Statistics and Its Interface Volume 11 (2018) 223–236

Bayesian-frequentist hybrid approach for skew-normal nonlinear mixed-effects joint models in the presence of covariates measured with errors

Gang Han, Yangxin Huang*, and Ao Yuan

It is a common practice to analyze complex longitudinal data using nonlinear mixed-effects (NLME) models. Existing methods often assume a normal model for the errors, which is not realistic. To explain between- and within-subject variations, covariates are usually introduced in such models to partially explain inter-subject variations, but some covariates may often be measured with substantial errors. Moreover, although statistical methods for analyzing longitudinal data have been evolving substantially, existing methods are either frequentist or full Bayesian, not taking into account scenarios where only part of the parameters have sound prior information available. In an attempt to take full advantage of both approaches, we adopt a Bayesian-frequentist hybrid (BFH) approach to NLME models with a skew-normal distribution in the presence of covariate measurement errors and jointly model the response and covariate processes. We illustrate the proposed method in a real example from an AIDS clinical trial by modeling the viral dynamics to compare potential models with different inference methods. Simulation studies are conducted to assess the performance of the proposed model and method.

AMS 2000 subject classifications: Primary 62F15; secondary 62P10.

Keywords and phrases: Bayesian-frequentist hybrid approach, Longitudinal data, Measurement errors, Mixed-effects models, Skew-normal distribution.

1. INTRODUCTION

Longitudinal data analysis methods have been the subject of extensive studies, and they have been successfully applied to various practical problems. However, there are still at least three questions to address: First, existing methods are either frequentist or Bayesian and both have their limitations in practice. For parameter estimation under general conditions, both methods are asymptotically equivalent, in that they have the same convergence rate with the same weak limit [23, 16]. But their finite sample size properties are different, and each method has its advantages and disadvantages: The former is simpler in modeling and often has computational advantage but cannot implement prior knowledge into the model. Also, the frequentist method requires slightly more conditions for estimation consistency than those required by the Bayesian method [25]. Some frequentist analyses of HIV viral load could lead to unpredictable convergence problems during numerical computing, e.g., [26, 30]. On the other hand, the Bayesian method can incorporate prior information into the model, and any admissible procedure can be formulated as either a Bayesian or a limit of Bayesian procedures [28], including the frequentist maximum likelihood estimate (MLE). It can gain small sample advantage when the prior information is sound, but could also be misleading when the prior information is not justified even if the non-informative prior is used. This is because the estimation from data of small sample sizes may not have enough data information to swap away the effect of the prior, and more or less advisably or inadvisably affected by the prior. Moreover, the full Bayesian inference is typically more time consuming due to numerical methods such as the Markov chain Monte Carlo (MCMC) algorithms, which may also generate additional approximation errors.

Second, most of the existing methods assume normal random errors for convenience [8, 14, 18, 30, 32]. This requires the variables to be "symmetrically" distributed. A violation of this assumption could lead to misleading inferences [3, 9, 11, 10, 15, 24, 27]. As shown in Figure 1(a), in fact, the observed viral load data (in log scale) in HIV studies are often far from being "symmetric."

Third, the validity of inference methods relies on an important requirement that variables are "perfectly" measured. In practice, however, collected data are often far from "perfect". Measurement error in model covariates is another typical feature of the HIV viral load data and ignoring this phenomenon may result in biased statistical inference. As an example, Figure 1(b) displays trajectories of longitudinal viral load measured from RNA levels in plasma (in $\log_{10}$ scale) in an AIDS clinical trial study indicating that the between- and within-subject variations appear to be large [17, 1]. To partially explain these variations, covariates such as CD4 counts are usually introduced in such models to explain inter-subject variations. CD4 counts are often measured with substantial errors. As a result, joint modeling of

*Corresponding author.

