The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 8

Simultaneous Bayesian Inference for Linear, Nonlinear and Semiparametric Mixed-Effects Models with Skew-Normality and Measurement Errors in Covariates

Yangxin Huang, *University of South Florida*

Ren Chen, *University of South Florida*

Getachew Dagne, *University of South Florida*

Recommended Citation:

Huang, Yangxin; Chen, Ren; and Dagne, Getachew (2011) "Simultaneous Bayesian Inference for Linear, Nonlinear and Semiparametric Mixed-Effects Models with Skew-Normality and Measurement Errors in Covariates," *The International Journal of Biostatistics*: Vol. 7 : Iss. 1, Article 8.

Available at: <http://www.bepress.com/ijb/vol7/iss1/8>

DOI: 10.2202/1557-4679.1292

©2011 Berkeley Electronic Press. All rights reserved.

Simultaneous Bayesian Inference for Linear, Nonlinear and Semiparametric Mixed-Effects Models with Skew-Normality and Measurement Errors in Covariates

Yangxin Huang, Ren Chen, and Getachew Dagne

Abstract

In recent years, various mixed-effects models have been suggested for estimating viral decay rates in HIV dynamic models for complex longitudinal data. Among those models are linear mixed-effects (LME), nonlinear mixed-effects (NLME), and semiparametric nonlinear mixed-effects (SNLME) models. However, a critical question is whether these models produce coherent estimates of viral decay rates, and if not, which model is appropriate and should be used in practice. In addition, one often assumes that a model random error is normally distributed, but the normality assumption may be unrealistic, particularly if the data exhibit skewness. Moreover, some covariates such as CD4 cell count may be often measured with substantial errors. This paper addresses these issues simultaneously by jointly modeling the response variable with skewness and a covariate process with measurement errors using a Bayesian approach to investigate how estimated parameters are changed or different under these three models. A real data set from an AIDS clinical trial study was used to illustrate the proposed models and methods. It was found that there was a significant incongruity in the estimated decay rates in viral loads based on the three mixed-effects models, suggesting that the decay rates estimated by using Bayesian LME or NLME joint models should be interpreted differently from those estimated by using Bayesian SNLME joint models. The findings also suggest that the Bayesian SNLME joint model is preferred to other models because an arbitrary data truncation is not necessary; and it is also shown that the models with a skew-normal distribution and/or measurement errors in covariate may achieve reliable results when the data exhibit skewness.

KEYWORDS: Bayesian analysis, covariate measurement errors, HIV dynamics, mixed-effects joint models, skew-normal distribution

Author Notes: The authors gratefully acknowledge the Editor and two anonymous referees for their insightful comments and helpful suggestions that led to a marked improvement of the article. We thank A5055 study investigators for allowing us to use the data from their study. This research was partially supported by NIAID/NIH grant R03 AI080338 and MSP/NSA grant H98230-09-1-0053 to Y. Huang.

1 Introduction

Modeling of HIV dynamics in AIDS research has greatly improved our understanding of the pathogenesis of HIV-1 infection and guided for the treatment of AIDS patients and evaluation of antiretroviral (ARV) therapies (Perelson, 1997; Wu and Ding, 1999). Such models often incorporate repeated measures over a period of treatment to assess rates of changes in viral load (number of HIV RNA copies in plasma). Recent research indicates that the first-phase decay rate of viral response to treatment is a potentially useful biomarker for ARV potency (Ding and Wu, 1999). Even though various statistical modeling and analysis methods have been applied for estimating the parameters of HIV dynamics such as linear and nonlinear regression (Perelson, 1997), linear mixed-effects (LME) and nonlinear mixed-effects (NLME) modeling approach (Wu and Ding, 1999; Wu, et al., 2004; Wu, 2004), nonparametric NLME modeling approach (Liu and Wu, 2007; Liang et al., 2003; Wu and Zhang, 2002; Wu, et al., 2004), joint model approach via Monte Carlo EM algorithm (Liu and Wu, 2007; Wu, 2002; Wu, 2004) and Bayesian NLME modeling approach via Markov Chain Monte Carlo (MCMC) procedure (Huang, 2006; Huang and Dagne, 2010), it is not clear which model should be used to estimate the first-phase decay rate. More importantly, in all of these models at least one of the following three important issues stand out.

Firstly, the common assumption of distributions for (within-subject) random error is normal in those statistical models. This assumption may lack the robustness against departures from normality and/or outliers as discussed by Verbeke and Lesaffre (1996), and may also lead to misleading results (Verbeke and Lesaffre, 1996; Ghosh et al., 2007). Very often in AIDS studies, the virologic responses exhibit skewness. For example, Figure 1 displays the histograms of repeated viral load (in ln scale) and CD4 cell count measurements for 44 subjects enrolled in the AIDS clinical trial study–A5055 (Acosta et al., 2004). It seems that for these data sets to be analyzed in this paper, the viral load responses are highly skewed even after ln-transformation. One way to incorporate skewness is to use a skew-normal (SN) distribution for (within-subject) random errors. Secondly, the mixed-effects models have been used in the previous studies to account for both between-subject and within-subject variations in viral load measurements which are associated with covariates including CD4 cell count. However, the covariates such as CD4 cell count which were considered in those studies are often measured with substantial errors and highly skewed as shown in Figure 1(bottom panel). Thirdly, a major challenge with these modeling approaches is the associated intensive computation burden in the inference, and in some cases it can even be computationally infeasible. Particularly, for nonlinear longitudinal models in the presence of measurement errors in covariates, the computational problem becomes much worse. In addi-

Figure 1: The histograms of viral load (in ln scale) and standardized CD4 cell count measured from day 0 to day 35 [Set (a): data cut at truncation day 35], day 84 [Set (b): data cut at truncation day 84] and the end of study period [Set (c): complete data] for 44 patients in an AIDS clinical trial study.

tion, there may exist the problem of non-convergence of these algorithms under the framework of likelihood estimation. To the best of our knowledge, there is little literature on simultaneously addressing measurement errors in covariates and a SN distribution for random errors to compare performance of the various mixed-effects models under the framework of Bayesian mixed-effects modeling approach. This article provides a unified approach to investigate SN Bayesian mixed-effects models with covariate measurement errors. It is noted that the random-effects could assume to have SN distribution. Huang and Dagne (2010) have studied NLME models where both model error and random-effects were assumed to have skew-normal distribution. It was found, in comparison of random-effects with normal and random-effects with SN, that the modeling results were very similar and not significantly different. Along with this finding, the random-effects are assumed to have a normal distribution in the mixed-effects models considered in this study.

In this paper, the main focus is to provide a comprehensive comparison of three mixed-effects models (LME, NLME and semiparametric NLME) with a SN distribution and measurement errors in covariates for estimated viral decay rates in viral dynamic models. We consider a multivariate SN distribution intro-

duced by Sahu et al.(2003) which is suitable for a Bayesian inference since it is built using conditional method and is defined in Appendix A. The rest of the paper is organized as follows. Section 2 presents a general modeling approach for SN Bayesian semiparametric nonlinear mixed-effects (SN-BSNLME) joint models which include three specific models as special cases to be discussed in Section 3. We describe data from an AIDS clinical study that motivated this research, discuss the specific models for HIV dynamics and reports the results obtained by using the three different methods or models. Finally, the paper concludes with some discussions in Section 4.

2 Bayesian inference on joint models with skew-normal distributions

2.1 Measurement error models with a skew-normal distribution

This section will briefly discuss measurement error joint models with a skew-normal distribution. Various covariate models were investigated in the literature (Carroll et al., 2006; Higgins et al., 1997; Liu and Wu, 2007; Wu, 2002). However, the fundamental assumption of distributions for measurement random errors is normal in these statistical covariate models and this assumption may lack the robustness against departures from normality and/or may violate the agreement with observed data. In this paper, we extend the covariate models to have a skew-normal distribution for measurement errors. For simplicity, we consider a single time-varying covariate. Let z_{ik} be the observed covariate value for individual i at time s_{ik} ($i = 1, \dots, n; k = 1, \dots, m_i$). Note that for each individual, we allow the covariate measurement times s_{ik} to differ from the response measurement times t_{ij} ($j = 1, \dots, n_i$). We consider the following LME covariate model with a SN distribution

$$z_{ik} = u_{ik}^T \alpha + v_{ik}^T a_i + \varepsilon_{ik} (\equiv z_{ik}^* + \varepsilon_{ik}), \quad \varepsilon_i \text{ iid} \sim SN_{m_i}(\mathbf{0}, \tau^2 I_{m_i}, \Delta(\delta_{\varepsilon_i})), \quad (1)$$

where $z_i = (z_{i1}, \dots, z_{im_i})^T$ with z_{ik} being the covariate value for individual i at time s_{ik} , u_{ik} and v_{ik} are $l \times 1$ design vectors, $\alpha = (\alpha_1, \dots, \alpha_l)^T$ and $a_i = (a_{i1}, \dots, a_{il})^T$ are unknown population (fixed-effects) and individual-specific (random-effects) parameter vectors, respectively, and $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})^T$ is a multivariate skew-normal distribution with ε_{ik} being the measurement error for individual i at time s_{ik} , τ^2 is the unknown within-individual variance. The $m_i \times m_i$ skewness diagonal matrix $\Delta(\delta_{\varepsilon_i}) = \text{diag}(\delta_{\varepsilon_{i1}}, \dots, \delta_{\varepsilon_{im_i}})$ and $m_i \times 1$ skewness parameter vector $\delta_{\varepsilon_i} = (\delta_{\varepsilon_{i1}}, \dots, \delta_{\varepsilon_{im_i}})^T$. In particular, if $\delta_{\varepsilon_{i1}} = \dots = \delta_{\varepsilon_{im_i}} \triangleq \delta_{\varepsilon}$, then $\Delta(\delta_{\varepsilon_i}) = \delta_{\varepsilon} I_{m_i}$ and $\delta_{\varepsilon_i} = \delta_{\varepsilon} \mathbf{1}_{m_i}$ with

