 58(2):220–233.

Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 48(1):1–17.

Walker, S. (1996). An EM algorithm for nonlinear random-effects model. *Biometrics*, 52(3):934–944.

Weiss, R. E. (2005). *Modeling longitudinal data*. Springer Texts in Statistics. Springer, New York.

Witten, G. Q. and Perelson, A. S. (2004). Modelling cellular-level interaction be-

tween the immune system and HIV. *South African Journal of Science*, 100:447–451.

Wu, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trials. *Statistical Methods in Medical Research*, 14(2):171–192.

Wu, H., Zhu, H., Miao, H., and Perelson, A. S. (2008). Parameter identifiability and estimation of HIV/AIDS dynamic models. *Bulletin of Mathematical Biology*, 70(3):785–799.

Wu, K. and Wu, L. (2007). Generalized linear mixed models with informative dropouts and missing covariates. *Metrika*, 66(1):1–18.

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, S. and Müller, H. G. (2011). Response-adaptive regression for longitudinal data. *Biometrics*, 67:852–860.

Xia, X. and Moog, C. H. (2003). Identifiability of nonlinear systems with application to HIV/AIDS models. *IEEE Transactions on Automatic Control*, 48(2):330–336.

Yu, Y. and Liang, H. (2013). Parameter estimation for HIV ODE models incorporating longitudinal structure. *Statistics and Its Interface*, 6:9–18.

Zhang, H. (2004). Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Statistical Methods in Medical Research*, 13:63–82.

Appendices

Appendix A

This is an R code that was used in creating longitudinal data and also for carrying out multiple imputation. It also includes elements used for making inferences in Section 6.4 of Chapter 6.

```
VL<-read.csv("VL_long_partial.csv")
VL_par<-reshape(VL,direction="long",idvar="ID",varying=
+colnames(VL)[-(1:6)])
CD<-read.csv("CD_long_partial.csv")
LonCD_par<-reshape(CD,direction="long",idvar="ID",varying=
+colnames(CD)[-(1:6)])
CD_par=sqrt(LonCD_par[,8])
HIV_par=data.frame(VL_par,CD_par)
library(arm)
library(mi)
IM_HIV<-mi (HIV_par,n.imp=8, n.iter=500 , add.noise=TRUE)
imputed.HIV<-mi.completed(IM_HIV)
write.csv(imputed.HIV,"Imputes_6.csv")
ind_MI<-read.csv("indiv_MI.csv")
.....................................
means_MI<-colMeans(ind_MI[,c(2:9)])
```

```
means_MI_1<-colMeans(ind_MI_1[,c(2:9)])
.....................................
mean_final<-1/6*(means_MI+means_MI_1+means_MI_2+means_MI_3+
+means_MI_4+means_MI_5)
sd_MI<-sapply(ind_MI[,c(2:9)],sd)
sd_MI_1<-sapply(ind_MI_1[,c(2:9)],sd)
.....................................
V_AVSD<-1/6*(sd_MI^2+sd_MI_1^2+sd_MI_2^2+sd_MI_3^2+sd_MI_4^2+sd_MI_5^2)
V_BIM<-(means_MI-mean_final)^2
V_BIM1<-(means_MI_1-mean_final)^2
.....................................
V_BIM_fin<-7/30*(V_BIM+V_BIM1+V_BIM2+V_BIM3+V_BIM4+V_BIM5)
sd_final<-sqrt(V_AVSD+V_BIM_fin)
.....................................
V_im=sd_final/sqrt(N)
imputed_par_error<-qnorm(0.975)*V_im
lower_imp_limit<-mean_final-imputed_par_error
upper_imp_limit<-mean_final+imputed_par_error
intervals<-data.frame(lower_imp_limit,upper_imp_limit)
.....................................
```

Appendix B

Part of the partially observed data (for two subjects) used in the analyses of Chapters 5 and 6 conducted in MONOLIX. The headers are patient's ID, time in weeks, marker measurements, whether censored or not, marker type (1 = Viral load, 2 = CD4+ T cell count) and the rest are covariates.

Pat	time	Y	Censor	Marker	Supp	Compl	Fac	Sex	Age
1	0	207	0	2	0	1	1	0	A
1	0	4.0208	0	1	0	1	1	0	A
1	14	410	0	2	0	1	1	0	A
1	14	2.6021	1	1	0	1	1	0	A
1	28	317	0	2	0	1	1	0	A
1	28	3.3034	0	1	0	1	1	0	A
1	42	.	.	2	0	1	1	0	A
1	42	.	.	1	0	1	1	0	A
1	56	.	.	2	0	1	1	0	A
1	56	.	.	1	0	1	1	0	A
1	70	.	.	2	0	1	1	0	A
1	70	.	.	1	0	1	1	0	A
1	84	.	.	2	0	1	1	0	A

