A Continuous Latent Factor Model for Non-ignorable Missing Data

Jun Zhang* Mark Reiser†

Abstract

Many longitudinal studies, especially in clinical trials, suffer from missing data issues. Most estimation procedures assume that the missing values are ignorable. However, this assumption leads to unrealistic simplification and is implausible for many cases. When non-ignorable missingness are preferred, classical pattern-mixture models with the data stratified according to a variety of missing patterns and a model specified for each stratum, are widely used for longitudinal data analysis. But this assumption usually results in under-identifiability because of the need to estimate many stratum-specific parameters. Further, pattern mixture models have the drawback that a large sample is usually required. In this paper, the continuous latent factor model is proposed and this novel approach overcomes limitations which exist in pattern mixture models by specifying a continuous latent factor. The advantages of this model, including small sample feasibility, are evaluated by comparing with Roy's pattern mixture model, based on simulations and an application on a clinical study of AIDS patients with advanced immune suppression.

Key Words: Missing data mechanism, missing patterns, nonignorability, continuous latent factor

1. Introduction

Missing values in multivariate studies pose many challenges. The primary research of interest focuses on accurate and efficient estimation of means and covariance structure in the population. The assumption and estimation of the covariance structure provide the foundation of many statistical models, for instance, structural equation modeling, principle component analysis, and so on. Literature on multivariate missing data methods was reviewed by Little and Rubin (2002), and Schafer (1997). For some frequentist statistical procedures, we may generally ignore the distribution of missingness only when the missing data are missing completely at random (MCAR), such as in the generalized estimation equations (GEE) estimation procedure. For likelihood or Bayes procedures, however, we may ignore the missing values when the missing data are missing at random (MAR), as in for example, the estimation procedure for linear mixed models. However, if missing at random in the data is questioned, and one suspects that the missing mechanism is NMAR, i.e. missingness may depend on missing values, then the joint modeling of the complete data and the missing indicators is required. The reason to follow this modeling method is that the resulting estimates of population parameters may be biased (Pirie and Leupker, 1988) unless these NMAR aspects of the data are taken into account in the analysis. Furthermore, the results of the study may not be feasible to generalize, because the observed respondents may not represent the target population. From a practical aspect, investigators could not point out whether violations of the MAR assumption are severe enough to result in a conclusions that are not valid.

Models for NMAR data have been proposed for a few decades, including selection models (Diggle and Kenward, 1994a), pattern-mixture models (Diggle and Kenward, 1994b), as well as shared-parameter models (Diggle and Kenward, 1994b). Many researchers have extended this field in the last decade. Some authors have incorporated latent class structure into pattern-mixture models to jointly describe the pattern of missingness and the outcome

*Bayer Healthcare Pharmaceuticals, 100 Bayer Blvd, Whippany, NJ 07981

†Arizona State University, Tempe, AZ 85287

of interest (Lin et al., 2004; Muthn et al., 2003; Roy, 2003). Lin et al. (2004) proposed a latent pattern-mixture model where the mixture patterns are formed from latent classes that link a longitudinal response with a missingness process. Roy (2003) investigated latent classes to model dropouts in longitudinal studies to effectively reduce the number of missing-data patterns. Muthen et al. (2003) also discussed how latent classes could be applied to non-ignorable missingness. Jung et al. (2011) extended traditional latent class models, where the classes are defined by the missingness indicators alone.

All the above extensions are from the family of pattern-mixture models, and these models stratify the data according to time to dropout or missing indicators alone and formulate a model for each stratum. This usually results in under-identifiability, since we need to estimate many pattern-specific parameters even though the eventual interest is usually on the marginal parameters. As the alternative, Guo et al. (2004) extended pattern-mixture to a random pattern-mixture model for longitudinal data with dropouts. The extended model works effectively on the case where a good surrogate for the dropout can be representative for the dropout process. In most real studies, however, it maybe impossible to find good measures for the missing mechanism. For instance, in a longitudinal study with many intermittent missing values, time to dropout is not necessarily a good measure, and it probably wouldn't capture most features of missingness. That is, this measurement can not represent for subjects who have drop-in responses. Instead, modeling for missing indicators is necessary in this case. Further, models other than the normal distribution will be required to describe the missingness process. The violation of joint multivariate normality will lead to an increase of computation difficulties. In the proposed new model, missing indicators are directly modeled with a continuous latent variable, and this latent factor is treated as a predictor for latent subject-level random effects in the primary model of interests. Some informative variables related with missingness (e.g. time to first missing, number of switches between observed and missing responses) will be served as covariates in the modeling of missing indicators. The detailed description of the new model will be given in next section.