$\mathbf{1}_{m_i} = (1, \dots, 1)^T$; this indicates that we are interested in skewness of overall data set which is the case to be used in real data analysis in the next section. $z_i^* = (z_{i1}^*, \dots, z_{im_i}^*)^T$ and $z_{ik}^* = u_{ik}^T \alpha + v_{ik}^T a_i$ may be viewed as the true (but unobservable) covariate values at time s_{ik} under normal distribution of model errors in which case skewness parameter $\delta_{e_{ik}} = 0$. We assume that $a_i \text{ iid } \sim N_l(0, \Sigma_a)$, where Σ_a is the unrestricted covariance matrix. Note that the model (1) may be interpreted as a skew-normal (SN) covariate measurement error model.

2.2 Skew-normal Bayesian semiparametric nonlinear mixed-effects joint models

In this section, we present the joint models and methods in general forms, illustrating that our methods may be applicable in other applications. Denote the number of subjects by n and the number of measurements on the i th subject by n_i . For the response process, we consider a general semiparametric NLME (SNLME) model incorporating possibly mismeasured time-varying covariates and model random errors with a skew-normal distribution.

$$\begin{aligned} y_{ij} &= g(t_{ij}, \beta_{ij}^\dagger, \phi(t_{ij})) + e_{ij}, \quad e_i \text{ iid } \sim SN_{n_i}(\mathbf{0}, \sigma^2 I_{n_i}, \Delta(\delta_{e_i})), \\ \beta_{ij}^\dagger &= d^\dagger[z_{ij}^*, \beta^\dagger, b_i^\dagger], \quad b_i^\dagger \text{ iid } \sim N_{s_3}(\mathbf{0}, \Sigma_b^\dagger), \\ \phi(t_{ij}) &= v[w(t_{ij}), h_i(t_{ij})], \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, n_i, \end{aligned} \quad (2)$$

where $y_i = (y_{i1}, \dots, y_{in_i})^T$ with y_{ij} being the response value for individual i at t_{ij} , $g(\cdot)$, $d^\dagger(\cdot)$ and $v(\cdot)$ are known parametric functions, $w(t)$ and $h_i(t)$ are unknown nonparametric smooth fixed-effects and random-effects functions, respectively, $h_i(t)$ are iid realizations of a zero-mean stochastic process, $\beta_{ij}^\dagger$ are $s_1 \times 1$ individual-specific time-dependent parameter vectors, $\beta^\dagger$ are $s_2 \times 1$ population parameter vectors ($s_2 \geq s_1$), σ^2 is the unknown within-subject variance, $e_i = (e_{i1}, \dots, e_{in_i})^T$ is the vector of random errors; $b_i^\dagger$ are $s_3 \times 1$ vector of random effects ($s_3 \leq s_1$) and $\Sigma_b^\dagger$ is the unrestricted covariance matrix. The $n_i \times n_i$ skewness diagonal matrix $\Delta(\delta_{e_i}) = \text{diag}(\delta_{e_{i1}}, \dots, \delta_{e_{in_i}})$ and the $n_i \times 1$ skewness parameter vector $\delta_{e_i} = (\delta_{e_{i1}}, \dots, \delta_{e_{in_i}})^T$. In particular, if $\delta_{e_{i1}} = \dots = \delta_{e_{in_i}} \triangleq \delta_e$, then $\Delta(\delta_{e_i}) = \delta_e I_{n_i}$ and $\delta_{e_i} = \delta_e \mathbf{1}_{n_i}$. In the model (2), we assume that the individual-specific parameters $\beta_{ij}^\dagger$ depend on the true (but unobservable) covariate z_{ij}^* rather than the observed covariate z_{ij} , which may be measured with errors.

Semiparametric NLME model (2) reverts to a parametric NLME model when the nonparametric parts $w(t)$ and $h_i(t)$ are constants. To fit model (2), we apply the regression spline method. The working principle is briefly described as follows and more details can be found in Wu and Zhang (2002). The main idea

of regression spline is to approximate $w(t)$ and $h_i(t)$ by using a linear combination of spline basis functions. For instance, $w(t)$ and $h_i(t)$ can be approximated by a linear combination of basis functions $\Psi_p(t) = \{\psi_0(t), \psi_1(t), \dots, \psi_{p-1}(t)\}^T$ and $\Phi_q(t) = \{\phi_0(t), \phi_1(t), \dots, \phi_{q-1}(t)\}^T$, respectively. That is,

$$w(t) \approx w_p(t) = \sum_{l=0}^{p-1} \mu_l \psi_l(t) = \Psi_p(t)^T \mu_p, \quad h_i(t) \approx h_{iq}(t) = \sum_{l=0}^{q-1} \xi_{il} \phi_l(t) = \Phi_q(t)^T \xi_{iq}, \quad (3)$$

where μ_p and ξ_{iq} ($q \leq p$) are the unknown vectors of fixed and random coefficients, respectively. For our model, we consider natural cubic spline bases with the percentile-based knots. To select an optimal degree of regression spline and numbers of knots, i.e., optimal sizes of p and q , the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) is often applied (Davidian and Giltinan, 1995; Wu and Zhang, 2002). Substituting $w(t)$ and $h_i(t)$ by their approximations $w_p(t)$ and $h_{iq}(t)$, we can approximate model (2) in a compact way as follows.

$$y_{ij} = g\left(t_{ij}, d^\dagger[z_{ij}^*, \beta^\dagger, b_i^\dagger], v[\Psi_p(t_{ij})^T \mu_p, \Phi_q(t_{ij})^T \xi_{iq}]\right) + e_{ij} \equiv g\left(t_{ij}, d[z_{ij}^*, \beta, b_i]\right) + e_{ij}, \quad (4)$$

where $\beta = (\beta^{\dagger T}, \mu_p^T)^T$ and $b_i = (b_i^{\dagger T}, \xi_{iq}^T)^T$ are the vectors of fixed-effects and random-effects, respectively, and $d(\cdot)$ is a known but possible nonlinear function. Thus, for given $\Psi_p(t)$ and $\Phi_q(t)$, we approximate the SN semiparametric NLME model (2) by the following SN parametric NLME model.

$$\begin{aligned} y_i &= g_i(t_{ij}, \beta_{ij}) + e_i, & e_i &\text{iid} \sim SN_{n_i}(\mathbf{0}, \sigma^2 I_{n_i}, \Delta(\delta_{e_i})), \\ \beta_{ij} &= d[z_{ij}^*, \beta, b_i], & b_i &\text{iid} \sim N_{s_4}(\mathbf{0}, \Sigma_b), \end{aligned} \quad (5)$$

where $s_4 = s_3 + q$, $g_i(t_{ij}, \beta_{ij}) = (g(t_{i1}, \beta_{i1}), \dots, g(t_{in_i}, \beta_{in_i}))^T$ with $g(\cdot)$ being a known linear or nonlinear function, and Σ_b is an unstructured covariance matrix.

Under Bayesian framework, we still need to specify prior distributions for unknown parameters in the models (1) and (5) as follows.

$$\begin{aligned} \alpha &\sim N_r(\alpha_0, \Lambda_1), \quad \tau^2 \sim IG(\omega_1, \omega_2), \quad \Sigma_a \sim IW(\Omega_1, \nu_1), \quad \delta_{\varepsilon_i} \sim N_{m_i}(\mathbf{0}, \Gamma_1), \\ \beta &\sim N_{s_5}(\beta_0, \Lambda_2), \quad \sigma^2 \sim IG(\omega_3, \omega_4), \quad \Sigma_b \sim IW(\Omega_2, \nu_2), \quad \delta_{e_i} \sim N_{n_i}(\mathbf{0}, \Gamma_2), \end{aligned} \quad (6)$$

where $s_5 = s_2 + p$, the mutually independent Inverse Gamma (IG), Normal (N) and Inverse Wishart (IW) prior distributions are chosen to facilitate computations (Davidian and Giltinan, 1995). The super-parameter matrices Λ_1 , Λ_2 , Ω_1 , Ω_2 , Γ_1 and Γ_2 can be assumed to be diagonal for convenient implementation.

