Bayesian Analysis (2012) **7**, Number 1, pp. 189–210

Simultaneous Bayesian Inference for Skew-Normal Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors

Yangxin Huang* and Getachew A. Dagne†

Abstract. Longitudinal data arise frequently in medical studies and it is a common practice to analyze such complex data with nonlinear mixed-effects (NLME) models which enable us to account for between-subject and within-subject variations. To partially explain the variations, covariates are usually introduced to these models. Some covariates, however, may be often measured with substantial errors. It is often the case that model random error is assumed to be distributed normally, but the normality assumption may not always give robust and reliable results, particularly if the data exhibit skewness. Although there has been considerable interest in accommodating either skewness or covariate measurement error in the literature, there is relatively little work that considers both features simultaneously. In this article, our objectives are to address simultaneous impact of skewness and covariate measurement error by jointly modeling the response and covariate processes under a general framework of Bayesian semiparametric nonlinear mixed-effects models. The method is illustrated in an AIDS data example to compare potential models which have different distributional specifications. The findings from this study suggest that the models with a skew-normal distribution may provide more reasonable results if the data exhibit skewness and/or have measurement errors in covariates.

Keywords: Bayesian approach, Covariate measurement errors, HIV/AIDS, Joint models, Longitudinal data, Semiparametric nonlinear mixed-effects models, Skew-normal distribution.

1 Introduction

Viral dynamic studies have a common structure in the sense that they use repeated measures over a treatment period to assess rates of changes in viral load over time. There has been substantial interest in estimating viral dynamic parameters in order to acquire more comprehensive understanding of the pathogenesis of HIV infection and to assess the effectiveness of antiretroviral treatment (Wu and Ding, 1999). Thus various statistical modeling and analysis methods have been used, in conjunction with HIV dynamic models and also pharmacokinetic (PK) and pharmacodynamic (PD) models, for

*Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, FL, yhuang@health.usf.edu

†Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, FL, gdagne@health.usf.edu

 DOI:10.1214/12-BA706

statistical inference and analysis. Some of those methods are linear mixed-effects (LME) and nonlinear mixed-effects (NLME) modeling (Lunn et al., 2002; Wu et al., 1998; Wu and Ding, 1999), nonparametric NLME modeling (Liu and Wu, 2007; Wu and Zhang, 2002), and Bayesian NLME modeling via Markov Chain Monte Carlo (MCMC)(Huang et al., 2006; Huang and Dagne, 2010; Lunn et al., 2002). However, there is relatively little work done on simultaneously accounting for the biases induced by the following three issues.

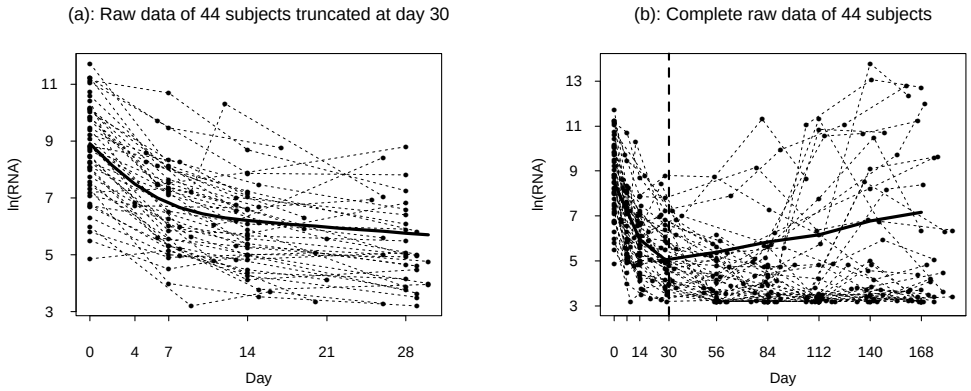

Figure 1: Profiles of HIV viral load measurements (in natural log scale) from a group of patients in an AIDS clinical study. Change in viral load during treatment is shown for day 0 to day 30 (a) with the solid curve being the estimate of viral load trajectory from a parametric model (B.9) and for day 0 to the end of study (b) with the solid curve being the estimate of viral load trajectory from a semiparametric model (B.10).

Firstly, current HIV dynamic models (Wu et al., 1998; Wu and Ding, 1999) are mostly developed to quantify short-term dynamics. For example, Figure 1(a) displays the early stage trajectories based on the first 30-day viral load data (in natural log scale) for 44 subjects enrolled in an AIDS clinical trial study (A5055) (Acosta et al., 2004), while Figure 1(b) includes the complete (long-term) viral load data of the same patients. In Figure 1(a) the solid curve is the average trajectory estimated from a parametric model (B.9), and in Figure 1(b) the solid curve is the population estimate obtained from a semiparametric model (B.10); both models are to be discussed in Section 3.1. These models fit only the early stage of the viral load trajectory (Figure 1(a)) and are limited to interpreting earlier stage HIV dynamic data from AIDS clinical trials. Moreover, as is seen from Figure 1(b), the viral load trajectory may change to different shapes at later stages. This general phenomenon in HIV dynamics, as illustrated in the example presented in Figure 1, has motivated us to investigate long-term HIV dynamic studies via a semiparametric model (B.10). Secondly, the commonly assumed distribution for model random errors is normal, but this assumption may lack the robustness against departures from normality and/or outliers and thus statistical inference and analysis with normal assumption may lead to misleading results (Verbeke and Lesaffre, 1996; Sahu et al., 2003). Specifically, the distributions of the outcomes

in virologic response are skewed with a number of outliers. Figure 2 displays the distributions of repeated viral load (in natural log scale) and standardized CD4 cell count measurements for 44 subjects enrolled in the A5055 (Acosta et al., 2004). It can be seen that, for this data set to be analyzed in this paper, both the viral load response (even after log-transformation) and CD4 covariate are highly skewed, and thus a normality assumption is not quite realistic. Thirdly, the NLME models have been used in the the literature to account for both between-subject and within-subject variations in response measurements associated with covariates (Wu et al., 1998; Wu and Ding, 1999). However, the covariates such as CD4 cell counts are often measured with substantial errors, and thus statistical inference ignoring measurement errors in covariates may result in biased estimation results. To the best of our knowledge, there is relatively little work done on simultaneously accounting for the biases induced by mismeasured covariates and misspecified model error distribution under a Bayesian framework in conjunction with semiparametric nonlinear mixed-effects (SNLME) models. It is not clear how covariate measurement error and skewness of data may interact and simultaneously influence inferential procedures.

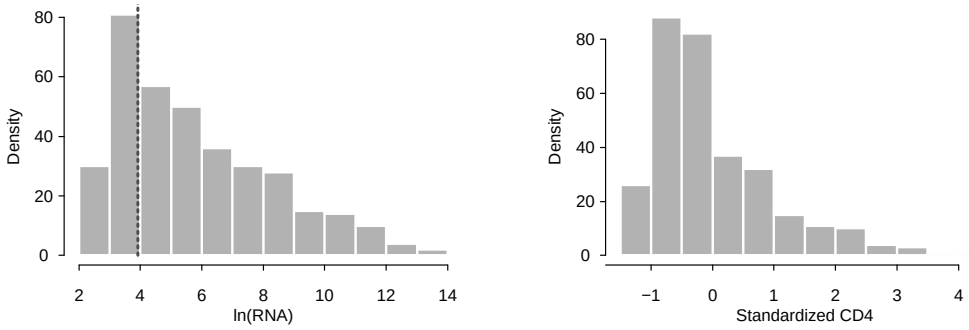

Figure 2: Histograms of viral load measured from RNA levels (in natural log scale) and standardized CD4 cell count in plasma for 44 patients in an AIDS clinical trial study.

The main goal of this article is to investigate the effects of skewness in response and covariate variables, and of measurement error in covariates, on statistical inference by introducing skew-normal (SN) Bayesian semiparametric nonlinear mixed-effects (SN-BSNLME) joint models. Specifically, we jointly consider an SN semiparametric nonlinear mixed-effects (SN-SNLME) model for response process and an SN linear mixed-effects (SN-LME) model for the covariate process. In formulating this joint model, we consider a multivariate SN distribution introduced by Sahu et al.(2003) which is suitable for a Bayesian analysis and also briefly discussed in Appendix A. The remainder of the article is organized as follows. Section 2 introduces model setup in general forms and investigates associated inference methods that simultaneously account for skewness and covariate measurement error. In Section 3, we discuss the specific models for HIV dynamics and CD4 covariate process, along with description of an AIDS clinical data set that is used to illustrate the proposed methods and then report the results. Finally, the paper concludes with discussion in Section 4.