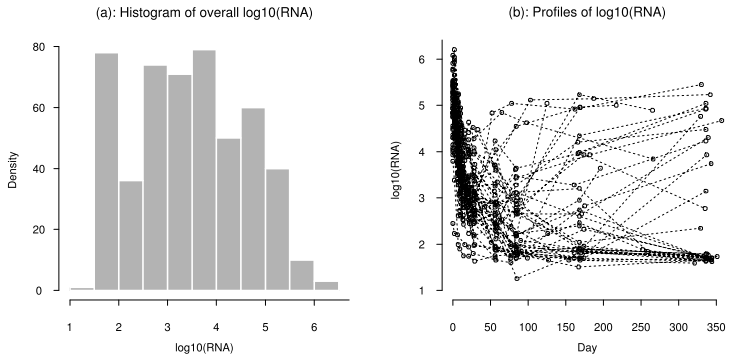

Figure 1. Histogram of longitudinal viral load measurements and trajectories of viral load measured from RNA levels in plasma (in $\log_{10}$ *scale) from 48 patients in an AIDS clinical trial study.*

CD4 cell counts and viral load is required to account for measurement errors and random effects from both sources.

To address the aforementioned issues, in this article we propose a Bayesian-frequentist hybrid (BFH) inference method for skew-normal nonlinear mixed-effects joint models of the viral load and CD4 cell count. Specifically, as a rule of thumb, for estimation of a d-dimensional parameter, it requires the sample size to be at least $10 \times d$ for the regression coefficients to be unbiased [7], and 30×2^d for the validity of asymptotic representation following the Bonferroni's rule, which can be difficult to achieve. The finite sample behavior of the inference is thus important. In practice, we have situations in which we have sound prior information on only part of the parameters, and no such information on the rest parameters. In this case, a full frequentist method will ignore the valuable information on part of the parameters, while a full Bayesian method requires prior on all the parameters, which may yield misleading results, due to the prior on the rest parameters for which we don't have sound prior information (even if with a non-informative prior on this part of parameters). Thus, to deal with the above situation, an approach that takes advantage of the two procedures and avoids their weakness is desirable. As a solution, we adopt a Bayesian-frequentist hybrid (BFH) approach proposed by [34] to nonlinear mixed-effects (NLME) models with a skew-normal (SN) distribution joint with a covariate model with measurement errors. As an extension of [34], who proposed a hybrid estimate for independent response variable-based regression models, we propose to jointly investigate the NLME model with an SN distribution for the response process and the linear mixed-effects (LME) model with an SN distribution for covariate measurement error process.

It is worthwhile to note that the BFH approach can not be replaced by the full Bayesian analysis with non-informative priors on parameters with no objective prior information. As an example, for n Bernoulli observations of parameter p and the sum of X, if we have prior $\pi(p)$ for p, then the posterior distribution of p will be proportional to $p^X(1-p)^{n-X} \times \pi(p)$. There can be potential bias for non-informative prior where $\pi(p)$ follows standard uniform distribution. For example, if $n = 6$ and $X = 1$ then the unbiased MLE of p is $1/6$. But with the uniform prior, the posterior of p is $Beta(2, 6)$. If we use the posterior mean as the estimate, then the Bayes estimate is $1/4$. The bias introduced by the uniform prior can be calculated as $1/4 - 1/6 = 1/12 \approx 8.33\%$.

In Section 2, we describe the data set that motivated this research and investigate SN-NLME joint models for HIV dynamics. The associated BFH inferential approach to obtain hybrid estimate (HE) is presented in Section 3, and the model implementation using a Monte Carlo EM algorithm is introduced in Section 4. In Section 5, we demonstrate the proposed method by applying it to the AIDS data described in Section 2 and report the analysis results. In Section 6 we conduct a simulation study to examine and compare the finite sample performances of the BFH and frequentist methods under the normal and skew-normal models.

2. THE DATA AND JOINT MODELS FOR HIV DYNAMICS

2.1 Data description

The AIDS clinical trial study in [17] consists of 53 HIV-1 infected patients who were treated by an antiretroviral regimen. Five patients who dropped out earlier and never returned to the study were excluded from the data analysis. The plasma HIV-1 RNA (viral load) is repeatedly quantified on days 0, 2, 7, 10, 14, 21, 28, 56, 84, 168 and 336 of follow-up after initiation of treatment and trajectories of viral load from 48 patients were depicted in Figure 1(b). The number of measurements for each individual varies from 7 to 11. The covariate CD4 cell count was also measured throughout the study on a similar scheme. A $\log_{10}$-transformation of viral load was used in the analysis to stabilize the variations of observations and speed up algorithm of estimation. In addition, to avoid too small or large estimates that may be unstable, we standardize the time-varying (covariate) CD4