We assume that e_i , ε_i , b_i and a_i are independent of each other. Following Sahu et al.(2003) and properties of skew-normal distribution, it can be shown that z_i and y_i in the models (1) and (5) follow the following distributions

$$\begin{aligned} y_i|a_i, b_i, w_{e_i}; \alpha, \beta, \sigma^2, \delta_{e_i} &\sim N_{n_i}(g_i + \Delta(\delta_{e_i})w_{e_i}, \sigma^2 I_{n_i}), \\ z_i|a_i, w_{\varepsilon_i}; \alpha, \tau^2, \delta_{\varepsilon_i} &\sim N_{m_i}(z_i^* + \Delta(\delta_{\varepsilon_i})w_{\varepsilon_i}, \tau^2 I_{m_i}), \\ w_{e_i} &\sim N_{n_i}(\mathbf{0}, I_{n_i})I(w_{e_i} > \mathbf{0}), \quad w_{\varepsilon_i} \sim N_{m_i}(\mathbf{0}, I_{m_i})I(w_{\varepsilon_i} > \mathbf{0}), \end{aligned} \quad (7)$$

where $I(w_k > 0)$ is an indicator function and $w_k = |\xi|$ with $\xi \sim N_k(0, I_k)$.

Let $\theta = \{\alpha, \beta, \tau^2, \sigma^2, \Sigma_a, \Sigma_b, \delta_{\varepsilon_i}, \delta_{e_i}; i = 1, \dots, n\}$ be the collection of unknown parameters in the models (1) and (5), and $f(\cdot|\cdot)$ and $\pi(\cdot)$ be a conditional density function and a prior density function, respectively. Denote the observed data by $\mathcal{D} = \{(y_i, z_i), i = 1, \dots, n\}$. We assume that $\alpha, \beta, \tau^2, \sigma^2, \Sigma_a, \Sigma_b, \delta_{\varepsilon_i}, \delta_{e_i}$ ($i = 1, \dots, n$) are independent of each other, and thus we have $\pi(\theta) = \pi(\alpha)\pi(\beta)\pi(\tau^2)\pi(\sigma^2)\pi(\Sigma_a)\pi(\Sigma_b)\prod_i \pi(\delta_{\varepsilon_i})\pi(\delta_{e_i})$. After we specify the models for the observed data and the prior distributions for the unknown model parameters, we can make statistical inference for the parameters based on their posterior distributions under Bayesian framework. The joint posterior density of θ based on the observed data can be given by

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \left\{ \prod_i^n \int \int f(y_i|a_i, b_i, w_{e_i}; \alpha, \beta, \sigma^2, \delta_{e_i}) f(w_{e_i}|w_{e_i} > \mathbf{0}) \times \right. \\ &\quad \left. f(z_i|a_i, w_{\varepsilon_i}; \alpha, \tau^2, \delta_{\varepsilon_i}) f(w_{\varepsilon_i}|w_{\varepsilon_i} > \mathbf{0}) f(a_i|\Sigma_a) f(b_i|\Sigma_b) da_i db_i \right\} \pi(\theta). \end{aligned} \quad (8)$$

In general, the integrals in (8) are of high dimension and do not have closed form. Analytic approximations to the integrals may not be sufficient accurate. Therefore, it is prohibitive to directly calculate the posterior distribution of θ based on the observed data. As an alternative, MCMC procedures can be used to sample based on (8) using the Gibbs sampler along with the Metropolis-Hasting (M-H) algorithm. The above representations based on the models are useful as it allows to implement easily using the WinBUGS codes (Lunn et al., 2000).

3 Analysis of AIDS clinical data

3.1 Data and models

We illustrate our methods using a real AIDS clinical data. The study consists of 44 HIV-infected patients who were treated with a potent ARV regimen. Viral load, CD4 count and other variables were repeatedly measured over a period of 24 weeks. RNA viral load was measured in copies/mL at study days 0, 7, 14, 28, 56, 84, 112, 140 and 168 of follow-up. The nucleic acid sequence-based amplification assay

(NASBA) was used to measure plasma HIV-1 RNA, with a lower limit of quantification, 50 copies/mL. the HIV-1 RNA measures below this limit are not considered reliable, therefore we simply imputed such values as 25 copies/mL (Acosta et al., 2004; Davidian and Giltinan, 1995). Covariates such as CD4 count including associated baseline data were also measured throughout the study on similar schemes. Figure 2 shows the measurements of viral load in natural ln scale and CD4 count for the three randomly selected patients. Both viral load and CD4 cell count trajectories exhibit distinctive and important patterns throughout the time course. We can see that the rate change in viral load appears to vary substantially across patients, reflecting both biological variation and systematic associations with subject-level covariates.

Figure 2: Profiles of viral load (response) in ln scale and CD4 cell count (covariate) for three randomly selected patients. The horizontal line is below the detectable level of viral load ($3.91 = \ln(50)$) and the two vertical lines represent truncation days 35 and 84, respectively.

As is evident from Figure 3(c), RNA levels varied widely during the early treatment stage. For some patients, the RNA level decreased rapidly with treatment (described as the first-phase decay rate); for others it decreased slowly. Studying the relation between baseline RNA and the first-phase decay rate will provide useful information for clinical decision-making and treatment individualization. The first-phase decay rate indicates the potency of ARV therapies (Ding and Wu, 1999). If we know the potency of a treatment at an earlier stage, we may be able to avoid the less potent regimens for a particular patient. This will help clinicians to select a treatment for their patients.

Viral dynamic models can be formulated through a system of ordinary differential equations (ODE) (Huang et al., 2006; Wu et al., 1998; Wu and Ding, 1999). In practice, some investigators have used a LME model simplified from an

Figure 3: Profiles of viral load in ln scale from an AIDS clinical trial study. Change in HIV-1 load, measured from RNA levels in plasma, with time during treatment from day 0 to (a) day 35, (b) day 84 and (c) the end of study period.

ODE system to fit viral load data from the very early time period such as displayed in Figure 3(a). Their response model can be described by the following linear model

$$y(t) = \ln\{V(t)\} = \beta_1 - \beta_2 t + e(t), \quad (9)$$

where $y(t)$ is the natural logarithm of the number of HIV-1 RNA copies per mL of plasma, $e(t)$ is the measurement error, β_2 is the first-phase viral decay rate which may represent the minimum turnover rate of productively infected cells (Perelson et al., 1997) and β_1 is macro-parameter with $\exp(\beta_1)$ being the baseline viral load at time $t = 0$.

Based on biological and clinical arguments, an effective model used to estimate viral dynamic parameters is the biphasic model approximated from a compartmental analysis-based an ODE system (Perelson et al., 1997; Wu and Ding, 1999). This model plays an important role in modeling HIV dynamics and has demonstrated promise in studying HIV response process. The model is described as follows.

$$y(t) = \ln\{V(t)\} = \ln\{\exp[\beta_1 - \beta_2 t] + \exp[\beta_3 - \beta_4 t]\} + e(t), \quad (10)$$

where β_4 is the second-phase viral decay rate which may represent the minimum turnover rate of latently or long-lived infected cells (Perelson et al., 1997). It is generally assumed that $\beta_2 > \beta_4$, which assures that the model is identifiable and is appropriate for empirical studies. It is of particular interest to estimate the viral decay rates β_2 and β_4 because they quantify the antiviral effect and hence can be used to assess the efficacy of the antiviral treatments (Ding and Wu, 1999). In estimating these decay rates, only the early segment of the viral load trajectory data (for example, data shown in Figure 3(b)) can be used (Perelson et al., 1997; Wu and

Ding, 1999), because the viral load trajectory may have a different shape in later stages (see Figure 3(c)).

Although the models (9) and (10) are widely used in HIV dynamic studies and have shown promise, there are some concerns. For example, when different models give different conclusions, how do we know which is correct and should be used? In our analysis of the clinical data shown above, the models (9) and (10) produce incongruous estimates of viral decay rates of β_2 and provide conflicting results on their correlations with baseline viral load. In addition, both models are applied to the early segment of the viral load data. That means one has to cut the data at some arbitrary time point. It is not obvious what time point should be the cut-point or whether we should use different cut-points. To avoid these problems, a semiparametric biexponential model was proposed as follows (Wu and Zhang, 2002).

$$y(t) = \ln\{V(t)\} = \ln\{\exp[\beta_1 - \beta_2 t] + \exp[\beta_3 - \beta_4(t)t]\} + e(t), \quad (11)$$

where the second-phase decay rate $\beta_4(t)$ is a smooth unknown function. Intuitively, model (11) is more reasonable because it assumes that the viral decay rate in the second-phase can vary with time as a result of drug resistance, pharmacokinetics, medication adherence and other relevant clinical factors likely to affect changes in the viral load during the late stage of treatment. Therefore, all data obtained during ARV treatment can be used by fitting model (11). We also assume that $\beta_2 > \beta_4(t)$ for all time in order to guarantee that there is the first phase of curve decay. This is a semiparametric model because of the mechanistic structure (two-exponential) with constant parameters ($\beta_1, \beta_2, \beta_3$) and a time-varying parameter ($\beta_4(t)$) to capture the time-varying effects of the treatment and over a longer period. Actually, by including long-term viral load data, the estimate of β_2 may be more accurate and reasonable compared with those obtained in previous studies (Perelson et al., 1997; Wu and Ding, 1999) after excluding long-term viral load data for modeling and analysis by some *ad hoc* rules (that is, the screening and inclusion of viral load data are quite arbitrary). In the mean time, this model enjoys the flexibility of a semiparametric function for the second-phase decay rate $\beta_4(t)$. The estimate of $\beta_4(t)$ provides not only an approximate turnover rate over time of long-lived/latently infected cells at the early stage of treatment as the standard parametric model does, but also more importantly describes how it may change over a long treatment period as driven by, presumably, drug resistance, non-compliance and other clinical determinants. Most importantly, the semiparametric model is capable of modeling long-term viral load data of which the trajectory may vary substantially among different patients (Wu and Zhang, 2002). It is noted that the three different models are applied to different segments of the viral dynamic data. Therefore, the standard

goodness-of-fit or model selection methods cannot be used to identify the appropriate model.