2. Models and Estimation

2.1 Proposed Model

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

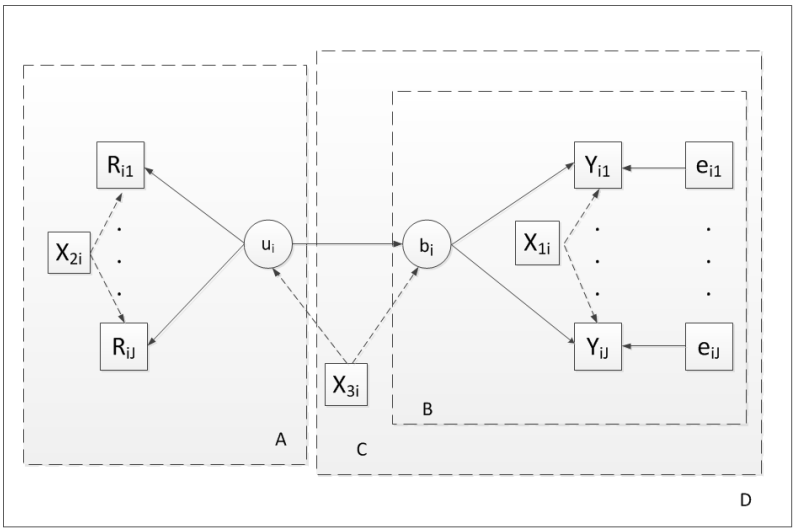

Figure 1: Proposed model diagram: observed quantities are described in squared boxes, latent quantities are in circled boxes

One of the fundamental assumptions of this new model is that $\mathbf{Y}_i$ is conditionally independent of $\mathbf{R}_i$ given the latent variables u_i and $\mathbf{b}_i$. This is a natural assumption when modeling relationships between variables measured with error, i.e., we want to model the relationship between the underlying variables, not the ones with error. Finally, we assume that $\mathbf{Y}_i$ is conditionally independent of u_i given $\mathbf{b}_i$, and likewise, $\mathbf{R}_i$ is conditionally independent of $\mathbf{b}_i$ given u_i. Hence, we introduce the following model for the joint distribution of the responses $\mathbf{Y}_i$ and missing indicators $\mathbf{R}_i$,

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

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

$$(\mathbf{Y}_i|\mathbf{b}_i,\ \mathbf{X}_{1i}) \sim_{ind} N_J(\mathbf{X}_{1i}\beta + \mathbf{Z}_{1i}\mathbf{b}_i,\ \Sigma_\epsilon) \tag{2}$$

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

$$u_i \sim_{ind} N_1(0,\ \sigma_u^2) \tag{4}$$

$$f(\mathbf{R}_i|u_i,\ \mathbf{X}_{2i}) = \prod_{j=1}^{J} \pi_{ij}^{r_{ij}} (1-\pi_{ij})^{1-r_{ij}} \tag{5}$$

A linear mixed model (growth curve) is used for the relationship between $\mathbf{Y}_i$ and $\mathbf{b}_i$ (model B in Figure 1), where $\mathbf{X}_{1i}$ is a known $(J \times p_1)$ design matrix containing fixed within-subject and between-subject covariates (including both time-invariate and time-varying covariates), with associated unknown $(p_1 \times 1)$ parameter vector β, $\mathbf{Z}_{1i}$ is a known $(J \times q)$ matrix for modeling random effects, and $\mathbf{b}_i$ is an unknown $(q \times 1)$ random coefficient vector. We specify $\mathbf{Y}_i = \mathbf{X}_{1i}\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$, where the random error term ϵ_i is a J-dimensional vector with $E(\epsilon_i) = \mathbf{0}$, $Var(\epsilon_i) = \Sigma_\epsilon$, and ϵ_i is assumed independent of $\mathbf{b}_i$. Furthermore, the $J \times J$ covariance matrix Σ_ϵ is assumed to be diagonal, that any correlations found in the