2 Bayesian approach to joint models of response and covariate processes

2.1 SNLME joint model with a skew-normal distribution

In this section, we present the joint models in general forms, illustrating that our models and methods may be applicable to other fields as well. Various covariate models were investigated in the literature (Carroll et al., 2006; Liu and Wu, 2007; Wu, 2002). For those covariate models, however, the commonly assumed distribution for random errors is normal and this assumption may lack robustness against departures from normality. We extend covariate models with measurement errors to have an SN distribution. For simplicity, we consider a single time-varying covariate with measurement errors. Denote the number of subjects by n and the number of measurements on the ith subject by n_i. Let z_{ij} be the observed covariate value for individual i at time t_{ij} $(i = 1, \ldots n; j = 1, \ldots, n_i)$. We consider the following LME covariate model with an SN distribution:

$$z_{ij} = \boldsymbol{u}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{v}_{ij}^T\boldsymbol{a}_i + \epsilon_{ij} \;\; (\equiv z_{ij}^* + \epsilon_{ij}), \boldsymbol{\epsilon}_i \sim \; iid \; SN_{n_i}\left(-\sqrt{2/\pi}\boldsymbol{\delta}_{\epsilon_i}, \tau^2\boldsymbol{I}_{n_i}, \boldsymbol{\Delta}(\boldsymbol{\delta}_{\epsilon_i})\right), \tag{B.1}$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{in_i})^T$ with z_{ij} being the covariate value for individual i at time t_{ij}, $\boldsymbol{z}_i^* = (z_{i1}^*, \ldots, z_{in_i}^*)^T$ and $z_{ij}^* = \boldsymbol{u}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{v}_{ij}^T\boldsymbol{a}_i$ may be viewed as the true (but unobservable) covariate value at time t_{ij}, $\boldsymbol{u}_{ij}$ and $\boldsymbol{v}_{ij}$ are $l \times 1$ design vectors, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^T$ and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{il})^T$ are unknown population (fixed-effects) and individual-specific (random-effects) parameter vectors, respectively, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ follows an SN distribution with ϵ_{ij} being the model error for individual i at time t_{ij}, τ^2 is the unknown within-individual scale parameter. The $n_i \times n_i$ skewness diagonal matrix $\boldsymbol{\Delta}(\boldsymbol{\delta}_{\epsilon_i}) = diag(\delta_{\epsilon_{i1}}, \ldots, \delta_{\epsilon_{in_i}})$ and $n_i \times 1$ skewness parameter vector $\boldsymbol{\delta}_{\epsilon_i} = (\delta_{\epsilon_{i1}}, \ldots, \delta_{\epsilon_{in_i}})^T$. In particular, if $\delta_{\epsilon_{i1}} = \cdots = \delta_{\epsilon_{in_i}} \hat{=} \delta_\epsilon$, then $\boldsymbol{\Delta}(\boldsymbol{\delta}_{\epsilon_i}) = \delta_\epsilon\boldsymbol{I}_{n_i}$ and $\boldsymbol{\delta}_{\epsilon_i} = \delta_\epsilon\boldsymbol{1}_{n_i}$ with $\boldsymbol{1}_{n_i} = (1, \ldots, 1)^T$; this indicates that we are interested in skewness of the overall data set, which is the case to be used in real data analysis below. We assume that $\boldsymbol{a}_i \sim \; iid \; N_l(0, \boldsymbol{\Sigma}_a)$, where $\boldsymbol{\Sigma}_a$ is the unrestricted covariance matrix. Note that the model (B.1) may be interpreted as an SN covariate measurement error model.

For the response process, we consider a general SNLME model which incorporates possibly mismeasured time-varying covariates and model random error with an SN distribution:

$$\begin{aligned} y_{ij} &= g(t_{ij}, \boldsymbol{\beta}_{ij}^\dagger, \phi(t_{ij})) + e_{ij}, \;\; \boldsymbol{e}_i \sim \; iid \; SN_{n_i}\left(-\sqrt{2/\pi}\boldsymbol{\delta}_{e_i}, \sigma^2\boldsymbol{I}_{n_i}, \boldsymbol{\Delta}(\boldsymbol{\delta}_{e_i})\right), \\ \boldsymbol{\beta}_{ij}^\dagger &= \boldsymbol{d}^\dagger[z_{ij}^*, \boldsymbol{\beta}^\dagger, \boldsymbol{b}_i^\dagger], \; \phi(t_{ij}) = v[w(t_{ij}), h_i(t_{ij})], \; \boldsymbol{b}_i^\dagger \sim \; iid \; N_{s_3}(\boldsymbol{0}, \boldsymbol{\Sigma}_b^\dagger), \end{aligned} \tag{B.2}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ with y_{ij} being the response value for individual i at t_{ij} $(i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, n_i)$, $g(\cdot)$, $\boldsymbol{d}^\dagger(\cdot)$ 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, $\boldsymbol{\beta}_{ij}^\dagger$ is $s_1 \times 1$ individual-specific time-dependent parameter vector, $\boldsymbol{\beta}^\dagger$ is $s_2 \times 1$ population parameter

vector ($s_2 \geq s_1$), σ^2 is the unknown within-subject scale parameter, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$ is the vector of random errors, $\boldsymbol{b}_i^\dagger$ is $s_3 \times 1$ vector of random effects ($s_3 \leq s_1$), $\boldsymbol{\Sigma}_b^\dagger$ is the unrestricted covariance matrix, the $n_i \times n_i$ skewness diagonal matrix $\boldsymbol{\Delta}(\boldsymbol{\delta}_{e_i}) = diag(\delta_{e_{i1}}, \ldots, \delta_{e_{in_i}})$ and the $n_i \times 1$ skewness parameter vector $\boldsymbol{\delta}_{e_i} = (\delta_{e_{i1}}, \ldots, \delta_{e_{in_i}})^T$. In particular, if $\delta_{e_{i1}} = \cdots = \delta_{e_{in_i}} \hat{=} \delta_e$, then $\boldsymbol{\Delta}(\boldsymbol{\delta}_{e_i}) = \delta_e \boldsymbol{I}_{n_i}$ and $\boldsymbol{\delta}_{e_i} = \delta_e \boldsymbol{1}_{n_i}$. In the model (B.2), we assume that the individual-specific parameters $\boldsymbol{\beta}_{ij}^\dagger$ depend on the true (but unobservable) value of covariate z_{ij}^* rather than the observed covariate z_{ij}, which may be measured with errors.

The SNLME model (B.2) reverts to an NLME model when the nonparametric parts $w(t)$ and $h_i(t)$ are constants. To fit model (B.2), we apply the regression spline method. The main idea of a regression spline is to approximate $w(t)$ and $h_i(t)$ by using a linear combination of spline basis functions (Wu and Zhang, 2002). For instance, $w(t)$ and $h_i(t)$ can be approximated by a linear combination of basis functions $\boldsymbol{\Psi}_p(t) = \{\psi_0(t), \psi_1(t), ..., \psi_{p-1}(t)\}^T$ and $\boldsymbol{\Phi}_q(t) = \{\phi_0(t), \phi_1(t), ..., \phi_{q-1}(t)\}^T$, respectively. That is,

$$w(t) \approx w_p(t) = \sum_{k=0}^{p-1} \mu_k \psi_k(t) = \boldsymbol{\Psi}_p(t)^T \boldsymbol{\mu}_p, \quad h_i(t) \approx h_{iq}(t) = \sum_{k=0}^{q-1} \xi_{ik} \phi_k(t) = \boldsymbol{\Phi}_q(t)^T \boldsymbol{\xi}_{iq}, \tag{B.3}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\xi}_{iq}$ ($q \leq p$) are the unknown vectors of fixed and random coefficients, respectively. Based on the assumption of $h_i(t)$, we can regard $\boldsymbol{\xi}_{iq}$ as *iid* realizations of a zero-mean random vector. 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 (B.2) in a compact way as follows:

$$y_{ij} = g\left(t_{ij}, \boldsymbol{d}^\dagger[z_{ij}^*, \boldsymbol{\beta}^\dagger, \boldsymbol{b}_i^\dagger], v[\boldsymbol{\Psi}_p(t_{ij})^T \boldsymbol{\mu}_p, \boldsymbol{\Phi}_q(t_{ij})^T \boldsymbol{\xi}_{iq}]\right) + e_{ij} \equiv g\left(t_{ij}, \boldsymbol{d}(z_{ij}^*, \boldsymbol{\beta}, \boldsymbol{b}_i)\right) + e_{ij}, \tag{B.4}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}^{\dagger T}, \boldsymbol{\mu}_p^T)^T$ and $\boldsymbol{b}_i = (\boldsymbol{b}_i^{\dagger T}, \boldsymbol{\xi}_{iq}^T)^T$ are the vectors of fixed-effects and random-effects, respectively, and $\boldsymbol{d}(\cdot)$ is a known but possibly nonlinear function. Thus, for given $\boldsymbol{\Psi}_p(t)$ and $\boldsymbol{\Phi}_q(t)$, we approximate the SN-SNLME model (B.2) by the following SN-NLME model:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{g}_i(\boldsymbol{t}_i, \boldsymbol{\beta}_i) + \boldsymbol{e}_i, & \boldsymbol{e}_i &\sim SN_{n_i}\left(-\sqrt{2/\pi}\boldsymbol{\delta}_{e_i}, \sigma^2 \boldsymbol{I}_{n_i}, \boldsymbol{\Delta}(\boldsymbol{e}_{e_i})\right), \\ \boldsymbol{\beta}_{ij} &= \boldsymbol{d}[z_{ij}^*, \boldsymbol{\beta}, \boldsymbol{b}_i], & \boldsymbol{b}_i &\sim N_{s_4}(\boldsymbol{0}, \boldsymbol{\Sigma}_b), \end{aligned} \tag{B.5}$$

where $s_4 = s_3 + q$, $\boldsymbol{g}_i(\boldsymbol{t}_i, \boldsymbol{\beta}_i) = (g(t_{i1}, \boldsymbol{\beta}_{i1}), \ldots, g(t_{in_i}, \boldsymbol{\beta}_{in_i}))^T$ with $g(\cdot)$ being a known nonlinear function, $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})^T$, $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_{i1}, \ldots, \boldsymbol{\beta}_{in_i})^T$ and $\boldsymbol{\Sigma}_b$ is an unstructured covariance matrix. We assume that $\boldsymbol{e}_i$, $\boldsymbol{\epsilon}_i$, $\boldsymbol{b}_i$ and $\boldsymbol{a}_i$ are independent of each other.

2.2 Simultaneous Bayesian inference for parameter estimation

In a longitudinal study, the longitudinal response and covariate processes are usually connected physically or biologically. Although a simultaneous inference method based on a joint likelihood for the covariate and response data may be favorable, the computation associated with the joint likelihood inference in joint models of longitudinal data can be extremely intensive and may lead to convergence problems (Liu and Wu, 2007; Wu, 2002). Here we propose a simultaneous Bayesian inference method for models (B.1) and (B.5) based on an MCMC procedure for the covariate and response data. The Bayesian joint modeling approach paves a way to alleviate the computational burdens and to overcome convergence problems.

To carry out a Bayesian inference, prior distributions for unknown parameters in the models (B.1) and (B.5) need to be assessed as follows:

$$\begin{array}{llll} \boldsymbol{\alpha} \sim N_l(\boldsymbol{\alpha}_0, \boldsymbol{\Lambda}_1), & \tau^2 \sim IG(\omega_1, \omega_2), & \boldsymbol{\Sigma}_a \sim IW(\boldsymbol{\Omega}_1, \nu_1), & \boldsymbol{\delta}_{\epsilon_i} \sim N_{n_i}(\mathbf{0}, \boldsymbol{\Gamma}_1), \\ \boldsymbol{\beta} \sim N_{s_5}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_2), & \sigma^2 \sim IG(\omega_3, \omega_4), & \boldsymbol{\Sigma}_b \sim IW(\boldsymbol{\Omega}_2, \nu_2), & \boldsymbol{\delta}_{e_i} \sim N_{n_i}(\mathbf{0}, \boldsymbol{\Gamma}_2), \end{array} \tag{B.6}$$

where $s_5 = s_2 + p$, and 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 $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Omega}_2$, $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ can be assumed to be diagonal for convenient implementation. Following Sahu et al.(2003) and properties of the SN distribution, in order to specify the models (B.1) and (B.5) for MCMC computation it can be shown by introducing two $n_i \times 1$ random variable vectors $\boldsymbol{w}_{e_i} = (w_{e_{i1}}, \ldots, w_{e_{in_i}})^T$ and $\boldsymbol{w}_{\epsilon_i} = (w_{\epsilon_{i1}}, \ldots, w_{\epsilon_{in_i}})^T$ based on the stochastic representation for the SN distribution (see Appendix A for details) that $\boldsymbol{z}_i$ and $\boldsymbol{y}_i$ follow the following distributions

$$\begin{aligned} &\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{w}_{\epsilon_i}; \boldsymbol{\alpha}, \tau^2, \boldsymbol{\delta}_{\epsilon_i} \sim N_{n_i}\left(\boldsymbol{z}_i^* + \boldsymbol{\Delta}(\boldsymbol{\delta}_{\epsilon_i})[\boldsymbol{w}_{\epsilon_i} - \sqrt{2/\pi}\mathbf{1}], \tau^2 \boldsymbol{I}_{n_i}\right), \\ &\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{w}_{e_i}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}_{e_i} \sim N_{n_i}\left(\boldsymbol{g}_i + \boldsymbol{\Delta}(\boldsymbol{\delta}_{e_i})[\boldsymbol{w}_{e_i} - \sqrt{2/\pi}\mathbf{1}], \sigma^2 \boldsymbol{I}_{n_i}\right), \\ &\boldsymbol{w}_{\epsilon_i} \sim N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i}) I(\boldsymbol{w}_{\epsilon_i} > \mathbf{0}),\ \boldsymbol{w}_{e_i} \sim N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i}) I(\boldsymbol{w}_{e_i} > \mathbf{0}), \end{aligned} \tag{B.7}$$

where $\mathbf{1} = (1, \ldots, 1)^T$, $I(\boldsymbol{w}_{e_i} > \mathbf{0})$ is an indicator function and $\boldsymbol{w}_{e_i} \sim N_{n_i}(0, \boldsymbol{I}_{n_i})$ truncated in the space $\boldsymbol{w}_{e_i} > \mathbf{0}$; $\boldsymbol{w}_{\epsilon_i}$ can be defined similarly. An important advantage of the above representations based on the hierarchical models (B.1) and (B.5) is that they can be very easily implemented using the freely available WinBUGS software (Lunn et al., 2000) and that the computational effort is equivalent to the one necessary to fit the normal version of the model.

Let $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_b, \boldsymbol{\delta}_{\epsilon_i}, \boldsymbol{\delta}_{e_i}; i = 1, \ldots, n\}$ be the collection of unknown parameters in models (B.1) and (B.5), and $f(\cdot|\cdot)$ and $\pi(\cdot)$ be a conditional density function and a prior density function, respectively. Denote the observed data by $\mathcal{D} = \{(\boldsymbol{y}_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$. We assume that $\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_b, \boldsymbol{\delta}_{\epsilon_i}, \boldsymbol{\delta}_{e_i}$ $(i = 1, \ldots, n)$ are independent of each other, and thus we have $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\tau^2)\pi(\sigma^2)\pi(\boldsymbol{\Sigma}_a)\pi(\boldsymbol{\Sigma}_b)$ $\prod_i \pi(\boldsymbol{\delta}_{\epsilon_i})\pi(\boldsymbol{\delta}_{e_i})$. 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 the Bayesian framework. The joint posterior density of $\boldsymbol{\theta}$ based on the observed data can be given by

$$f(\boldsymbol{\theta}|\mathcal{D}) \propto \{\prod_{i}^{n} \int\int f(\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{w}_{e_i}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}_{e_i})\ f(\boldsymbol{w}_{e_i}|\boldsymbol{w}_{e_i} > \mathbf{0}) \times f(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{w}_{\epsilon_i}; \boldsymbol{\alpha}, \tau^2, \boldsymbol{\delta}_{\epsilon_i})\ f(\boldsymbol{w}_{\epsilon_i}|\boldsymbol{w}_{\epsilon_i} > \mathbf{0})\ f(\boldsymbol{a}_i|\boldsymbol{\Sigma}_a)\ f(\boldsymbol{b}_i|\boldsymbol{\Sigma}_b) d\boldsymbol{a}_i d\boldsymbol{b}_i\} \pi(\boldsymbol{\theta}). \tag{B.8}$$