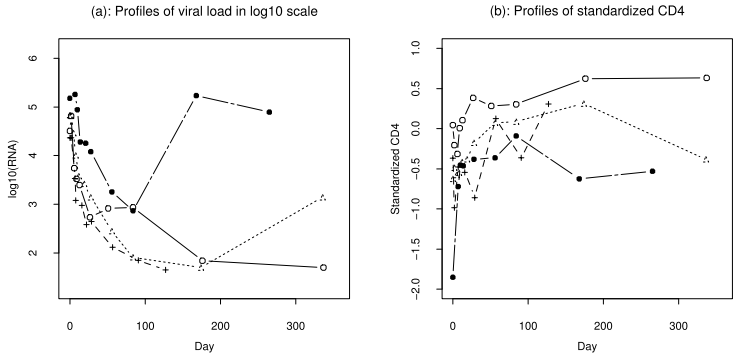

Figure 2. Profiles of viral load (response) in $\log_{10}$ *scale and standardized CD4 (covariate) for four randomly selected patients. The horizontal line is below the detectable level of viral load (2=*$\log_{10}(100)$*).*

cell count and rescale the original time t (in days) so that the time scale is between 0 and 1.

Figure 2 shows the measurements of HIV viral load in $\log_{10}$ scale and CD4 cell count for four randomly selected patients. All trajectories of viral load and CD4 cell count exhibit distinctive and important patterns throughout the time course. The rate change in viral load appears to vary substantially across patients, reflecting both biological variation and systematic associations with subject-level covariates. The detailed descriptions of the study and data can be found in [17].

2.2 Skew-normal NLME joint models for HIV dynamics

As shown in Appendix B, nonlinear mixed-effects models based on the two-compartment equation (B.2) are powerful tools for modeling HIV viral dynamics [29] and offer almost equal performance to capture the early segment of viral load trajectories [30]. One of our objectives in this paper is to investigate how the extended equation (B.3) performs when the complete viral load data including viral rebound are employed for modeling. We next develop such model structure and the associated inferential method.

Denote the number of subjects by n and the number of longitudinal measurements on the ith subject by n_i. Let $y_{ij} = y_{ij}(t_{ij})$ denote the $\log_{10}$-transformation of the viral load value $V(t_{ij})$ for individual i at time t_{ij} $(i = 1, \ldots n; j = 1, \ldots, n_i)$. It was noted that the viral decay rates may vary over time because they depend on some phenomenologic parameters that hide considerable microscopic complexity and change over time [20]. Negative values of the decay rates may correspond to viral increase and lead to viral rebound. [29] suggested that variation in the dynamic parameters may be partially associated with CD4 cell count and other covariates. For the viral load response process, we consider the following NLME model with an SN distribution incorporating possibly time-varying CD4 covariate with measurement errors:

(1)
$$
\begin{aligned}
y_{ij} &= \log_{10}(V(t_{ij})) = \log_{10}(e^{p_{1i}-\lambda_{1i}t_{ij}} + e^{p_{2i}-\lambda_{2ij}t_{ij}}) + e_{ij} \\
p_{1i} &= \beta_1 + b_{1i}, \quad \lambda_{1i} = \beta_2 + \beta_3 z_{i0} + b_{2i} \\
p_{2i} &= \beta_4 + b_{3i}, \quad \lambda_{2ij} = \beta_5 + \beta_6 z^*_{ij} + b_{4i}
\end{aligned}
$$

where z^*_{ij} is a summary of the true (but unobservable) CD4 covariate value at time t_{ij} (see below in detail) and z_{i0} is the baseline CD4 cell count; $\exp(p_{i1}) + \exp(p_{i2})$ is the baseline viral load; λ_{i1} and λ_{ij2} are the first- and second-phase viral decay rates, respectively; $\boldsymbol{\beta}_{ij} = (p_{i1}, p_{i2}, \lambda_{i1}, \lambda_{ij2})'$ is a vector of individual parameters for the i^{th} subject at time t_{ij} and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)'$ is a vector of population parameters; the vector of model errors $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})' \overset{\text{iid}}{\sim} SN_{n_i}\left(-\sqrt{2/\pi}\delta_1 \mathbf{1}_{n_i}, \sigma_1^2 \boldsymbol{I}_{n_i}, \delta_1 \boldsymbol{I}_{n_i}\right)$, which follows a multivariate SN distribution with unknown variance parameter σ_1^2, and skewness parameter δ_1, where $\mathbf{1}_{n_i} = (1, \ldots, 1)'$; the vector of random-effects $\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})' \overset{\text{iid}}{\sim} N_4(\mathbf{0}, \boldsymbol{\Sigma}_b)$, and $\boldsymbol{\Sigma}_b$ is covariance matrix.