observation vector $\mathbf{Y}_i$ are due to their relationship with common $\mathbf{b}_i$ and not due to some spurious correlation between ϵ_i. A continuous latent variable model is assumed for the relationship between $\mathbf{R}_i$ and u_i (model A in Figure 4) with $\pi_{ij} = Pr(r_{ij} = 1)$ representing the probability that the response for subject i at time point j is missing. We apply the logit link for the probability of the missingness, i.e., $log(\frac{\pi_{ij}(u_i,\, \mathbf{X}_{2i})}{1-\pi_{ij}(u_i,\, \mathbf{X}_{2i})}) = u_i - \tau_j \equiv X_{2i}\alpha + Z_{2i}u_i$, where τ_j are unknown parameters for determining an observation at time point j is missing. As discussed earlier, this relationship is equivalent to a random logistic regression, with appropriate design matrices $\mathbf{X}_{2i}$ and Z_{2i}. A latent variable regression, $\mathbf{b}_i = \mathbf{X}_{3i}^{'}\gamma + \zeta_i$, is used to establish the relationship between latent variable $\mathbf{b}_i$ and u_i, where $\mathbf{X}_{3i}^{'} = [\mathbf{X}_{3i} \;\; u_i]$ is a $p_3 + 1$ dimensional vector combining $\mathbf{X}_{3i}$ and u_i, γ is the $(p_3 + 1) \times q$ unknown regression coefficients for $\mathbf{X}_{3i}^{'}$ and the $q \times q$ matrix Ψ determines variance-covariance structure for error term ζ_i. Finally the latent continuous variable u_i is assumed to be normally distributed with mean 0 and variance σ_u^2.

Note that the maximum likelihood (ML) estimation of the model (2) - (4) requires the maximization of the observed likelihood, after integrating out missing data $\mathbf{Y}^{mis}$ and latent variables $\mathbf{b}$ and $\mathbf{u}$ from complete-data likelihood function. Detail of the ML estimation technique will be given in next section.

2.2 Maximum Likelihood Estimation

The main objective of this section is to obtain the ML estimate of parameters in the model and standard errors on the basis of the observed data $\mathbf{Y^{obs}}$ and $\mathbf{R}$. The ML approach is an important statistical procedure which has many optimal properties such as consistency, efficiency, etc. Furthermore, it is also the foundation of many important statistical methods, for instance, the likelihood ratio test, statistical diagnostics such as Cook's distance and local influence analysis, among others. To perform ML estimation, the computational difficulty arises because of the need to integrate over continuous latent factor $\mathbf{u}$, random subject-level effects $\mathbf{b}$, as well as missing responses $\mathbf{Y}^{mis}$. The classic Expectation-Maximization (EM) algorithm provides a tool for obtaining maximum likelihood estimates under models that yield intractable likelihood equations. The EM algorithm is an iterative routine requiring two steps in each iteration: computation of a particular conditional expectation of the log-likelihood (E-step) and maximization of this expectation over the parameters of interest (M-step). In our situations, in addition to the real missing data $\mathbf{Y}^{mis}$, we will treat the latent variables $\mathbf{b}$ and $\mathbf{u}$ as missing data. However, due to the complexities associated with the missing data structure and the nonlinearity part of the model (model A in Figure 1), the E-step of the algorithm, which involves the computations of high-dimensional complicated integrals induced by the conditional expectations, is intractable. To solve this difficulty, we propose to approximate the conditional expectations by sample means of the observations simulated from the appropriate conditional distributions, which is known as Monte Carlo Expectation Maximization algorithm. We will develop a hybrid algorithm that combines two advanced computational tools in statistics, namely the Gibbs sampler (Geman and Geman, 1984) and the Metropolis Hastings (MH) algorithm (Hastings, 1970) for simulating the observations. The M-step does not require intensive computations due to the distinctness of parameters in the proposed model. Hense, the proposed algorithm is a Monte Carlo EM (MCEM) type algorithm (Wei and Tanner, 1990). The description of the observed likelihood function is given in the following.

Given the parametric model (2) - (4) and the i.i.d. $J \times 1$ variables $\mathbf{Y}_i$ and $\mathbf{R}_i$, for $i = 1, \ldots, n$, estimation of the model parameters can proceed via the maximum likelihood method. Let $\mathbf{W}_i = (\mathbf{Y}_i^{obs}, \mathbf{R}_i)$ be the observed quantities, $\mathbf{d}_i = (\mathbf{Y}_i^{mis}, \mathbf{b}_i, u_i)$ be the missing quantities, and $\theta = (\alpha, \beta, \tau_j, \gamma, \Psi, \sigma_u^2, \Sigma_\epsilon)$ be the vector of parameters relating $\mathbf{W}_i$

with $\mathbf{d}_i$ and covariates $\mathbf{X}_i$. With Birch's regularity conditions (Birch, 1964) for parameter vector θ, the observed likelihood function for the model (2) - (4) can be written as

$$L_o(\theta|\mathbf{Y^{obs}}, \mathbf{R}) = \prod_{i=1}^{n} f(\mathbf{W}_i|\mathbf{X}; \theta) = \prod_{i=1}^{n} \int f(\mathbf{W}_i, \mathbf{d}_i|\mathbf{X}_i; \theta) d\mathbf{d}_i \tag{6}$$