In general, the integrals in (B.8) are of high dimension and do not have closed form. Analytic approximations to the integrals may not be sufficiently accurate. Therefore, it is prohibitive to directly calculate the posterior distribution of $\boldsymbol{\theta}$ based on the observed data. As an alternative, MCMC procedures can be used to sample based on (B.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 An Application to AIDS Studies

3.1 HIV dynamic models

Viral dynamic models can be formulated through a system of ordinary differential equations (ODE) for response variable, HIV RNA copies (viral load) (Huang et al., 2006; Laveille et al., 2011; Wu et al., 1998; Wu and Ding, 1999). The biexponential model derived from a system of ODE based on a compartmental analysis (Wu and Ding, 1999) is the most popular model for HIV dynamics:

$$y(t) = \ln\{V(t)\} + e(t) = \ln\{\exp[\rho_1 - \lambda_1 t] + \exp[\rho_2 - \lambda_2 t]\} + e(t), \tag{B.9}$$

where $V(t)$ is the plasma HIV-1 RNA levels (viral load) at time t, λ_1 and λ_2 are called the first- and second-phase viral decay rates, which may represent the minimum turnover rate of productively infected cells and that of latently or long-lived infected cells, respectively, $\exp(\rho_1) + \exp(\rho_2)$ is the baseline viral load at time $t = 0$, which is related to the parameters ρ_1 and ρ_2. It is of particular interest to estimate the viral decay rates λ_1 and λ_2 because they quantify the antiviral effect and, hence, can be used to assess the efficacy of the antiviral treatment. In estimating these decay rates, only the early segment of the viral load trajectory data has been used (Wu and Ding, 1999). Since the viral load trajectory may change to different shapes in the late stages, it may not be reasonable to assume that the second-phase decay rate remains constant during long-term treatment. To model the long-term HIV dynamics, a semiparametric biexponential model can be constructed as follows (Wu and Zhang, 2002):

$$y(t) = \ln\{V(t)\} + e(t) = \ln\{\exp[\rho_1 - \lambda_1 t] + \exp[\rho_2 - \lambda_2(t) t]\} + e(t), \tag{B.10}$$

where the second-phase decay rate $\lambda_2(t)$ is an unknown smooth function. Intuitively, model (B.10) is more reasonable because it assumes that the decay rate can vary with time as a result of drug resistance, pharmacokinetics, drug adherence and other relevant clinical factors. Therefore, all data obtained during antiretroviral (ARV) treatment can be used to fit model (B.10). This is a semiparametric model because of the mechanistic

structure (two-exponential) with constant parameters $(\lambda_1, \rho_1, \rho_2)$ and a time-varying parameter $(\lambda_2(t))$ to capture the time-varying effects of the treatment over a longer period. Actually, by including long-term viral load data, the estimate of λ_1 can be more accurate and reasonable compared with those obtained in previous studies (Wu and Ding, 1999) after excluding long-term viral load data for modeling and analysis by some *ad hoc* rules. In the mean time, this model enjoys the flexibility of a semiparametric function for the second-phase decay rate $\lambda_2(t)$. The estimate of $\lambda_2(t)$ 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 drug resistance, non-compliance and other clinical determinants.

To model the covariate CD4 process, in the absence of theoretical rationale for the CD4 trajectories, we consider empirical polynomial LME models for the CD4 process, and choose the best model based on AIC and BIC values. Specifically, we consider the covariate model (B.1) with $\boldsymbol{u}_{ij} = \boldsymbol{v}_{ij} = (1, t_{ij}, \ldots, t_{ij}^{l-1})^T$ and focus on linear $(l = 2)$, quadratic $(l = 3)$ and cubic $(l = 4)$ polynomials. The resulting AIC (BIC) values are 799.03 (821.74), 703.56 (744.42) and 766.18 (782.08), respectively. We thus adopted the following quadratic polynomial SN-LME model (B.1) for the observed CD4 process:

$$z_{ij} = (\alpha_1 + a_{i1}) + (\alpha_2 + a_{i2})t_{ij} + (\alpha_3 + a_{i3})t_{ij}^2 + \epsilon_{ij}, \tag{B.11}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T \sim \ iid\ SN_{n_i}\left(-\sqrt{2/\pi}\delta_\epsilon \mathbf{1}_{n_i}, \tau^2 \boldsymbol{I}_{n_i}, \delta_\epsilon \boldsymbol{I}_{n_i}\right)$, $z_{ij}^* = (\alpha_1 + a_{i1}) + (\alpha_2 + a_{i2})t_{ij} + (\alpha_3 + a_{i3})t_{ij}^2$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ is a population (fixed-effects) parameter vector, and $\boldsymbol{a}_i = (a_{i1}, a_{i2}, a_{i3})^T$ is an individual-specific (random-effects) vector with normal distribution $N_3(\mathbf{0}, \boldsymbol{\Sigma}_a)$.

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 (B.3) with $q \leq p$. The values of p and q are determined based on the AIC/BIC which suggest the following function for $\beta_{ij4}(t_{ij})$ with $p = 3$ and $q = 1$ in the model (B.3):

$$\lambda_{ij2}(t_{ij}) = w(t_{ij}) + h_i(t_{ij}) \approx \mu_0 + \mu_1\psi_1(t_{ij}) + \mu_2\psi_2(t_{ij}) + \xi_{i0}, \tag{B.12}$$

where $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2)^T$ and $\xi_{i0} \equiv b_{i4}$.

Because viral load is measured on each subject repeatedly over the study period, the measurements obtained from the same subject may be correlated, but they are assumed independent between patients. In the present context of a semiparametric model, it is straightforward to introduce the SNLME model in conjunction with the HIV dynamic model (B.10) as follows:

$$\begin{aligned} &y_{ij} = \ln\{\exp[\rho_{i1} - \lambda_{ij1}t_{ij}] + \exp[\rho_{i2} - \lambda_{ij2}(t_{ij})t_{ij}]\} + e_{ij}, \\ &\rho_{i1} = \beta_1 + b_{i1}, \quad \lambda_{ij1} = \beta_2 + \beta_3 z_{ij}^* + b_{i2}, \\ &\rho_{i2} = \beta_4 + b_{i3}, \quad \lambda_{ij2}(t_{ij}) = w(t_{ij}) + h_i(t_{ij}), \end{aligned} \tag{B.13}$$

where y_{ij} is the natural log-transformation of the viral load for the i^{th} subject at time t_{ij} $(i = 1, 2, ..., n,\ j = 1, 2, ..., n_i)$, within-individual random error $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T \sim$

$SN_{n_i}\left(-\sqrt{2/\pi}\delta_e \mathbf{1}_{n_i}, \sigma^2 \boldsymbol{I}_{n_i}, \delta_e \boldsymbol{I}_{n_i}\right)$, the covariate value z_{ij}^* is referred to as a summary of the true (but unobservable) CD4 value for the i^{th} subject at time t_{ij} in association with the model (B.11), $\boldsymbol{\beta}_{ij} = (\rho_{i1}, \lambda_{ij1}, \rho_{i2}, \lambda_{ij2})^T$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \boldsymbol{\mu}_3^T)^T$ are individual parameters for the i^{th} subject and population parameters, respectively, the random-effects $\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})^T \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}_b)$. We note that the model (B.13) accommodates a time-varying covariate CD4 into the first-phase decay rate and specifies an unknown nonparametric smooth function for the second-phase decay rate for capturing viral rebound.