To characterize skewness appeared often in viral loads and CD4 cell counts, we will develop SN Bayesian mixed-effects joint models in conjunction with the three dynamic response models and LME covariate model. To model the covariate CD4 process, we consider empirical polynomial LME models and choose the best model (1) with quadratic ($l = 3$) based on AIC/BIC model selection criteria as suggested by Liu and Wu (2007) and Wu (2002). We thus adapted the quadratic polynomial as the SN covariate model (1) with $u_{ik} = v_{ik} = (1, s_{ik}, s_{ik}^2)^T$ for the CD4 trajectory as follows.

$$z_{ik} = z_{ik}^* + \varepsilon_{ik}, \quad \varepsilon_i \sim SN_{m_i}(\mathbf{0}, \tau^2 I_{m_i}, \delta_\varepsilon I_{m_i}), \quad (12)$$

where $z_{ik}^* = (\alpha_1 + a_{i1}) + (\alpha_2 + a_{i2})s_{ik} + (\alpha_3 + a_{i3})s_{ik}^2$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ is population (fixed-effects) parameter vector, and $a_i = (a_{i1}, a_{i2}, a_{i3})^T$ is individual-specific (random-effects) vector with normal distribution $N_3(\mathbf{0}, \Sigma_a)$.

We employ the linear combinations of natural cubic splines with percentile-based knots to approximate the nonparametric functions $w(t)$ and $h_i(t)$. Following studies in (Liu and Wu, 2007; Wu and Zhang, 2002), we set $\psi_0(t) = \phi_0(t) \equiv 1$ and take the same natural cubic splines in the approximations (3) with $q \leq p$. The values of p and q are determined based on the AIC/BIC which suggest the following function for $w(t_{ij}) + h_i(t_{ij})$ with $p = 3$ and $q = 1$ in the model (14) below.

$$w(t_{ij}) + h_i(t_{ij}) \approx \mu_0 \psi_0(t_{ij}) + \mu_1 \psi_1(t_{ij}) + \mu_2 \psi_2(t_{ij}) + \xi_{i0}, \quad (13)$$

where $\mu_3 = (\mu_0, \mu_1, \mu_2)^T$ and $\xi_{i0} \equiv b_{i4}$. Special cases of the model (2), which are offered to jointly model HIV dynamics in the presence of CD4 covariate process with measurement errors described in the model (12), are discussed below.

Model I: SN Bayesian semiparametric nonlinear mixed-effects (SN-BSNLME) joint model (5) based on the semiparametric biexponential model (11) in conjunction with the SN covariate model (12) along with specified prior distributions (6) can be expressed as

$$\begin{aligned} y_{ij} &= \ln \{ \exp[\beta_{i1} - \beta_{i2}t_{ij}] + \exp[\beta_{i3} - \beta_{ij4}(t_{ij})t_{ij}] \} + e_{ij}, \\ e_i &\sim SN_{n_i}(\mathbf{0}, \sigma^2 I_{n_i}, \delta_e I_{n_i}), \\ \beta_{i1} &= \beta_1 + b_{i1}, \quad \beta_{i2} = \beta_2 + b_{i2}, \\ \beta_{i3} &= \beta_3 + b_{i3}, \quad \beta_{ij4}(t_{ij}) = \beta_4 + \beta_5 z_{ij}^* + w(t_{ij}) + h_i(t_{ij}), \end{aligned} \quad (14)$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \mu_3^T)^T$, $b_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})^T \sim N_4(\mathbf{0}, \Sigma_b)$. We can see that the SN-BSNLME joint model above reverts to a SN Bayesian nonlinear mixed-effects (SN-BNLME) models when the nonparametric parts $w(t)$ and $h_i(t)$ become

constants. More specifically, the SN-BNLME model reduces to a SN Bayesian linear mixed-effects (SN-BLME) model when the response function is linear. Thus, we next present the following two simplified mixed-effects models.

Model II: SN Bayesian Nonlinear Mixed-Effects (SN-BNLME) joint model based on the biexponential model (10) in conjunction with the SN covariate model (12) can be expressed as

$$\begin{aligned} y_{ij} &= \ln \{ \exp[\beta_{i1} - \beta_{i2}t_{ij}] + \exp[\beta_{i3} - \beta_{ij4}t_{ij}] \} + e_{ij}, \\ e_i &\sim SN_{n_i}(\mathbf{0}, \sigma^2 I_{n_i}, \delta_e I_{n_i}), \\ \beta_{i1} &= \beta_1 + b_{i1}, \quad \beta_{i2} = \beta_2 + b_{i2}, \\ \beta_{i3} &= \beta_3 + b_{i3}, \quad \beta_{ij4} = \beta_4 + \beta_5 z_{ij}^* + b_{i4}, \end{aligned} \quad (15)$$

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ and $b_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})^T \sim N_4(\mathbf{0}, \Sigma_b)$.

Model III: SN Bayesian Linear Mixed-Effects (SN-BLME) model based on the linear model (9) can be written as

$$\begin{aligned} y_{ij} &= \beta_{i1} - \beta_{i2}t_{ij} + e_{ij}, \quad e_i \sim SN_{n_i}(\mathbf{0}, \sigma^2 I_{n_i}, \delta_e I_{n_i}), \\ \beta_{i1} &= \beta_1 + b_{i1}, \quad \beta_{i2} = \beta_2 + b_{i2}, \end{aligned} \quad (16)$$

where $\beta = (\beta_1, \beta_2)^T$ and $b_i = (b_{i1}, b_{i2})^T \sim N_2(\mathbf{0}, \Sigma_b)$.

3.2 Results of analysis

In this section, we report the results of our analysis of the three data sets (mentioned in Figure 3) using SN-BLME, SN-BNLME and SN-BSNLME joint models, respectively. A natural $\ln$ transformation for viral load data was used in the analysis in order to stabilize the variation of measurement error and speed up estimation algorithm. To avoid very small (large) estimates which may be unstable, we standardize the time-varying covariate CD4 cell counts (each CD4 value is subtracted by mean 375.46 and divided by standard deviation 228.57) and rescale the original time t (in days) so that the time scale is between 0 and 1. To fit the SN-BLME model and the SN-BNLME joint model, we included only the viral load data from day 0 to day 35 (Figure 3(a)) and day 0 to day 84 (Figure 3(b)), respectively, because the SN-BLME model could only be used to fit linear trajectories of viral load and the SN-BNLME assumptions might be violated after long-term treatment if there are rebounds of viral load (i.e., we excluded the significant rebound data from the analysis). In fitting the SN-BSNLME joint model, we use all viral load data shown in Figure 3(c).

To carry out Bayesian inference, we need to specify the values of the hyperparameters in the prior distributions. In Bayesian approach, we only need to specify

the priors at the population level. We take weakly informative prior distributions for all the parameters. In particular, (i) fixed-effects were taken to be independent normal distribution $N(0, 100)$ for each component of the population parameter vectors α and β . (ii) For the scale parameters σ^2 and τ^2 we assume a limiting non-informative inverse gamma prior distribution, $IG(0.01, 0.01)$ so that the distribution has mean 1 and variance 100. (iii) The priors for the variance-covariance matrices of the random-effects Σ_a and Σ_b are taken to be inverse Wishart distributions $IW(\Omega_1, \nu_1)$ and $IW(\Omega_2, \nu_2)$, where degree of freedom $\nu_1 = \nu_2 = 5$, and Ω_1 and Ω_2 are diagonal matrices with diagonal elements being 0.01. (iv) For each of the skewness parameters δ_e and δ_ε , we choose independent normal distribution $N(0, 100)$, where we specify that $\delta_{e_i} = \delta_e \mathbf{1}_{n_i}$ and $\delta_{\varepsilon_i} = \delta_\varepsilon \mathbf{1}_{m_i}$ to indicate that we are interested in skewness of both overall viral load data and overall CD4 cell count data.