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

$$\begin{aligned}
&L_o(\theta|\mathbf{Y^{obs}}, \mathbf{R}) = \prod_{i=1}^{n} \\
&\iiint \frac{1}{\sqrt{2\pi}}|\Sigma_\epsilon|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i)^T\Sigma_\epsilon^{-1}(\mathbf{Y}_i^{com} - \mathbf{X}_{1i}\beta - \mathbf{Z}_{1i}\mathbf{b}_i)\right\} \\
&\frac{1}{\sqrt{2\pi}}|\Sigma_b|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{b}_i - \mathbf{X}_{3i}'\gamma)^T\Sigma_b^{-1}(\mathbf{b}_i - \mathbf{X}_{3i}'\gamma)\right\} \frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left\{-\frac{u_i^2}{2\sigma_u^2}\right\} \\
&\left\{\prod_{j=1}^{J}\left(\frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)}\right)^{r_{ij}}\left(1 - \frac{exp(X_{2i}\alpha + Z_{2i}u_i)}{1 + exp(X_{2i}\alpha + Z_{2i}u_i)}\right)^{1-r_{ij}}\right\} du_i d\mathbf{b}_i d\mathbf{Y}_i^{mis}
\end{aligned} \tag{7}$$

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

In order to obtain valid ML estimates, one needs to investigate the convergence of the EM algorithm. However, in our case, determining the convergence of the MCEM algorithm is not straightforward. Meng and Schilling (1996) pointed out that the log-likelihood function can 'zigzag' along the iterates even without implementation or numerical errors, due to the variability introduced by simulation at the E-step. Further to evaluate the observed-data log-likelihood function, some numerical method has to be used because of a closed form is lacking. In the absence of accurate evaluation of the observed-data log-likelihood function, we could not judge whether any large fluctuation is due to the implementation errors, to the numerical errors in computing the log-likelihood values, or to non-convergence of the MCEM algorithm. We will implement bridge sampling to solve this problem, as suggested by Meng and Schilling (1996). Standard error estimates of the ML estimates can be obtained by applying the formula of Louis. (Louis, 1982)

3. Applications

3.1 Simulation Studies

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

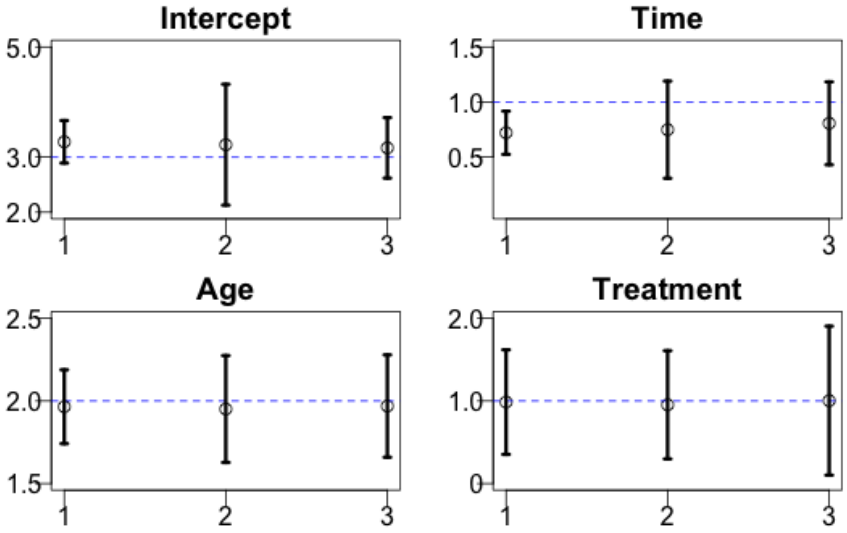

Figure 2: Point estimates and confidence interval (credible interval for Bayesian estimates) for fixed effects from simulated repeated-measure model with higher missing probability. The study sample size is 80. The true values are indicated by the dashed lines. 1, ignorable model; 2, Roy's model; 3,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

Figures 2 and 3 summarize point estimates and standard errors for both fixed and random effects. As expected, CLFM produced the best results for both cases, in terms of MSE. More specifically, one can observe that the ignorable likelihood approach tends to underestimate fixed intercept and slope in the model; furthermore 95 percent confidence intervals obtained from this approach do not cover the true values.