1	84	.	.	1	0	1	1	0	A
1	98	.	.	2	0	1	1	0	A
1	98	.	.	1	0	1	1	0	A
2	0	443	0	2	1	1	1	1	B
2	0	4.8902	0	1	1	1	1	1	B
2	14	600	0	2	1	1	1	1	B
2	14	3.1482	0	1	1	1	1	1	B
2	28	530	0	2	1	1	1	1	B
2	28	4.399	0	1	1	1	1	1	B
2	42	540	0	2	1	1	1	1	B
2	42	4.7296	0	1	1	1	1	1	B
2	56	730	0	2	1	1	1	1	B
2	56	2.6493	0	1	1	1	1	1	B
2	70	.	.	2	1	1	1	1	B
2	70	.	.	1	1	1	1	1	B
2	84	.	.	2	1	1	1	1	B
2	84	.	.	1	1	1	1	1	B
2	98	.	.	2	1	1	1	1	B
2	98	.	.	1	1	1	1	1	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Appendix C

This is part of the MONOLIX code that was used in the estimation of parameters that characterise the HIV dynamical system.

```
<project mlxVersion="4.1.0.125" name="Luwanda_project.xmlx">
<covariateDefinitionList>
<covariateDefinition columnName="Supp" type="continuous"/>
<covariateDefinition columnName="Compl" type="continuous"/>
<covariateDefinition columnName="Fac" type="continuous"/>
<covariateDefinition columnName="Age" type="categorical">
.....................................
<groupList>
 <group name="A" reference="true"/>
<group name="B"/><group name="C"/>
<group name="D"/>
</groupList></covariateDefinition>
</covariateDefinitionList>
<data columnDelimiter="\t" headers="Pat,time,Y,
Censo,Marker,COV,COV,COV,IGNORE,CAT"
<models><statisticalModels>
.....................................
```

```
 <parameterList>
<parameter name="a" transformation="L">
<intercept initialization="10"/>
<betaList>
<beta covariate="Supp" initialization="0"/>
<beta covariate="Compl" initialization="0"/>
<beta covariate="Fac" initialization="0"/>
<beta covariate="Age" initialization="0"/>
</betaList>
.....................................
<variability initialization="1"
<level="1" levelName="IIV"/>
</parameter>
<parameter name="gamma0" transformation="L">
<intercept initialization="0.015"/><variability
initialization="1" level="1" levelName="IIV"/>
</parameter><parameter name="p" transformation="G">
<intercept initialization="0.5"/>
.....................................
.....................................
```

Appendix D

This is part of the R code that was used in carrying out covariates analyses used in Chapter 2 and also for computation of confident limits of the system parameters.

```
attach(HIV_par)
print(xtabs(~SUPPLEMENTARY + COMPLIANCE, data = HIV_par))
#THE CODES BELOW GIVE MARGINAL TOTALS/PROPORTIONS
margin.table(tab_suppcom, 1)#SUMMED OVER COMPLIANCE
margin.table(tab_suppcom, 2)#SUMMED OVER SUPPLEMENTARY
prop.table(tab_suppcom,2)#GIVES COLUMN PERCENTAGES
tab_ga<-table(GENDER,AGE)
tab_ga
prop.table(tab_ga)
margin.table(prop.table(tab_ga),2)*100
chisq.test(tab_ga)
.....................................
tab_gf<-table(GENDER,FACILITY)
tab_gf
chisq.test(tab_gf)
margin.table(tab_gf,2)#SUM FOR FACILITY
.....................................
```

```
##CONFIDENCE INTERVALS OF PARS IN AGE GROUPS FOR PAPER 3
partial_means<-read.csv("indiv_pars_partial.csv")
attach(partial_means)
library(lattice)
library(Matrix)
.................
#INTERVALS FOR A
partial_means_A1<-AGE_A*partial_means[,c(2:19)]
partial_means_A<-read.csv("partial_means_a2.csv")
means_parta<-colMeans(partial_means_A[,c(2:12)])
means_parta
sd_parta<-sapply(partial_means_A[,c(2:12)],sd)
V_a=sd_parta/sqrt(sum(AGE_A))
part_errora<-qnorm(0.975)*V_a/sqrt(sum(AGE_A))
lower_limita<-means_parta-part_errora
upper_limita<-means_parta+part_errora
intervalsa<-data.frame(lower_limita,upper_limita)
intervalsa
.......................................
## CREATED FOR TABLE 3 OF PAPER 3
FOR RESUBMISSION 02.02.14 (AFTERNOON) ##RESULTS FOR LME WERE
DEPARTING SIGNIFICANTLY FROM THOSE FOUND BYE #DIRECT EVALUATION G
##MONOLIX; HENCE CHOICE OF GLM fitted_d2<-glm(d~
SUPP+COMPL+FAC,family = gaussian(link=log), +data=partial_means)
summary(fitted_d2) fitted_p2<-glm(p~ SUPP+COMPL+FAC,family =
gaussian(link=log), +data=partial_means) summary(fitted_p2)
```

```
.....................................
#CODE FOR COMPUTATION OF INTER-INDIVIDUAL VARIABILITY
#22/9/13 (EVENING) etas_sd<-read.csv("indiv_eta_par.csv")
attach(etas_sd) F_b<-AGE_B*etas_sd[,c(2:12)]
write.csv(F_b,"F_bB.csv") etas_sd_B<-read.csv("F_bB.csv")
etas_meanB<-colMeans(etas_sd_B) etas_meanB
sd_etaB<-sapply(etas_sd_B,sd) V_etas_B=sd_etaB/sqrt(sum(AGE_B))
V_etas_B ....................................
```