3.2 Analysis of AIDS data

We illustrate our methods using a real AIDS clinical data (Acosta et al., 2004). The study consists of 44 HIV-infected patients who were treated with a potent ARV regimen. 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 was used to measure RNA viral load, with a lower limit of quantification (BDL) of 50 copies/mL. The viral load measures below this limit are not considered reliable, and they are considered left-censored. For dealing with left-censoring, substitution methods, such as BDL or BDL/2 which produces biased results, have been suggested in the literature (Davidian and Giltinan, 1995). The relatively better methods than substitution methods are to use a Bayesian method, which treats the observations below the detection limit as missing values, and simultaneously predicts them based on a predictive distribution (Gelman et al., 2003) or a maximum likelihood method (Tobin, 1958), which takes into account the proportion of observation below BDL and observed values above BDL. We will use the Bayesian method in the analysis next due to the rationale of the paper. Covariates such as CD4 and CD8 cell counts were also measured throughout the study on a similar scheme. Figure 3 shows the measurements of viral load in natural log scale and CD4 cell count for three randomly selected patients. Both viral load and CD4 cell count trajectories exhibit distinctive and important patterns throughout the time course.

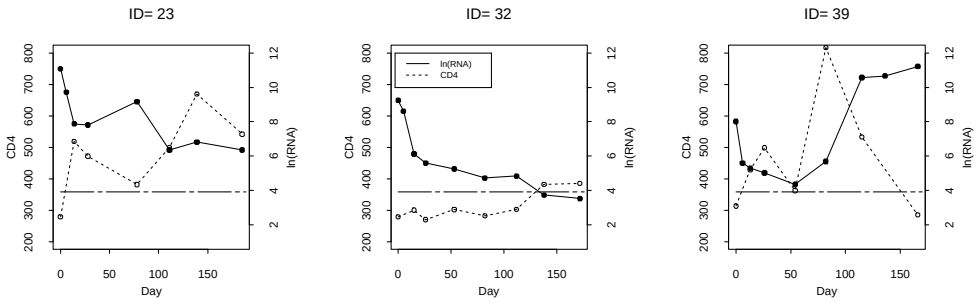

Figure 3: Profiles of viral load (response: solid curve) in natural log scale and CD4 cell count (covariate: dotted curve) for three randomly selected patients. The horizontal line is below the detectable level of viral load (3.91=ln(50)).

In this study, CD4 measurements are known with nonnegligible errors and ignor-

ing covariate measurement errors can lead to severely misleading results in a statistical inference. Some of the CD4 values were missing at viral load measurement time t_{ij}, possibly due to different CD4 measurement schemes as designed in the study (for example, CD4 measurements were missed at day 7 displayed in Figure 3). Thus we assume that the missing data in CD4 are missing at random (MAR) in the sense of Rubin (1976), so that the missing data mechanism can be ignored in the analysis. A natural log-transformation for viral load data was used in the analysis in order to stabilize the variation of measurement error and speed up the estimation algorithm. To avoid very small (large) estimates which may be unstable, we standardize the time-varying covariate CD4 cell counts (from each CD4 value we subtract the mean 375.46 and divide by the standard deviation 228.57) and rescale the original time t (in days) so that the time scale is between 0 and 1. As shown in Figure 2, the distributions of viral load in natural log scale and CD4 cell count clearly indicate their asymmetric nature and it seems adequate fitting a joint model with the SN distribution to the data set. Along with this consideration, the following two statistical models with different distributions of random errors for both the response model (B.13) and the covariate model (B.11) are employed to compare their performance:

- **Model I:** A model with independent multivariate normal distributions of random errors for both the covariate model and the response model;
- **Model II:** A model with independent multivariate SN distributions of random errors for both the covariate model and the response model.

We will investigate the following two scenarios. Firstly, we investigate how an SN distribution for model error contributes to the efficiency of parameter estimation in comparison with a normal distribution, which is a special case of the SN distribution with zero skewness. Secondly, we also estimate the model parameters by using the 'naive' method, which does not separate the measurement errors from the true CD4 values. That is, the 'naive' method only uses the observed CD4 values z_{ij} rather than true (unobservable) CD4 values z_{ij}^* in the response model (B.13). Thus we use it 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.

To carry out a Bayesian inference, we need to specify the values of the hyperparameters in the prior distributions. We assume weakly informative prior distributions for all the parameters. In particular, (i) fixed-effects are taken to be independent normal $N(0, 100)$ for each component of the population parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. (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 $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$ are taken to be inverse Wishart distributions $IW(\boldsymbol{\Omega}_1, \nu_1)$ and $IW(\boldsymbol{\Omega}_2, \nu_2)$, where the degree of freedom $\nu_1 = \nu_2 = 4$, and $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are diagonal matrices with diagonal elements being 0.01. (iv) For each of the skewness parameters δ_e and δ_ϵ, we choose an independent normal distribution $N(0, 100)$, where we specify that $\boldsymbol{\delta}_{e_i} = \delta_e \mathbf{1}_{n_i}$ and

$\boldsymbol{\delta}_{\epsilon_i} = \delta_\epsilon \mathbf{1}_{m_i}$ to indicate that we are interested in skewness of both overall viral load data and overall CD4 cell count data.

The MCMC sampler is implemented using WinBUGS software (Lunn et al., 2000), and the program codes are available in Appendix B, where the initial values were chosen from previous studies (Liu and Wu, 2007; Wu, 2002). In particular, the MCMC scheme for drawing samples from the full conditional posterior distributions of all parameters in both the response and covariate models is obtained by iterating between the following two steps: (i) the Gibbs sampler is used to update $\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_b, \delta_\epsilon, \delta_e$; (ii) we update $\boldsymbol{b}_i$ and $\boldsymbol{a}_i$ $(i = 1, 2, \cdots, n)$ using the Metroplis-Hastings (M-H) algorithm. After convergence was achieved using standard tools within WinBUGS (such as trace plots), we retain the final MCMC samples for making statistical inference for the unknown parameters. After an initial 50,000 burn-in iterations, every 40th MCMC sample is retained from the next 400,000. Thus we obtain 10,000 samples from the posterior distributions of the unknown parameters for further statistical inference. See 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.

3.3 Comparison of modeling results

The SN-BSNLME joint modeling approach in conjunction with the NLME response model (B.13) and the covariate model (B.11) with different distribution specifications for the random errors was used to fit the viral load and CD4 data simultaneously. Table 1 presents the population posterior mean (PM), the corresponding standard deviation (SD) and 95% credible interval for fixed-effects parameters. The following findings are observed for the estimated results. For parameter estimates of the response model, (i) for both models the coefficient parameter β_2 of time t has posterior means which are positive and their corresponding 95% credible intervals (CI) do not contain zero. (ii) For the coefficient parameter β_3, where the corresponding covariate (true CD4 values) is interacted with time t, the posterior mean for Model I is smaller than that based on Model II. The results indicate that the covariate CD4 effect (β_3) may be underestimated if a normal distribution is assumed. (iii) For the scale parameter σ^2, the posterior mean (0.57) for Model II is much smaller than the estimate (1.87) of Model I. For parameter estimates of the covariate model, (i) the estimate of intercept α_1 based on Model I is larger in absolute value than that based on Model II; however, the estimates of the coefficients α_2 and α_3 are comparable for both Models. (ii) For the scale parameter τ^2, the estimated value (0.08) based on Model II is smaller than that (0.13) based on Model I.

From the model fitting results, we have seen that, in general, both Model I and Model II provided a reasonably good fit to the observed data for most patients in our study, although the fitting for a few patients (<7%) was not completely satisfactory due to unusual viral load fluctuation patterns for these patients, particularly for Model I. To assess the goodness-of-fit of the proposed models, Figure 4 presents the diagnosis plots

Model		α_1	α_2	α_3	β_1	β_2	β_3	β_4	σ^2	τ^2	δ_e	δ_ϵ
I(JM)	PM	-0.21	0.64	-0.26	8.26	25.9	7.20	2.19	1.87	0.13	–	–
	L_{CI}	-0.46	0.13	-0.80	7.72	19.2	0.03	-0.24	1.49	0.10	–	–
	U_{CI}	0.03	1.15	0.27	8.79	33.4	15.4	4.33	2.38	0.13	–	–
	SD	0.13	0.26	0.28	0.28	3.58	3.86	1.17	0.24	0.15	–	–
II(JM)	PM	-0.21	0.67	-0.33	8.13	22.2	9.63	2.68	0.57	0.08	1.86	0.24
	L_{CI}	-0.49	0.12	-0.94	7.57	16.2	3.59	0.57	0.01	0.03	0.52	0.06
	U_{CI}	0.06	1.26	0.25	8.68	28.3	17.5	4.53	1.98	0.14	2.54	0.54
	SD	0.14	0.29	0.31	0.28	3.11	3.58	1.00	0.53	0.03	0.68	0.26
II(NM)	PM	–	–	–	8.08	17.2	3.47	-0.34	0.41	–	2.08	–
	L_{CI}	–	–	–	7.54	12.7	0.96	-0.45	0.01	–	0.39	–
	U_{CI}	–	–	–	8.66	22.4	7.99	3.50	1.89	–	2.61	–
	SD	–	–	–	0.29	2.63	2.38	1.23	0.47	–	0.59	–

Table 1: Summary of estimated posterior means (PM) of population (fixed-effects), scale and skewness parameters, corresponding standard deviation (SD) and lower limit (L_{CI}) and upper limit (U_{CI}) of 95% equal-tail credible intervals (CI) from the four models based on the joint modeling(JM) approach and the 'naive' method (NM) associated with the model II.