The MCMC sampler was implemented using WinBUGS software (Lunn et al., 2000), and the program codes are available in Appendix B. In particular, the MCMC scheme for drawing samples from the posterior distributions of all parameters in the both response and covariate models is obtained by iterating between the following two steps: (i) Gibbs sampler is used to update $\alpha, \beta, \tau^2, \sigma^2, \Sigma_a, \Sigma_b, \delta_e, \delta_\varepsilon$; (ii) we update b_i and a_i ($i = 1, 2, \dots, n$) using the Metropolis-Hastings (M-H) algorithm. After collecting the final MCMC samples, we are able to draw statistical inference for the unknown parameters. Specifically, we are interested in the posterior means and quantiles. See the articles (Huang et al., 2006; Lunn et al., 2000) for detailed discussions of the Bayesian modeling approach and the implementation of the MCMC procedures, including the choice of the hyper-parameters, the iterative MCMC algorithm, the choice of proposal density related to M-H sampling, sensitivity analysis and convergence diagnostics. When the MCMC implementation is applied to the actual clinical data, convergence of the generated samples is assessed using standard tools within WinBUGS software (such as trace plots). After convergence was achieved, one long chain was run which may be more efficient with the following considerations. We propose that, after an initial number of 50,000 burn-in iterations, every 40th MCMC sample is retained from the next 400,000. Thus we obtain 10,000 samples of targeted posterior distributions of the unknown parameters for statistical inference.

We will investigate the following two scenarios. Firstly, As shown in Figure 1, the histograms of viral load and CD4 cell count clearly indicate their asymmetric nature and it seems adequate fitting a SN model to the data set. Since a normal distribution is a special case of a SN distribution when skewness parameter is zero, we will investigate how a SN distribution for random error contributes to modeling results and parameter estimation in comparison with a normal distribution for random error. Secondly, we also estimate the model parameters by using the

‘naive’ method, which does not separate the measurement errors from the true CD4 values (i.e., completely ignores measurement errors in CD4 values in the modeling). That is, the ‘naive’ method only uses the observed CD4 values z_{ij} rather than true (unobservable) CD4 values z_{ij}^* in the response Models I–II in which case the joint models revert to regular models without covariate models involved for inference. We use the ‘naive’ method as a comparison to the joint modeling method proposed in Section 2. This comparison attempts to investigate how the measurement errors in CD4 contribute to parameter estimation.

3.2.1 Comparison of results between models with normal and SN distributions

In this section, we investigate how a SN distribution contributes to modeling results in comparison with a normal distribution for random error. The population posterior mean (PM), the corresponding standard deviation (SD) and 95% credible interval (CI) for fixed-effect parameters based on the models with a SN or normal random error are presented in Table 1. The following findings are observed based on the estimated results. (i) Firstly, in the response models for the most interested parameters $(\beta_2, \beta_4, \beta_5)$, β_2 based on the three models with a normal random error are larger than the corresponding that based on the three models with a SN random error; all the estimates are statistically significant since the 95% CIs do not contain zero. Secondly, for (β_4, β_5) , it can be seen that the estimates based on the SN-BNLME and SN-BSNLME models are substantially different from those based on the N-BNLME and N-BSNLME models. Thirdly, for the variance parameter σ^2 , its estimated values (0.04, 0.08, 0.09) based on the three models with a SN random error are much smaller than those (0.96, 0.54, 1.31) based on the three models with a normal random error. Finally, for the skewness parameter δ_e , we found that δ_e associated with the three models with a SN random error is estimated to be significantly positive; this confirms the positive skewness of the viral load data in ln scale as shown in Figure 1, and the estimates of the skewness parameter δ_e based on SN-BLME model (1.57), SN-BNLME model (1.15) and SN-BSNLME model (2.03) are fairly high. (ii) For estimated parameters in the CD4 covariate models, the estimates of intercept α_1 based on the N-BNLME and N-BSNLME models are larger than those based on SN-BNLME and SN-BSNLME models, however the estimates of coefficients α_2 and α_3 are very similar between SN and normal models. For the variance parameter τ^2 , the estimated values (0.10, 0.13) based on the N-BNLME and N-BSNLME models are larger than those (0.01, 0.08) based on SN-BNLME and SN-BSNLME models. The estimates of the skewness parameter δ_e

based on SN-BNLME and SN-BSNLME models are significantly positive which is consistent with positive skewness of the CD4 cell count data as shown in Figure 1.

Table 1: A summary of the estimated posterior mean (PM) of interested population (fixed-effects) and precision parameters, the corresponding standard deviation (SD) and lower limit (L_{CI}) and upper limit (U_{CI}) of 95% equal-tail credible interval (CI) from the normal (N) and SN distributions of random error based on the joint modeling(JM) approach and the naive method (NM) associated with the SN models.

Model(Method)		α_1	α_2	α_3	β_1	β_2	β_3	β_4	β_5	τ^2	σ^2	δ_e	δ_e
N-BLME (JM)	PM	–	–	–	8.01	3.70	–	–	–	–	0.96	–	–
	L_{CI}	–	–	–	7.52	3.16	–	–	–	–	0.74	–	–
	U_{CI}	–	–	–	8.51	4.22	–	–	–	–	1.25	–	–
	SD	–	–	–	0.25	0.27	–	–	–	–	0.13	–	–
SN-BLME (JM)	PM	–	–	–	5.42	3.43	–	–	–	–	0.04	–	1.57
	L_{CI}	–	–	–	4.80	2.85	–	–	–	–	0.01	–	1.34
	U_{CI}	–	–	–	6.03	4.04	–	–	–	–	0.18	–	1.83
	SD	–	–	–	0.31	0.30	–	–	–	–	0.05	–	0.13
N-BNLME (JM)	PM	-0.20	0.68	-0.42	8.52	21.0	5.59	1.26	0.04	0.10	0.54	–	–
	L_{CI}	-0.47	0.15	-0.98	7.96	16.6	5.01	0.60	0.01	0.08	0.40	–	–
	U_{CI}	0.07	1.19	0.14	9.09	26.3	6.15	1.91	0.33	0.13	0.73	–	–
	SD	0.14	0.26	0.28	0.28	2.44	0.29	0.33	0.19	0.13	0.08	–	–
SN-BNLME (JM)	PM	-1.01	0.70	-0.50	6.70	20.9	3.82	1.35	0.02	0.01	0.08	0.53	1.15
	L_{CI}	-1.30	0.24	-0.99	5.97	16.6	3.03	0.64	0.01	0.01	0.01	0.38	0.79
	U_{CI}	-0.70	1.16	0.04	7.50	26.0	4.60	1.97	0.22	0.05	0.33	0.65	1.42
	SD	0.16	0.24	0.27	0.39	2.39	0.39	0.34	0.15	0.01	0.09	0.07	0.16
N-BSNLME (JM)	PM	-0.22	0.66	-0.28	8.29	26.6	3.66	-2.56	0.88	0.13	1.31	–	–
	L_{CI}	-0.48	0.13	-0.85	7.75	17.5	1.12	-10.6	0.31	0.10	1.06	–	–
	U_{CI}	0.03	1.20	0.29	8.83	36.2	5.27	2.90	1.77	0.15	1.62	–	–
	SD	0.13	0.27	0.29	0.28	4.77	1.02	3.33	0.45	0.01	0.14	–	–
SN-BSNLME (JM)	PM	-0.62	0.65	-0.27	4.60	17.1	-0.09	-4.44	0.15	0.08	0.09	0.25	2.03
	L_{CI}	-1.13	0.10	-0.85	3.76	9.97	-2.76	-10.9	0.06	0.04	0.01	0.16	1.64
	U_{CI}	0.37	1.19	0.31	5.63	23.0	1.50	-0.60	0.94	0.13	0.38	0.52	2.34
	SD	0.41	0.27	0.29	0.47	3.41	1.14	2.83	0.39	0.03	0.10	0.25	0.17
SN-BSNLME (NM)	PM	–	–	–	5.34	28.1	1.57	0.39	0.78	–	0.12	–	1.84
	L_{CI}	–	–	–	4.55	20.7	0.22	-3.98	0.17	–	0.01	–	1.45
	U_{CI}	–	–	–	6.18	36.7	2.73	3.57	1.37	–	0.43	–	2.18
	SD	–	–	–	0.41	3.99	0.64	1.92	0.32	–	0.12	–	0.18

Figure 4 displays the three randomly selected individual estimates of viral load trajectories along with the associated 95% CIs on each fitted value obtained based on the BLME (left), BNLME (center) and BSNLME (right) models with a Normal (dotted line) or SN (solid line) random error, respectively. The following findings are observed from joint modeling results. (i) The estimated individual trajectories for the models where the random error is assumed to be SN fit the originally observed values much closer than those for the corresponding models where the random error is assumed to be normal. (ii) Overall, the 95% CI associated with each of fitted values from the normal models is wider than that from the corresponding SN models. (iii) All the 95% CIs from three SN models cover the true (observed) viral load values, while some of 95% CIs from three normal models do

Figure 4: The individual estimates of viral load trajectories for three randomly selected patients based on the BLME (left), BNLME (center) and BSNLME (right) models with a normal (dotted line) or SN (solid line) random error. The respective vertical dotted line (normal) ended with ‘o’ and solid line (SN) ended with ‘•’ on each fitted value are the 95% credible interval (CI) associated with the fitted value. The observed values are indicated by sign crosses (×).

not. For example, for patient 39 whose observed value at day 115 is 10.57, the 95% CI from the SN-BSNLME model is (9.90,11.02) with the fitted value 10.51, while the corresponding 95% CI from the N-BSNLME model is (7.19, 9.53) with the fitted value 8.33 which does not cover the observed value 10.57.