In Figure 3, true values of variance components in random effects, including $\sigma^2_{b_0}$, $\sigma^2_{b_1}$ and $\sigma_{b_0 b_1}$, are labeled as blue dotted lines, and red lines represent the non-significant level. Based on these plots, the variance components are indicated to be non-significant from the ignorable likelihood approach and Roy's model, but CLFM shows the correct result. In summary, CLFM can correct bias and generate efficient estimators when missing values are not ignorable in a study that contains substantial of missingness.

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

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

3.2.1 Description of Study

Patients in an AIDS Clinical Trial Group (ACTG) Study 193 A were randomized to dual or triple combinations of HIV-1 reverse transcriptase inhibitors. Specifically, HIV patients were randomized to one of four daily regimens containing 600 mg of zidovudine: zidovudine plus 2.25 mg of zalcitabine; zidovudine plus 400 mg of didanosine; zidovudine alternating monthly with 400 mg didanosine; or zidovudine plus 400 mg of didanosine plus

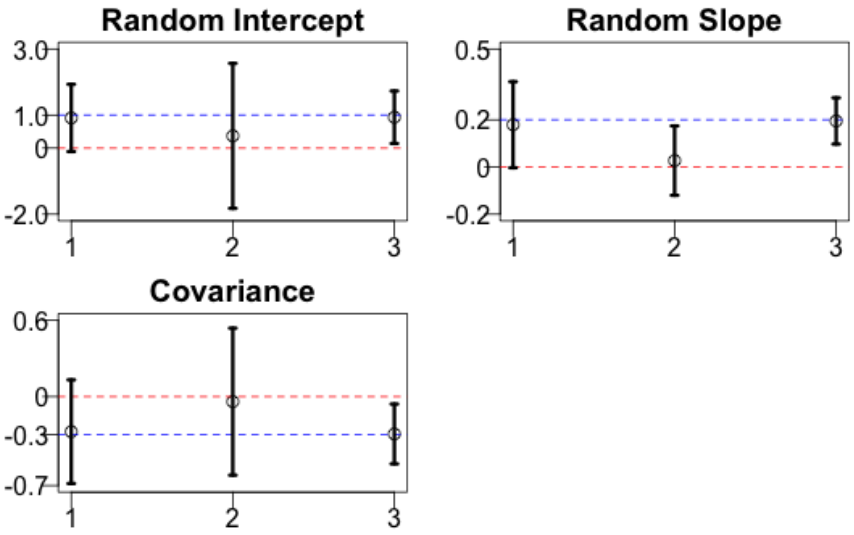

Figure 3: Point estimates and confidence interval (credible interval for Bayesian estimates) for random effects from simulated repeated-measure model with higher missing probability. The study sample size is 80. The true values are indicated by the blue dashed lines, the non-significant level is indicated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

400 mg of nevirapine (triple therapy). In this study, we focus on the comparison of the first three treatment regimens (dual therapy) with the forth (triple therapy)as described in Fitzmaurice's work. (Fitzmaurice and Laird, 2004)

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to unequal measurements and also CD4 counts have missing data that were caused by skipped visits and dropout. Table 1 presents four randomly selected subjects. The number of measurements of CD4 counts during the first 40 weeks of follow-up varied from 1 to 9, with a median of 4, based on the available data. The goal in this study is to compare the dual and triple therapy groups in terms of short-term changes in CD4 counts from baseline to week 40. The responses of interest are based on log transformation CD4 counts, log(CD4 counts + 1), available on 1309 patients.

Figure 4 describes the trend in the mean response in the dual and triple therapy groups via lowess smoothed curves on observed data. The curves reveal a modest decline in the mean response during the first 16 weeks for the dual therapy group, followed by a steeper decline from week 16 to week 40. By comparison, the mean response increases during the first 16 weeks and declines after for the triple therapy group. The rate of decline from week 16 to week 40 appears to be similar for the two groups. However, one has to notice that there is a substantial amount of missing data in the study, therefore the plot of the mean response over time can be potentially misleading, unless the data are missing completely at random (MCAR). Moreover, based on a small random sample of individuals, we observed that those with drop-out tend to have large CD4 counts. In other words, there is a trend that a patient in the study tended to skip a visit due to a large magnitude of current CD4 count. That is, a patient tends to skip a visit because of no treatment benefits or side effects. When

Table 1: Data example on log CD4 counts for four randomly selected subjects from ACTG study 193A

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

data are missing due to this reason, a plot of the mean response over time can be deceptive. Figure 5 describes observed responses at different visit points in each group. Almost all patients from both groups are treated at baseline and their CD4 count data are collected. There are two sharp decreases in response rate, one is from week 0 to week 8 and the other is from week 32 to week 40. Approaching the end of the study, most patients are dropping out from study, and response rates at week 40 are close to 20 percent for both treatments. The missing information can substantially influence the analysis and even bias findings. In the example, we will implement CLFM which assumes missing data are not ignorable, and compare with the conventional model that ignores missingness.