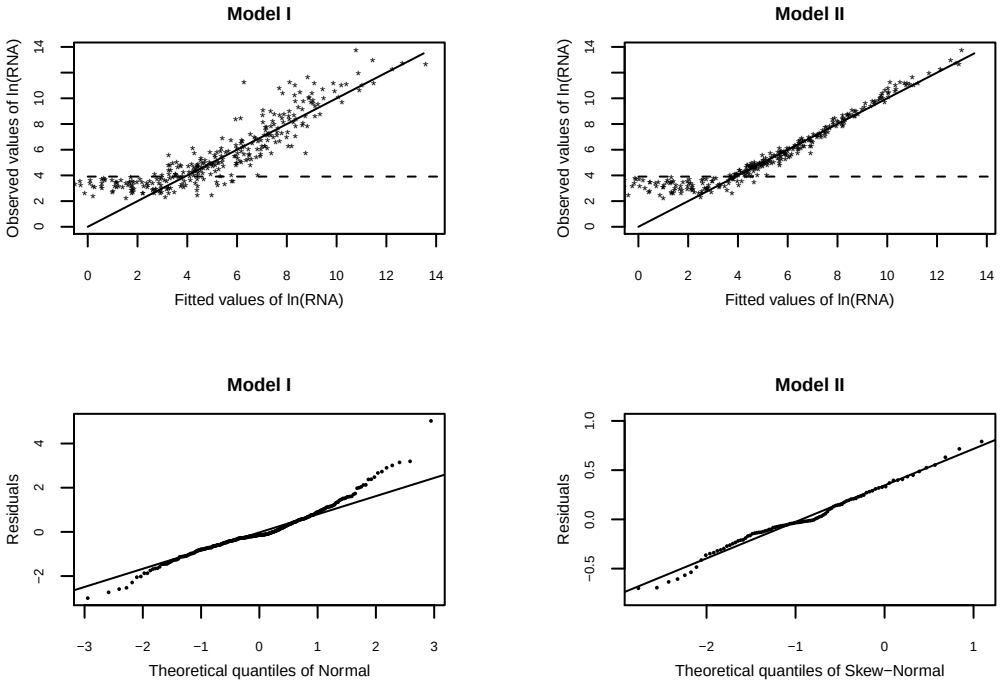

Figure 4: Goodness-of-fit: Observed values versus fitted values of ln(RNA) (top panel) and SN or normal Q-Q plots with line (bottom panel).

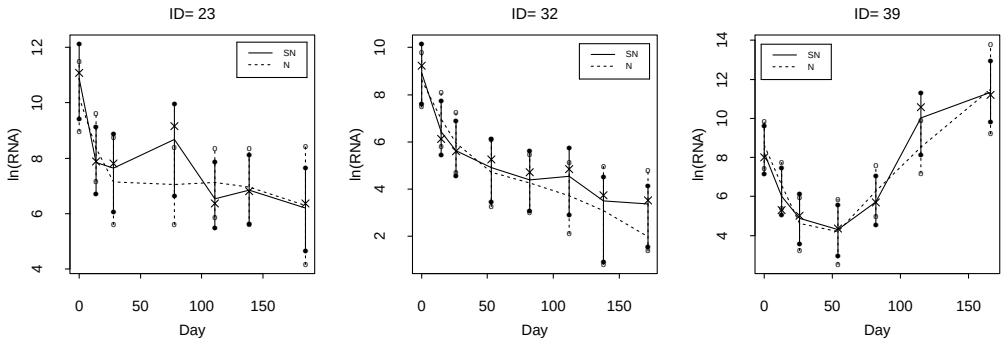

Figure 5: The individual fitted curves of viral load for three randomly selected patients based on the joint models with a normal (dotted line) or SN (solid line) random error. The respective vertical dotted line (normal) ended with '$\circ$' and solid line (SN) ended with '$\bullet$' on each fitted value are the 95% credible interval (CI) associated with the fitted value. The observed values are indicated by cross sign ($\times$).

of the observed values versus the fitted values (upper panel) and SN or normal Q-Q plots (lower panel) from Models I and II. It can be seen from Figure 4 (upper panel) that the SN model provided a better fit to observed data, compared with the normal model. This result can be also explained by examining the SN or normal Q-Q plots of the residuals (lower panel) where both plots show the existence of outliers, but it is clearly seen that Model II only has a few negative outliers and, thus, fits observed data better than Model I. This finding is further confirmed by their standardized residual sums of squares (RSS) which are 45.56 (SN random error) and 279.67 (normal random error).

Figure 5 displays three randomly selected individual estimates of viral load trajectories along with the associated 95% confidence interval on each fitted value obtained by using the joint modeling approach based on Models I (normal) and II (SN). The following findings are observed from the joint modeling results. (i) The estimated individual trajectories for Model II fit the originally observed values more closely than those for Model I. (ii) Overall, the 95% CI associated with predicted values from Model I is wider than the corresponding 95% CI from Model II. (iii) All the 95% CIs from Model II cover the observed viral load values, while some of 95% CIs from Model I do not; for example, for patient 39 whose observed value at day 112 is 10.57, the corresponding 95% CI from Model II is (8.78, 11.18) with the fitted value 10.04, while the corresponding 95% CI from Model I is (6.67, 9.83) with the fitted value 8.17 which does not cover the observed value 10.57.

For selecting the better 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 two models with the specification of different distributions for the random error, DIC can be used to find out how assumption of an SN distribution contributes to virologic responses and parameter estimation in comparison with that of a normal distribution. We find that the DIC value (1172.90) for Model I (with normal random error) is larger than that (673.34) for Model II (with SN random error). As mentioned before, it is hard to tell which model is 'correct' but which one fits data better. Furthermore, the model which fits data better may be more accurate to describe the mechanism of HIV infection and CD4 changing process, and thus needs more attention for patient treatment. Therefore, based on the DIC criterion, the results indicate that Model II is the better fitting model, supporting the contention of a departure from normality. These results are consistent with those in diagnosis of the goodness-of-fit displayed in Figure 4 indicating that Model II outperforms Model I. In summary, our results may suggest that it is very important to assume an SN distribution for the response model and the CD4 covariate model in order to achieve reliable results, in particular if the data exhibit skewness. Along with these observations, we will further report our findings in details only for the better Model II in Section 3.4.

3.4 Estimation results based on Model II

The estimated results presented in Table 1 based on a better model (Model II) indicate that the population CD4 trajectory may be approximated by the quadratic polynomial $\hat{z}(t) = 228.57(-0.21 + 0.67t - 0.33t^2) + 375.46$, where $z(t)$ is in the original CD4 scale. Figure 6 shows the estimated first- and second-phase viral decay rates of change ($\hat{\lambda}_1$ and $\hat{\lambda}_2(t)$) in viral load and their correlation relationship. Thus, the population viral load process may be approximated by $\hat{V}(t) = \exp(8.13 - \hat{\lambda}_1 t) + \exp(2.68 - \hat{\lambda}_2(t)t)$. Since the first-phase viral decay rate (λ_1) is significantly associated with the true CD4 values (due to the statistically significant estimate of β_3), this suggests that the viral load change $V(t)$ may be significantly associated with the true CD4 values. Note that, although the true association described above may be complicated, the simple approximation considered here may provide a rough guidance and point to further research.