We also investigate the model fitting results for each of the three mixed-effects models with normal and SN random errors, respectively. We have seen that, in general, all the models provided a reasonably good fit to the observed data for most patients in our study, although the fitting for a few patients was not completely satisfactory due to unusual viral load fluctuation patterns for these patients. To assess the goodness-of-fit of each of the three mixed-effects models with normal and SN random errors, the diagnosis plots of the observed values versus the fitted val-

Figure 5: The observed values versus fitted values of $\ln(\text{RNA})$ based on the BLME (left), BNLME (center) and BSNLME (right) models with a normal or SN random error.

ues are presented in Figure 5. It was seen from Figure 5 that the models where the random error is assumed to be SN provided better fit to observed data, compared with the models where the random error is assumed to be normal. This result can be also explained by examining the SN or normal Q-Q plots of the residuals (Figure 6) where all plots show the existence of outliers. From these plots, it can be seen that the SN models only have few negative outliers and thus fit observed data better than the corresponding normal models. This finding is further confirmed by their residual sum of squares (RSS). That is, for the BLME model the RSSs are 3.62 (SN random error) and 117.37 (normal random error); for the BNLME model the RSSs are 6.64 (SN random error) and 63.78 (normal random error); for the BSNLME model the RSSs are 7.49 (SN random error) and 312.58 (normal random error). As one of referees pointed out, we also checked the residual distributions by plotting densities of residuals for the three models with SN or normal distributions (the plots are not shown). Based on those plots, the residuals appear to be symmetrically distributed for the models with normal distribution, while the residuals follow a positive skew-distribution for the models with SN distribution, confirming the assumptions of the residual distributions.

Figure 6: SN or normal Q-Q plots of residuals with lines based on the BLME (left), BNLME (center) and BSNLME (right) models with a normal or SN random error.

For selecting the best model that fits the data adequately, a Bayesian selection criterion is used. This criterion, known as deviance information criterion (DIC), was first suggested in a recent publication by Spiegelhalter et al.(2002). As with other model selection criteria, we caution that DIC is not intended for identification of the ‘correct’ model, but rather merely as a method of comparing a collection of alternative formulations. In each of the three models with the specification of different distributions for the random error, DIC can be used to find out how assumption of a SN distribution contributes to virologic responses and parameter estimation in comparison with that of a normal distribution. We found that the DIC values, 335.14 (SN-BLME), 607.24 (SN-NLME) and 1051.21 (SN-SNLME) for the three models with a SN random error are smaller than the corresponding those, 524.54 (normal-BLME), 701.17 (normal-NLME) and 1355.45 (normal-SNLME) for the three models with a normal random error, respectively.

As mentioned before, it is hard to tell which model is ‘correct’ but which one fits data better. Therefore, based on the DIC criterion, the results indicate that each of the three models with a SN random error fits the data better, supporting the contention of a departure from normality. These results are consistent with those from diagnosis of the goodness-of-fit displayed in Figure 5. In summary, our results may suggest that it is very important to assume a SN distribution for the response models in order to achieve reliable results, in particular if data exhibit skewness.

Along with these observations, we further report our results in details only for the three models with a SN random error.

3.2.2 Results of analysis based on the SN models

The population estimates of the viral dynamic parameters presented in Table 1 based on the three (SN-BLME, SN-BNLME, SN-BSNLME) models indicate that the estimates of β_1 from the different models agree fairly well. However, the estimates (20.9 and 17.1) of the first decay rate β_2 by SN-BNLME and SN-BSNLME modeling methods are significantly different from that (3.43) obtained by SN-BLME modeling method. Although the estimates of β_2 by SN-BNLME and SN-BSNLME modeling methods are comparable, one problem is that we considered only 84-day data for SN-BNLME model fitting. This means that only 68% of the data from the 168-day period were included due to arbitrary truncation of data. Therefore, the SN-BNLME modeling may not be efficient. In this case, we prefer to use the SN-BSNLME model in which a smooth function of treatment time is incorporated into the second-phase decay rate to better catch rebound viral load data and all data measured during the treatment period can be used.

In the SN-BLME and SN-BNLME model fittings, individual curves (solid lines in Figure 4) for each subject follow a similar trend; that is, the trajectories of viral load decay in all 3 subjects decrease rapidly in the first-phase, then flatten in general. When the entire treatment period is considered, the viral loads of subject 39 rebound after the second-phase, whereas the viral loads of subjects 23 and 32 remain low until the end of the treatment period. Obviously, the SN-BLME and SN-BNLME model fittings are reasonable for data cutting at days 35 and 84, but they do not represent data measured over the whole treatment period.

It is also worth noting the differences among the estimated values of the first-phase decay rate (β_{i2}) for each subject based on the three models. We can see from Figure 7 (top panel) that the individual estimated values of β_{i2} obtained by the SN-BNLME model fitting are consistently much greater than those obtained by the SN-BLME model fitting, but are slightly different from those obtained by the SN-BSNLME model fitting. Although for each model the individual estimates of $\ln(\text{RNA})$ levels approximately follow the observed values, the differences in β_{i2} values obtained with each model may suggest a completely different relation between β_{i2} and baseline viral loads. When we investigate the relation between the estimated individual first-phase decay rates β_{i2} and baseline $\ln(\text{RNA})$ levels, the results are incongruous. The correlations between the subject-specific viral decay rates β_{i2} estimated by each method and baseline $\ln(\text{RNA})$ levels are shown in Figure 7 (bottom panel). The subject-specific estimates of β_{i2} obtained from the SN-

Figure 7: Correlations between baseline $\ln(\text{RNA})$ levels and the subject-specific first phase viral decay rates, β_{i2} estimated by using each of the three different methods. The solid lines are robust (MM-estimator) linear regression fit. The correlation coefficients (r) and p -values were obtained from Spearman rank correlation test.

BSNLME and SN-BLME modeling methods show significantly positive correlations ($r_I = 0.727$ and $r_{III} = 0.874$ with p -value $p < 0.0001$) with baseline $\ln(\text{RNA})$ levels. However, the estimates of β_{i2} obtained by using the SN-BNLME modeling method are negatively correlated with baseline $\ln(\text{RNA})$ levels ($r_{II} = -0.952$ with p -value $p < 0.0001$).

The incongruity in the individual estimates of the first-phase viral decay rates and their correlations with baseline $\ln(\text{RNA})$ levels, as determined by the SN-BLME, SN-BNLME and SN-BSNLME modeling methods, is significantly different with the following two observed scenarios: (i) Although the individual estimates of β_{i2} obtained by the both SN-BLME and SN-BSNLME modeling methods are positively correlated with baseline $\ln(\text{RNA})$ levels, the the individual estimates of β_{i2} from the SN-BSNLME method are, in general, five times larger than those from the SN-BLME method. (ii) The individual estimates of β_{i2} by SN-BNLME and SN-BSNLME modeling methods are fairly comparable, but the correlations between baseline $\ln(\text{RNA})$ levels and the subject-specific viral decay rates β_{i2} estimated by these two methods are completely opposite. These inconsistencies are presumably caused by data truncation. From the above results, it may suggest that the estimates obtained from the both SN-BLME and SN-BNLME models might be not reliable and the estimates based on the SN-BSNLME model may be favorable.

To fit the SN-BLME model, we truncated the data at day 35 in this study. However, it is not clear where one should cut the data between the first- and second-phases of decay. Also, different subjects may have different change points between the two phases. For example, truncation at day 35 may cause data from the second-phase to be included with first-phase data. It is for this reason that SN-BLME models underestimate the first-phase decay rates (β_2). The SN-BSNLME modeling method is preferable to the SN-BLME and SN-BNLME modeling methods, especially for sparse individual data. We believe that estimates obtained from the SN-BSNLME modeling method and their correlations with baseline $\ln(\text{RNA})$ levels are reliable since the complete data are used. Conversely, the both SN-BLME and SN-BNLME model fittings may result in misleading conclusions, as shown above, perhaps because it is impossible to find a truncation point to data for these two model fittings that would be suitable for all patients. The estimated values of β_2 would be affected by the inclusion of second-phase data if truncation occurred too late, and by the loss of data if truncation was too early.