In the following we describe a model for the mean response that enables the rates of change before and after week 16 to differ within and between groups, and this model was also been adopted by Fitzmaurice and Laird (2004) in their work. Specifically, one could assume that each patient has a piecewise linear spline with a knot at week 16. That is, the response trajectory of each patient can be described with an intercept and two slopes–one slope for the changes in response before week 16, another slope for the changes in response after week 16. Further, we assume the average slopes for changes in response before and after week 16 are allowed to vary by group. Because this is a randomized study, the mean response at baseline is assumed to be the same in the two groups, as supported by Figure 4. Hence instead of the conventional growth curve model, we applied a special growth curve model to capture changing trends of responses on CD4 counts.

3.2.2 Model Specification

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

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

where $Group_i = 1$ if the ith subject is randomized to triple therapy, and $Group_i = 0$ otherwise; $(t_{ij} - 16)_+ = t_{ij} - 16$ if $t_{ij} > 16$ and $(t_{ij} - 16)_+ = 0$ if $t_{ij} \leq 16$; b_{1i}, b_{2i} and

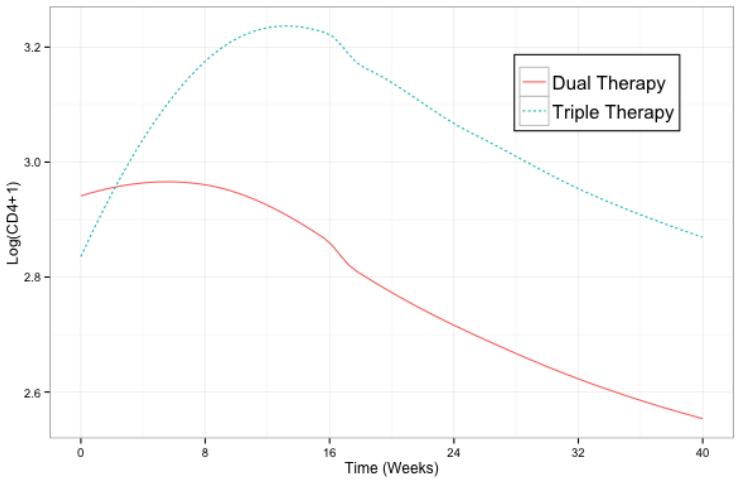

Figure 4: Lowess smoothed curves of log(CD4 + 1) against time (in weeks), for subject in the dual and triple therapy groups in ACTG study 193A

b_{3i} are random effects in this splined growth curve model. In this model, $(\beta_1 + b_{1i})$ is the intercept for the ith subject and has an interpretation as the true log CD4 count as baseline, i.e. when $t_{ij} = 0$. Similarly, $\beta_2 + b_{2i}$ is the ith subject's slope, or rate of change in log CD4 counts from baseline to week 16, if this patient is randomized to dual therapy; $(\beta_2 + \beta_4 + b_{2i})$ is the ith subject's slope if randomized to triple therapy. Finally, the ith subject's slope from week 16 to week 40 is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$ if randomized to dual therapy and $\{(\beta_2 + \beta_3 + \beta_4 + \beta_5) + (b_{2i} + b_{3i})\}$ if randomized to triple therapy. The model described above will be fitted without incorporating missing data. In order to fit CLFM, one has to specify the model for the missing part. Assume that $\mathbf{R}$ is a missing indicator matrix where its (i, j)th element $r_{ij} = 1$ if Y_{ij} is missing and $r_{ij} = 0$ if it is observed. Within a framework of CLFM, we incorporate information on missing values through modeling the missing information matrix $\mathbf{R}$ with time location parameters, and a continuous latent factor $\mathbf{u}$. Further, there are strong indications which support a application of this model. Based on Figure 5 one can see that the response variable tends to be missing over time. In other words, time locations are good indicators for explaining missing data. From Figure 5 one might also notice that the two therapies have identical missing proportions which suggests a group effect for therapies is not necessary in modeling $\mathbf{R}$. The continuous latent factor $\mathbf{u}$ is used to describe individuals' variability in missingness, and two regression parameters γ_1 and γ_2 are specified to provide information on random intercept $\mathbf{b}_0$ and slope $\mathbf{b}_1$, in order to correct estimation bias. A third regression parameter was also explored which links $\mathbf{u}$ with $\mathbf{b}_3$, but analysis results showed that this parameter is not significant. Hence we exclude this parameter in the final results. To estimate CLFM, we adopt two approaches: MECM to obtain ML estimates and full Bayesian estimates with specified conjugate priors. Point estimates and corresponding standard errors from a Bayesian perspective are summarized by posterior mean and standard deviation. Roy's model is also implemented by summarized missing patterns from $\mathbf{R}$ into three latent classes. (The number of latent classes for Roy's model is determined by information criteria)