The analysis results suggest that the first-phase and second-phase viral decay rates are always positive and negative, interactively, and they show a significantly negative correlation ($r = -0.976$ with p-value $p = 0.0089$). The results in Figure 6 indicate that the first-phase (the second-phase) decay rate increases (decreases) at the early stage and then decreases (increases) at the late stage. This finding is biologically meaningful and may reflect the viral load trajectory shown in Figure 1(b) on the rapid decay phase and then a slow growth phase. The true CD4 process has a significantly positive effect on the first-phase viral decay rate; this finding confirms that the CD4 covariate may be a significant predictor on the first-phase viral decay rate during the process. More rapid increase in CD4 cell count may be associated with faster viral decay in the early stage. This may be explained by the fact that higher CD4 cell count suggests a higher turnover rate of lymphocyte cells, which may cause a positive correlation between viral decay and the CD4 cell count. In addition, the posterior means (0.57) of the within-subject scale parameter (σ^2) based on Model II is much smaller than that (1.87) based on Model

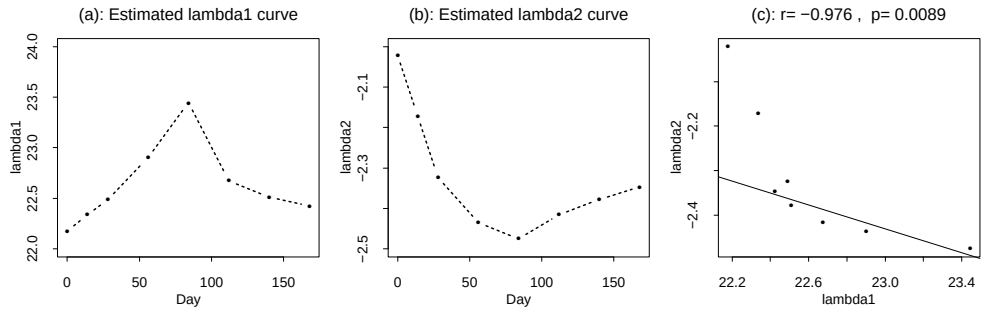

Figure 6: The estimated curves of the first and second phase viral decay rates and their correlation. The solid line in (c) is a robust (MM-estimator) linear regression fit. The correlation coefficient (r) and p-values are obtained from a Spearman rank correlation test.

I; it indicates that gain in significant efficiency for the skew-normal model relative to the normal model is observed for the scale estimation. This is expected because high variability, heaviness of the tails and the skewness are interrelated to a certain extent.

For the estimates of the skewness parameters, we found that the response model skewness parameter δ_e based on Model II is estimated to be significantly positive and fairly large (1.86). This confirms the positive skewness of the viral load as shown in Figure 2 (left panel). We also check the residual distributions and plot density estimates of residuals for both models (plots not shown here). It was indicated that the results coincide with our assumption of residual distributions. That is, the residual appears to be symmetrically distributed for Model I, while the residual follows a fairly positive skew-distribution for Model II. The estimate of the covariate model skewness parameter δ_ϵ is 0.24 based on Model II which is consistent with the significantly positive skewness of the CD4 cell count (see the right panel in Figure 2). Thus, it may suggest that accounting for skewness is required to model the data when the data exhibit skewness.

We further compare two methods for estimation based on Model II: the proposed joint modeling method and the 'naive' method where the raw (observed) CD4 values z_{ij}, rather than the true (unobservable) CD4 values z^*_{ij}, are substituted in the model (B.13). This naive method ignores measurement errors in CD4 values and treats the observed CD4 as true values. The results of the naive method associated with Model II are shown in Table 1. It can be seen that the estimates of the parameters β_1 and β_2 are similar for the two methods. However, there are important differences in the estimates for the parameters β_3 and β_4. The naive method, which ignores measurement errors, may substantially underestimate the covariate CD4 effect. The joint modeling method appears to give larger standard deviations (SD) for the model parameters, probably because it incorporates the variation from fitting the CD4 process. Thus, the difference of the naive estimates and the joint modeling estimates, due to whether or not we ignore potential CD4 measurement errors, indicates that CD4 measurement errors can not be ignored in the analysis.

4 Discussion

For longitudinal data with heavy tail characteristics of viral load response and CD4 covariate, we have developed a general SN-BSNLME joint model with a skew-normal distribution and measurement errors in covariates that may be preferred over those with a normal distribution or 'naive' method in the sense that it produces more reliable parameter estimates. The proposed method may have a significant impact on AIDS research because, in the presence of skewness in the data and measurement errors in covariates, appropriate statistical inference is important for making robust conclusions and reliable clinical decisions. We believe that, to the best of our knowledge, this is the first attempt in working on such general distributional structure for SN-BSNLME models. Our proposed method is quite general and so can be used in other applications. This kind of skew-normal modeling approach is important in many biostatistical application areas, allowing accurate inference of parameters while adjusting for the data with skewness.

The foregoing results indicate that in a two-phase HIV dynamic model, the analysis results suggest that there may be a significantly positive relation between the first-phase viral decay and the covariate CD4 values. This finding is consistent with those reported by Liu and Wu (2007). Our result may be partially explained by the fact that the higher CD4 value suggests a higher turnover rate of hymphocyte cells. This result is very interesting and clinically important. Since the viral decay rates may reflect the efficacy of antiretroviral treatment, the higher CD4 value may need less potent drug efficacy to suppress virus replication so that a more potent drug regimen may not be necessary to avoid side-effects of drug use. This also confirms the fact from the modeling point of view that more rapid increase in CD4 cell count may be associated with faster viral decay, whereas more rapid decrease in CD4 cell count may be associated with earlier viral rebound. These findings may help improve understanding of the pathogenesis of HIV infection and evaluation of antiretroviral treatments.

The results indicate that with the skew-normality 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 the 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. This paper combined new technologies in mathematical modeling and statistical inference with advances in HIV/AIDS dynamics 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. We also notice that both the skew-normal and skew-t distributions are in the class of skew-elliptical distributions. Thus one may consider a skew-t distribution as an alternative in this study. An associate editor pointed out that the study of the models separately considers occasion variation and assay error. While interesting, this issue requires additional efforts and more data information. These problems, however, are beyond the focus of this article, but a further study may be warranted. We are actively investigating these problems, and hope that we can report

these interesting results in the near future.

References

Acosta, E. P., Wu, H., Walawander, A., Eron, J., Pettinelli, C., Yu, S., Neath, D., Ferguson, E., Saah, A. J., Kuritzkes, D. R., and Gerber, J. G. (2004). "Comparison of two indinavir/ritonavir regimens in treatment-experienced HIV-infected individuals." *Journal of Acquired Immune Deficiency Syndromes*, 37: 1358–1366.

Arellano-Valle, R. B. and Azzalini, A. (2006). "On the Unification of Families of Skew-normal Distributions." *Scandinavian Journal of Statistics*, 33: 561–574.

Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2007). "Bayesian Inference for Skew-normal Linear Mixed Models." *Journal of Applied Statistics*, 34: 663–682.

Azzalini, A. and Capitanio, A. (1999). "Statistical Applications of the Multivariate Skew Normal Distributions." *Journal of Royal Statistical Society, Series B*, 61: 579–602.

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*. Chapman and Hall: London., 2nd edition.

D., S., 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 Statistics Society, Series B*, 64: 583–639.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall: London.

Gelman, A., Carlin, J. B., Stern, S., H, and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall: London., 2nd edition.

Huang, Y. and Dagne, G. (2010). "Skew-normal Bayesian Nonlinear Mixed-effects Models with Application to AIDS Studies." *Statistics in Medicine*, 29: 2384–2398.

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.

Laveille, M., Sanson, A., Fermin, A. K., and Mentre, F. (2011). "Maximium Likelihood Estimation of Long-term HIV Dynamic Models and Antiviral Response." *Biometrics*, 67: 250–259.

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., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002). "Bayesian Analysis of Population PK/PD Models: General Concepts and Software." *Journal of Pharmacokinetics and Pharmacodynamics*, 29: 271–307.

Lunn, D. J., Thomas, A., Best, N., and Spiegehalter, D. (2000). "WinBUGS – a Bayesian Modelling Framework: Concepts, Structure, an Extensibility." *Statistics and Computing*, 10: 325–337.

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika*, 63: 581–592.

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.

Tobin, J. (1958). "Estimation of Relationships for Limited Dependent Variables." *Econometrica*, 26: 24–36.