For comparison, as an example, we also employed the ‘naive’ method-based the SN-BSNLME model to estimate the model parameters presented in Table 1 using the observed CD4 values and ignoring the CD4 measurement errors. It can be seen from the estimated results that the estimates of the parameters from the naive method are significantly larger than those from the joint modeling method. It indicates that the naive method may produce overestimated results with substantial biases; in particular, the estimated covariate CD4 effect (β_5) from the naive method is 5 times greater than that from the joint modeling method. The joint modeling method appears to give larger SD in most cases, probably because it incorporates the variation from fitting the CD4 covariate process. Further, the estimate of the model skewness parameter δ_e for the naive method is slightly smaller than that for the joint modeling method; this result suggests that the naive method may underestimate the skewness parameter due to ignoring measurement errors in CD4 values. Thus the difference of the naive estimates and the joint modeling estimates, due to whether or not ignoring potential CD4 measurement errors in conjunction with the SN-BSNLME model, indicates that it is important to take the measurement errors into account in the analysis when covariates are measured with errors.

4 Discussion

For viral dynamic models with skewness characteristics of viral load responses and CD4 measurement errors in covariate, we have investigated and compared the three Bayesian mixed-effects models with a SN distribution that may be preferred over those with a normal distribution in the sense that it produces less biased parameter

estimates and provides better fit to observed data. The proposed method may have a significant impact on AIDS research because, in the presence of skewness in the data, appropriate statistical inference for HIV dynamics is important for making robust conclusions and reliable clinical decisions. Our proposed method is quite general and so can be used to other applications. This kind of SN modeling approach is important in many biostatistical application areas, allowing accurate inference of parameters while adjusting for the data with skewness. The SN distribution is shown to provide an alternative to normal (symmetric) distribution that is often assumed in statistical models. The results indicate that with SN distribution assumption, there is potential to gain efficiency and accuracy in estimating certain parameters when the normality assumption does not hold in the data. The models considered in this paper can be easily fitted using MCMC procedure. Moreover, the proposed modeling approach is fitted using the WinBUGS package that is available publicly. This makes our approach quite powerful and accessible to practicing statisticians in the fields.

To estimate the viral dynamic parameters and study the relation of the baseline level of HIV-1 RNA in plasma with the decay rate of the first-phase of response to treatment, we compared the results of SN-BLME, SN-BNLME and SN-BSNLME model fittings, and found that the both SN-BLME and SN-BNLME model fittings in short-term dynamics may result in misleading conclusions due to data truncation. Of particular interest is that when long-term dynamics are considered, SN-BNLME may also become unreliable because of the complexity of the second-phase of decay. The foregoing results indicate that in a two-phase HIV dynamic model, the first decay rate may remain constant, while the second decay rate may change which depends on time-varying CD4 covariate during the period of study. The analysis results suggest that there may be a significantly positive correlation between the first-phase viral decay and the baseline HIV-1 RNA levels based on the SN-BSNLME modeling method. This finding is consistent with those reported in publications (Notermans et al., 1998; Wu et al., 2004). This positive correlation may be partially explained by the fact that the higher baseline viral load value, which is equivalent to the lower baseline CD4 value due to a negative relation between these two baseline factors, suggests a lower turnover rate of lymphocyte cells, which may lead to a positive correlation between the first-phase viral decay rate (β_{i2}) and the baseline HIV-1 RNA levels. Higher baseline HIV-1 RNA levels reflect more productively infected cells distributed at different sites; thus, greater drug potency or exposure may be required to achieve a similar decay rate to that seen in patients with lower levels of viral replication. This finding is very interesting and clinically important. Since the viral decay rates may reflect the efficacy of ARV treatment, the lower baseline HIV-1 RNA levels may need less potent drug efficacy to suppress virus replication so that a strong treatment strategy is not nec-

essary to avoid side-effect of drug use. This may help improve understanding of the pathogenesis of HIV infection and evaluation of ARV treatments.

We would point out that the problems we have addressed in this paper cannot be resolved by the standard goodness-of-fit or model selection methods, which are usually used when applying different models to the same data set. From Figure 4, we can see that all three models (BLME, BNLME and BSNLME) fitted the corresponding data very well. Goodness-of-fit or model selection methods are unable to identify the right model. We have found that the different models should not be applied to the same, entire data set, but applied simultaneously to appropriate segments of the data set such that the results will be biologically meaningful.

In conclusion, BLME fitting may be misleading and its use should be avoided; BNLME fitting may work well but is subject to data truncation; BSNLME fitting works in a similar way to BNLME fitting but has no data-screening problems associated with its use. Care is necessary in the implementation of BNLME and BSNLME fittings. With the introduction of SN distribution in the models, the estimated results suggest that the skewness parameters in viral load and CD4 cell count are estimated to be significantly positive for each of the three models. This confirms the positive skewness of the viral load and CD4 data presented in Figure 1. Thus, we may conclude that accounting for significant skewness is required when one models a data set which exhibits skewness.

This paper combines new technologies in mathematical modeling and statistical inference with advances in viral dynamics and ARV treatment to quantify complex HIV disease mechanisms. The complex nature of HIV/AIDS will naturally pose some challenges such as nonignorable missing data and data with detection limit problems. Furthermore, other more flexible distributions for model errors such as skew- t distributions (Sahu et al., 2003), normal/independent distributions (Lange and Sinsheimer, 1993) or nonparametric distributions such as an unspecified distribution that has a Dirichlet process prior (Ferguson, 1973) can be assumed. These complicated problems are beyond the purpose of this article, but a further study may be warranted. We are actively investigating these problems and associated simulation studies now. We hope that we could report these interesting results in the near future.

Appendices

A. Multivariate skew-normal distributions

Recently, there has been an increasing interest in finding more flexible methods to represent features of the data as adequately as possible and to reduce unrealistic assumptions. One approach for data modeling consists in constructing flexible

parametric classes of multivariate distributions that is different from the normal distribution. The skew-elliptical distribution is an attractive class of asymmetric thick-tailed parametric structure which includes the skew-normal (SN) distribution as a special case. Different versions of the multivariate SN distributions have been considered and used in the literature (Arellano-Valle et al., 2005; Azzalini and Capitanio, 1999; Arellano-Valle and Genton, 2005; Arellano-Valle and Azzalini, 2006; Azzalini and Dalla-Valle, 1996; Sahu et al., 2003 and among others). These studies demonstrated that the SN distribution has reasonable flexibility in real data fitting, while it maintains some convenient formal properties of the normal density. For more detailed discussions on properties and theories of SN distribution and its potential applications as well as differences among various versions of SN distributions, see References listed above.

In this paper, we consider a multivariate SN distribution introduced by Sahu et al.(2003) which is suitable for straightforward Bayesian analysis through hierarchical representations since it is built using conditional method. In particular, it is relatively easy to implement and provides an interesting alternative to other computationally challenging parametric or nonparametric models (Ferguson, 1973; Lange and Sinsheimer, 1993; Müller and Rosner, 1997). For completeness, this section is started by briefly summarizing the multivariate SN distribution that will be used in defining the SN joint models considered in this paper.

Figure 8: The univariate skew-normal density functions with skewness parameter $\delta = 0, -5, 5$, respectively.

An m -dimensional random vector Y follows an m variate SN distribution with location vector μ , $m \times m$ positive (diagonal) dispersion matrix Σ and $m \times m$ skewness matrix $\Delta(\delta) = \text{diag}(\delta_1, \delta_2, \dots, \delta_m)$, if its probability density function (pdf) is given by

$$f(y|\mu, \Sigma, \delta) = 2^m |A|^{-1/2} \phi_m[A^{-1/2} y_* | I_m] \Phi_m[\Delta(\delta) A^{-1} y_* | I_m - \Delta(\delta) A^{-1} \Delta(\delta)], \quad (17)$$

where $y_* = y - \mu$, $\delta = (\delta_1, \delta_2, \dots, \delta_m)^T$ is a skewness parameter vector, $A = \Sigma + \Delta^2(\delta)$, $\phi_m(y|V)$ and $\Phi_m(y|V)$ denote the pdf and the cumulative distribution function (cdf), respectively, of $N_m(\mathbf{0}, V)$. We denote this by $Y \sim SN_m(\mu, \Sigma, \Delta(\delta))$. The mean and covariance matrix are given by $E(Y) = \mu + \sqrt{2/\pi}\delta$, $cov(Y) = \Sigma + (1 - 2/\pi)\Delta^2(\delta)$. As claimed by Sahu et al.(2003), this multivariate SN pdf is very convenient to implement in a Bayesian framework, since it is built using conditional method. It is noted that when $\delta = \mathbf{0}$, the SN distribution reduces to usual normal distribution. In order to better understand the shape of a SN distribution as a referee suggested, plots of a SN density as a function of the skewness parameter with $\delta = -5, 0, 5$ are shown in Figure 8.

B. WinBUGS codes for the three modeling approaches

We enclose WinBUGS program code based on Model I (SN-BSNLME model) for reference. WinBUGS program codes of Models II and III can be easily modified from it or are available from authors upon request.