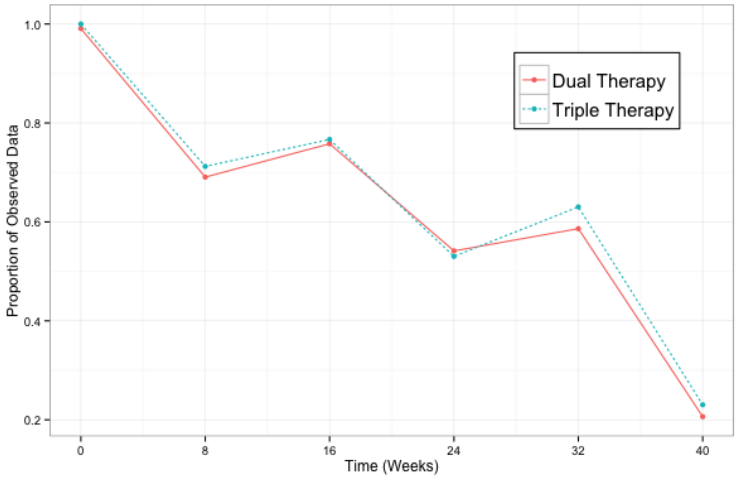

Figure 5: Proportions of observed responses in the dual and triple therapy groups in ACTG study 193A

3.2.3 Summary of Analyses under MAR and MNAR

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

baseline to week 16, subjects with higher log CD4 counts tend to be missing. CLFM involves non-ignorable missing data in the analysis, and the average of log CD4 counts tend to recover to a higher value. Hence, the decrease in the mean of the log CD4 counts from baseline to week 16 is not significant, when non-ignorable missing data are considered. By comparison, in the triple therapy group, there is a significant increase in the mean response. Based on the ignorable approach, the estimated change during the first 16 weeks in the triple therapy group is 0.31, $(16 \times (-0.0073 + 0.0269))$; the estimated slope for the triple therapy group is 0.0196 with a standard error 0.0033. In terms of the untransformed scale, it corresponds to an approximate 35 percent increase in CD4 counts. In CLFM, a similar estimate is obtained: the corresponding estimated change is 0.36. $(16 \times (-0.0047 + 0.0273))$; the estimated slope for the triple therapy group is 0.0226, and it corresponds to an approximate 40 percent increase in CD4 counts.

The loess curves in Figure 4 suggest that the rate of decline from week 16 to week 40 is similar for the two groups. The null hypothesis of no treatment group difference in the rates of change in log CD4 counts from week 16 to week 40 can be expressed as $H_0 : \beta_4 + \beta_5 = 0$. The estimates of β_4 and β_5 from all approaches appear to support the null hypothesis since they are of similar magnitude but with opposite signs. In the work of Fitzmaurice and Laird (2004), a test of the null hypothesis, $H_0 : \beta_4 + \beta_5 = 0$, is given and a Wald statistic is yielded with $W^2 = 0.07$, with 1 degree of freedom. The corresponding p value is greater than 0.75 based on the ignorable likelihood approach. DIC comparison for the Bayesian version of CLFM also suggests that two groups have similar rate of decline from week 16 to week 40. The Wald tests for Roy's model and MCEM version of CLFM further indicate this parallel change profiles after week 16, with both p-values are greater than 0.6.

The estimated variances of the random effects in Table 2 indicate that there is substantial individual variability in baseline CD4 counts and the rates of change in CD4 counts. For instance, in the triple therapy group, many patients show increases in CD4 counts during the first 16 weeks, but some patients have declining CD4 counts. Specifically, approximately 95 percent of patients are expected to have changes in log CD4 counts from baseline to week 16 between -0.64 and 1.27. Hence, approximately 26 percent of patients are expected to have decreases in CD4 counts during the first 16 weeks of triple therapy, based on the ignorable likelihood approach; by comparison, a larger variability from patient to patient is indicated by CLFM. 95 percent of patients are expected to have changes in log CD4 counts from baseline to week 16 between -1.15 and 1.87, and correspondingly approximately 30 percent of patients are expected to decrease CD4 counts from CLFM. Substantial components of variability due to measurement error are also suggested from all models.