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, H., Ding, A. A., and De Gruttola, V. (1998). "Estimation of HIV dynamic parameters." *Statistics in Medicine*, 17: 2463–2485.

Wu, H. and Zhang, J. T. (2002). "The study of long-term HIV dynamics using semiparametric nonlinear mixed-effects models." *Statistics in Medicine*, 21: 3565–3675.

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.

Acknowledgments

The authors thank the Editor, an Associate Editor and one anonymous referee for their insightful comments and suggestions that led to a marked improvement of the article. We gratefully acknowledge A5055 study investigators for allowing us to use the clinical data from their study. This research was partially supported by NIAID/NIH grant AI080338 to Huang and by NIMH grant R01MH040859-22 to Dagne.

Appendix 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 are 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., 2006, 2007; Azzalini et al., 1996, 1999; 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 the SN distribution and its potential applications as well as differences among various versions of SN distributions, see References listed above.

In this work, 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 a conditional method. In particular, it is relatively easy to implement and provides an interesting alternative to other computationally challenging parametric or nonparametric models. 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. An m-dimensional random vector $\boldsymbol{Y}$ follows an m variate SN distribution with location vector $\boldsymbol{\mu}$, $m \times m$ positive (diagonal) dispersion matrix $\boldsymbol{\Sigma}$ and $m \times m$ skewness matrix $\boldsymbol{\Delta}(\boldsymbol{\delta}) = diag(\delta_1, \delta_2, \ldots, \delta_m)$, if its probability density function (pdf) is given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = 2^m |\boldsymbol{A}|^{-1/2} \phi_m[\boldsymbol{A}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})|\boldsymbol{I}_m] \Phi_m[\boldsymbol{\Delta}(\boldsymbol{\delta})\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})|\boldsymbol{I}_m - \boldsymbol{\Delta}(\boldsymbol{\delta})\boldsymbol{A}^{-1}\boldsymbol{\Delta}(\boldsymbol{\delta})],$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)^T$ is a skewness parameter vector, $\boldsymbol{A} = \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2(\boldsymbol{\delta})$, $\phi_m(\boldsymbol{y}|\boldsymbol{V})$ and $\Phi_m(\boldsymbol{y}|\boldsymbol{V})$ denote the pdf and the cumulative distribution function (cdf), respectively, of $N_m(\boldsymbol{0}, \boldsymbol{V})$. We denote this by $\boldsymbol{Y} \sim SN_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}(\boldsymbol{\delta}))$. The mean and covariance matrix are given by $E(\boldsymbol{Y}) = \boldsymbol{\mu} + \sqrt{2/\pi}\boldsymbol{\delta}$, $cov(\boldsymbol{Y}) = \boldsymbol{\Sigma} + (1 - 2/\pi)\boldsymbol{\Delta}^2(\boldsymbol{\delta})$. An appealing feature of the pdf $f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$ is that it gives an independent marginal when $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$. This pdf thus reduces to

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = \prod_{i=1}^m \left[\frac{2}{\sqrt{\sigma_i^2 + \delta_i^2}} \phi \left\{ \frac{y_i - \mu_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right\} \Phi \left\{ \frac{\delta_i}{\sigma_i} \frac{y_i - \mu_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right\} \right],$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. In order to have a zero mean vector, we should assume the location parameter $\boldsymbol{\mu} = -\sqrt{2/\pi}\boldsymbol{\delta}$, which is what we assume in this paper. By Proposition 1 of Arellano-Valle et al.(2007), the SN distribution of $\boldsymbol{Y}$ has a convenient stochastic representation as follows:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta}(\boldsymbol{\delta})|X_0| + \boldsymbol{\Sigma}^{1/2} X_1,$$

where X_0 and X_1 are two independent $N_m(\boldsymbol{0}, \boldsymbol{I}_m)$ random vectors. Note that the expression above provides a convenient device for random number generation and for implementation. Let $\boldsymbol{w} = |X_0|$; then $\boldsymbol{w}$ follows an m-dimensional standard normal distribution $N_m(\boldsymbol{0}, \boldsymbol{I}_m)$ truncated in the space $\boldsymbol{w} > \boldsymbol{0}$ (i.e., the standard half-normal

distribution). Thus, following Sahu et al.(2003), a hierarchical representation of the expression above is given by

$$\boldsymbol{Y}|\boldsymbol{w} \sim N_m(\boldsymbol{\mu} + \boldsymbol{\Delta}(\boldsymbol{\delta})\boldsymbol{w}, \boldsymbol{\Sigma}),\ \boldsymbol{w} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)\boldsymbol{I}(\boldsymbol{w} > \boldsymbol{0}).$$

It is noted that when $\boldsymbol{\delta} = \boldsymbol{0}$, the SN distribution reduces to usual normal distribution. To better understand the shape of an SN distribution, plots of the univariate SN density as a function of the skewness parameter can be found in (Huang and Dagne, 2010).

Appendix B. WinBUGS code for Model II: SN-BSNLME joint models

```
## Variables in the dataset
# y[,1] = serial number
# y[,2] = arm
# y[,3] = time(day)
# y[,4] = id
# y[,5] = rna
# y[,6] = cd4
# y[,7] = logerna
# y[,8] = cij (censoring indicator)
# y[,9] = cd4 (standardized)
# y[,10]= time(day) (rescaled time between 0 and 1 dividing by max)
# Z[,1:3] (base functions)
# Begin of model
model
{
for (i in 1:n)
{
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],Omega2[,])
a[i,1:3]~dmnorm(a3[i,1:3],Omega3[,])
}
for (j in 1 : N)
{
# (1) Modelling CD4 via measurement errors model with SN
z.star[j]<-(alpha[1]+a[y[j,4],1])+(alpha[2]+a[y[j,4],2])*y[j,10]+
(alpha[3]+a[y[j,4],3])*y[j,10]*y[j,10]+delta2*(w2[j]-0.798)
w2[j]~ dnorm(0,1)I(0,)
y[j,9]~ dnorm(z.star[j],tau2)
# (2) SNLME response model with SN incorporating covariate measurement error
```

```
betai1[j] <- beta[1]+ b[y[j,4],1]
betai2[j] <- beta[2]+beta[3]*z.star[j]+b[y[j,4],2]
betai3[j] <- beta[4]+ b[y[j,4],3]
betaij4[j]<- mu.not[1]+mu.not[2]*Z[j,2]+mu.not[3]*Z[j,3]+b[y[j,4],4]
dm1[j]<- betai1[j]-step(betai2[j]-betaij4[j])*betai2[j]*y[j,10]
dm2[j]<- betai3[j]-step(betai2[j]-betaij4[j])*betaij4[j]*y[j,10]
dm3[j]<- exp(dm1[j])
dm4[j]<- exp(dm2[j])
dm5[j]<- dm3[j] +dm4[j]
upper.limit[j]<-Below.detection*y[j,8]+upper.bound*(1-y[j,8])
mu[j]<-log(dm5[j]) + delta*(w[j]-0.798)
w[j]~ dnorm(0, 1 ) I(0,)
y[j,7]~ dnorm(mu[j], tau)I(,upper.limit[j])
♯ Residuals
fit[j]<-mu[j]
resid[j]<-y[j,7]-fit[j]
}
♯ Prior distributions of the hyper-parameters
♯ (1) Coefficients
for (l in 1:4){beta[l]~ dnorm(0,1.0E-2)}
for (l in 1:3){mu.not[l]~ dnorm(0,1.0E-2)
alpha[l]~ dnorm(0,1.0E-2)}
♯ (2) Covariance matrix in random effects
Omega2[1:4,1:4] ~ dwish(R2[,],4)
v2[1:4,1:4]<- inverse(Omega2[,])
Omega3[1:3,1:3] ~ dwish(R3[,],3)
v3[1:3,1:3]<- inverse(Omega3[,])
♯ (3) Skewness parameters
delta ~ dnorm(0.0, 0.01)
delta2 ~ dnorm(0.0, 0.01)
♯ (4) Precision parameters
tau ~ dgamma(0.01,0.01)
sigma.tau<- 1/tau
tau2 ~ dgamma(0.01,0.01)
sigma.tau2 <- 1/tau2
} ♯ End of model
♯ Data inputed
list(n=44,N=310, Below.detection=3.912, upper.bound=500,
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)))
```