Model I: SN Bayesian Semiparametric Nonlinear Mixed-Effects (SN-BSNLME)

Model in conjunction with the semiparametric biexponential model (11).

Variables in the datasets “rnacd48” and “Z-matrix” for spline:

##y[,1] = serial number

##y[,2] = arm

##y[,3] = patid

##y[,4] = id

##y[,5] = time (day)

##y[,6] = logrna

##y[,7] = rna

##y[,8] = cd4

##y[,9] = cd8

##[,10] = logerna

##y[,11] = cd4 (standardized)

##y[,12] = cd8 (standardized)

##y[,13] = time(day) (rescaled between 0 and 1 by dividing by max)

##y[,14] = cd4.baseline (standardized)

##Z[,] = Values of spline basis functions

##Begin of model

model {

for (i in 1:44){

a2[i,1] < -0

a2[i,2] < -0

```

a2[i,3]<-0
a2[i,4]<-0
a3[i,1]<-0
a3[i,2]<-0
a3[i,3]<-0
b[i,1:4] ~dmnorm(a2[i,1:4],Sigma2[,]) # random effects for main model
a[i,1:3] ~dmnorm(a3[i,1:3],Sigma3[,]) # random effects for covariate model
}
for (j in 1 : 310) {
## Modeling true CD4 via measurement errors model
z.star[j]< -(alpha[1]+a[y[j,4],1])+(alpha[2]+a[y[j,4],2])*y[j,13]+(alpha[3]+
a[y[j,4],3])*y[j,13]*y[j,13]+delta2*w2[j] ## Skew-normal
w2[j] ~ dnorm(0, 1 ) I(0,)
# z.star[j]< -(alpha[1]+a[y[j,4],1])+(alpha[2]+a[y[j,4],2])*y[j,13]+
# (alpha[3]+a[y[j,4],3])*y[j,13]*y[j,13] ## Normal
y[j,11] ~ dnorm(z.star[j],tau2) ## y[j,11]=observed cd4
## Viral load response model associated with true CD4 covariate
betai1[j] <- beta[1]+ b[y[j,4],1] ## y[,4]=id
betai2[j] <- beta[2]+b[y[j,4],2]
betai3[j] <- beta[3]+ b[y[j,4],3]
betaij4[j] <- beta[4]+beta[5]*z.star[j]+mu.not[1]*Z[j,2]+mu.not[2]*Z[j,3]+b[y[j,4],4]
dm1[j] <- betai1[j]-step(betai2[j]-betaij4[j])*betai2[j]*y[j,13]
dm2[j] <- betai3[j]-step(betai2[j]-betaij4[j])*betaij4[j]*y[j,13]
dm3[j]<- exp(dm1[j])
dm4[j]<- exp(dm2[j])
dm5[j]<- dm3[j] +dm4[j]
mu[j] <- log(dm5[j]) + delta*w[j] ## Skew-normal
w[j] ~ dnorm(0, 1 ) I(0,)
# mu[j] <- log(dm5[j]) ## Normal
y[j,10] ~ dnorm(mu[j], tau) ## y[,10]=logerna
## Fitted values and residues
fit[j] <- mu[j]
res[j] <- y[j,10]-fit[j]
}
## Prior distributions of the hyperparameters
# (1) Coefficients
for (l in 1:5)
{ beta[l]~ dnorm(0,1.0E-2) }
for (l in 1:2)
{ mu.not[l]~ dnorm(0,1.0E-2)}

```

```

for (k in 1:3){ alpha[k]~ dnorm(0,1.0E-2) }
# (2) Precision parameters
tau ~ dgamma(0.01,0.01)
sigma.tau <- 1/tau
tau2 ~ dgamma(0.01,0.01)
sigma.tau2 <- 1/tau2
# (3) Variance-covariance matrices
Sigma2[1:4,1:4] ~ dwish(R2[,],5)
v2[1:4,1:4] <- inverse(Sigma2[,])
Sigma3[1:3,1:3]~ dwish(R3[,],5)
v3[1:3,1:3] <- inverse(Sigma3[,])
# (4) Skewness parameters
delta ~ dnorm(0.0, 0.01)
delta2 ~ dnorm(0.0, 0.01)
} ## End of model
## Part of data
list(R2=structure( .Data = c(1, 0,0,0,0,1,0,0,0,0,1,0,0,0,0,1), .Dim = c(4, 4)),
R3 = structure( .Data = c(1, 0,0,0,1,0,0,0,1), .Dim = c(3, 3)))
## Initial values
list(beta=c(5.4,3.4,4.0,1.5,4.2),alpha=c(0,0,0),mu.not=c(0,0),
delta=0.5,delta2=0.5,tau=1,tau2=1,
Sigma2= structure( .Data = c(1.229, .043,-.750,.170, .043,.090,.002,-.013,
-.750,-.002,1.059,-.214, .170,-.013,-.214, .120), .Dim = c(4, 4)),
Sigma3= structure( .Data = c(.1, 0,0, 0,.1,0, 0,0,.1), .Dim = c(3, 3)))
## End of Program

```

References

- Acosta, E. P., Wu, H., Hammer, S.M., Yu, S., Kuritzkes, D. R., Walawander, A., Eron, J.J., Fichtenbaum, C.J, Pettinelli, C., Neath, D., Ferguson, E., Saan, A. J., and Gerber, J. G. for the Adult ACTG 5055 Protocol Team. (2004). Comparison of two indinavir/ritonavir regimens in treatment of HIV-infected individuals. *Journal of Acquired Immune Deficiency Syndromes* **37**, 1358–1366.
- Arellano-Valle, R.B., Bolfarine, H., and Lachos, V.H. (2005). Skew-normal linear mixed models. *Journal of Data Science*, **3**, 415–438.
- Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of Royal Statistical Society, Series B*, **67**, 579–602.

- Arellano-Valle, R.B., and Genton, M.G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, **96**, 93–116.
- Arellano-Valle, R.B., and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.
- Azzalini, A., and Dalla-Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**, 715–726.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective 2nd edition* London: Chapman and Hall.
- Ding, A. A., and Wu, H. (1999). Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. *Mathematical Biosciences*, **160**, 63–82.
- Davidian, M., and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, London: Chapman & Hall.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.
- Ghosh, P., Branco, M.D., and Chakraborty, H. (2007). Bivariate random effect model using skew normal distribution with application to HIV-RNA. *Statistics in Medicine* **26**, 1255–1267.
- Higgins, M., Davidian, M., and Giltinan D. M. (1997). A two-step approach to measurement error in time-dependent covariates in nonlinear mixed-effects models, with application to IGF-I pharmacokinetics. *Journal of the American Statistical association* **92**, 436–448.
- 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 Dagne G. (2010). Skew-normal Bayesian nonlinear mixed-effects models with application to AIDS studies. *Statistics in Medicine*, **29**, 2384–2398.
- Lange, K.L. and Sinsheimer, J.S. (1993). Normal/independent distributions and their applications in robust regression. *J. Comput. Graph. Stat* **2**, 175–198.
- Liang, H., Wu, H. and Carroll, R. (2003). The relationship between virologic and immunologic responses in aids clinical research using mixed-effect varying-coefficient semiparametric models with measurement error. *Biostatistics*, **4**, 297–312.
- Liu, W., and Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics*, **63**, 342–350.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, an extensibility. *Statistics and Computing*, **10**, 325–337.

- Müller, P., and Rosner, G.L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, **92**, 1279–1292.
- Notermans, D. W., Goudsmit, J., Danner, S. A., De Wolf, F., Perelson, A.S., and Mittler, J. (1998). Rate of HIV-1 decline following antiretroviral therapy is related to viral load at baseline and drug regimen. *AIDS* **12**, 1483–1490.
- Perelson, A. S. and Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D.D. (1997). Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**, 188–191.
- Sahu, S.K., Dey, D.K., and Branco, M.D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics* **31**, 129–150.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.
- Verbeke, G., and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.
- Wu, H., and Ding, A. A. (1999). Population HIV-1 dynamics *in vivo*: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410–418.
- Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of the American Statistical Association*, **97**, 955–964.
- Wu, H., Zhang, J.T. (2002). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine*, **21**, 3655–3675.
- Wu, H., Lathey, J., Ruan, P., Douglas, S.D., Spector, S.A., Lindsey, J., Hughes, M.D., Rudy, B.J., Flynn, P.M., and PACTG 381 TEAM. (2004). Relationship of plasma HIV-1 RNA dynamics to baseline factors and virological responses to Highly Active Antiretroviral Therapy (HAART) in adolescents infected through risk behavior. *Journal of Infectious Diseases* **189**, 593–601.
- Wu, H., Zhao, C., and Liang, H. (2004). Comparison of linear, nonlinear and semi-parametric mixed-effects models for estimating HIV dynamic parameters. *Biometrical Journal* **6**, 233–245.
- Wu, L. (2004). Simultaneous inference for longitudinal data with detection limits and covariates measured with errors, with application to AIDS studies. *Statistics in Medicine*, **23**, 1715–1731.