In this study, missing data are potentially not ignorable with analyzing a random selected subsample, especially for the first 16 weeks. To evaluate effectiveness of treatment therapies, we compared three approaches, including the ignorable model which assumes missing data are MAR, Roy's model that handles non-ignorable missing data from pattern-mixture perspective, and CLFM with NMAR assumption. Controversial results on change rates of log CD4 counts at dual therapy group during first 16 weeks were obtained; that is, ignorable suggested there is a significant decrease in log CD4 counts, whereas both Roy's model and CLFM indicated this decrease is not substantial. This disagreement is due to those potential non-ignorable missing values. However, all approaches supported that triple therapy has similar change rate on log CD4 counts from week 16 to week 40, compare with dual therapy group. Further, with incorporating missing values, efficacy for both therapy groups is shown to be more substantial from CLFM, which can be seen from the log CD4 counts at week 40. Compared with Roy's model, the proposed CLFM is more

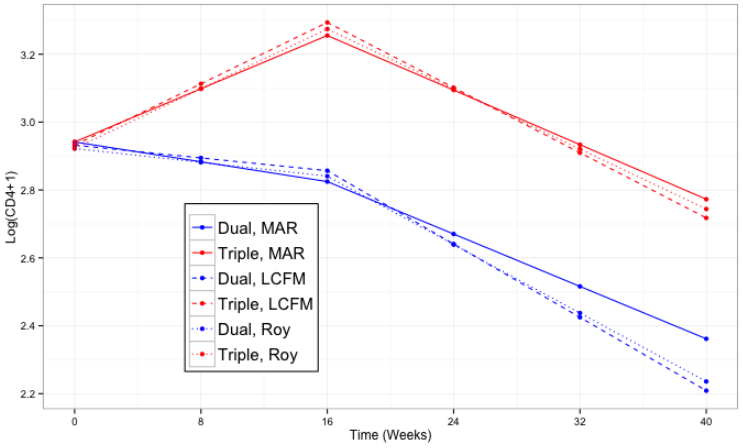

Figure 6: Fitted response curve in the dual and triple therapy groups in ACTG study 193A

flexible in extending the model with a more general distribution.

4. Conclusion and Discussion

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

We believe that this new approach will avoid untestable missing mechanism assumptions from selection models, and also believe that the new model will be more appealing to social behavioral and clinical researchers than pattern-mixture models, because the new model eliminates over-parameterizations issues. Further, the continuous latent factor provides an intuitive description of the response patterns in the study, and offers a feasible way to test conditional independence assumptions. For researchers who are interested in implementing CLFM model, we encourage them to compare latent factor models on missing indicator matrix with either constant slope or heterogeneous slopes and choose the one with better fitting in CLFM, based on information criteria or the likelihood ratio test. Lastly, CLFM is more feasible for small samples. With the truth that the underlying missing mechanism for missing data is unknown, (that is whether missingness is due to MAR or NMAR), we take this new method primarily as a tool for sensitivity analysis. In the case that a researcher cannot determine the distribution of missing data, the most responsible and objective approach to proceed is to explore and present alternative results from different plausible models.

In this paper, we have explored the proposed CLFM under the assumption of a multivariate normal distribution for the complete data. The normal model is an intuitive and natural starting point for this method, but it also has limitations. Many longitudinal studies will have discrete responses, such as measuring the total number of bleeding counts in a

Hemophilia study; or even binary responses. In the future, we will be extending our method to more flexible models for multivariate discrete responses. One promising approach is the Bayesian estimation approach which allows these extensions more straightforward.

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

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

Bibliography

M.W. Birch. A new proof of the pearson-fisher theorem. *Annals of Mathematical Statistics*, 35:818–824, 1964.

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

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

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

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

Wensheng Guo, Sarah J. Ratcliffe, and Thomas T. Ten Have. A random pattern-mixture model for longitudinal data with dropouts. *Journal of the American Statistical Association*, 99(468):pp. 929–937, 2004.

W.K. Hastings. Monte carlo sampling methods using markov chains and their application. *Biometrika*, 57:97–109, 1970.

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

Haiqun Lin, Charles E. McCulloch, and Robert A. Rosenheck. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, 60 (2):295–305, 2004.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 2002.

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

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

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

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

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

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

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

Table 2: Estimated regression coefficients (fixed effects) and variance components (random effects) for the log CD4 counts from a MAR model, Roy's model and CLFM in both approaches

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