Models for the covariate process are needed to incorporate measurement errors in covariates. Following the study by [33], we extend a linear mixed-effects (LME) model using the SN distribution for the CD4 process. In the absence of theoretical rationales, low-order polynomial LME models may be considered and standard model selection methods can be used to choose the best model. For simplicity, we consider a single time-varying covariate. Let z_{ij} be the observed covariate value for individual i at time t_{ij} $(i = 1, \ldots n; j = 1, \ldots, n_i)$. In the presence of covariate measurement errors, we consider the following covariate LME model with an SN distribution.

(2)
$$
\begin{aligned}
z_{ij} &= \boldsymbol{u}'_{ij}\boldsymbol{\alpha} + \boldsymbol{v}'_{ij}\boldsymbol{a}_i + \epsilon_{ij} \quad (\equiv z^*_{ij} + \epsilon_{ij}), \\
\boldsymbol{\epsilon}_i &\overset{\text{iid}}{\sim} SN_{n_i}\left(-\sqrt{2/\pi}\delta_1 \mathbf{1}_{n_i}, \sigma_2^2 \boldsymbol{I}_{n_i}, \delta_2 \boldsymbol{I}_{n_i}\right),
\end{aligned}
$$

where $z^*_{ij} = \boldsymbol{u}'_{ij}\boldsymbol{\alpha} + \boldsymbol{v}'_{ij}\boldsymbol{a}_i$ may be viewed as the true (but unobservable) CD4 covariate value at time t_{ij}; $\boldsymbol{u}_{ij} = \boldsymbol{u}_{ij}(t_{ij})$ and $\boldsymbol{v}_{ij} = \boldsymbol{v}_{ij}(t_{ij})$ are $l \times 1$ design vectors; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)'$

and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{il})'$ are unknown population (fixed-effects) and individual-specific (random-effects) parameter vectors, respectively. The random-effects $\boldsymbol{a}_i$, which are introduced to account for large inter-individual variations in the CD4 process, follow the multivariate normal distribution $N_l(\mathbf{0}, \boldsymbol{\Sigma}_a)$, where $\boldsymbol{\Sigma}_a$ is covariance matrix. We assume $\boldsymbol{e}_i, \boldsymbol{b}_i, \boldsymbol{\epsilon}_i$ and $\boldsymbol{a}_i$ are independent of each other.

Specifically, we consider the covariate model (2) with $\boldsymbol{u}_{ij} = \boldsymbol{v}_{ij} = (1, t_{ij}, \ldots, t_{ij}^{l-1})'$ and focus on linear ($l = 2$), quadratic ($l = 3$) and cubic ($l = 4$) polynomials to choose the best model based on AIC and BIC values. The resulting AIC (BIC) values are 890.0 (931.7), 773.4 (801.3) and 845.2 (887.1), respectively. Thus, we adopt the following quadratic polynomial LME model for the CD4 process:

$$z_{ij} = (\alpha_1 + a_{i1}) + (\alpha_2 + a_{i2})t_{ij} + (\alpha_3 + a_{i3})t_{ij}^2 + \epsilon_{ij}, \tag{3}$$

where $z_{ij}^* = (\alpha_1 + a_{i1}) + (\alpha_2 + a_{i2})t_{ij} + (\alpha_3 + a_{i3})t_{ij}^2$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)'$ is a vector of population (fixed-effects) parameters and individual-specific random-effects $\boldsymbol{a}_i = (a_{i1}, a_{i2}, a_{i3})' \overset{\text{iid}}{\sim} N_3(\mathbf{0}, \boldsymbol{\Sigma}_a)$.

3. BAYESIAN-FREQUENTIST HYBRID ESTIMATE

In a longitudinal study, such as the AIDS study described in Section 2, 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 with non-normality and measurement error may be favorable, the computation associated with the joint likelihood inference in such models with SN distribution for longitudinal data can be extremely intensive and, particularly, may lead to convergence problems [33]. Here, we propose a hybrid inferential method for the response model (1) joint with the covariate model (3) to estimate all parameters simultaneously.

3.1 Hybrid estimate given a loss function

Let $\mathbf{y}_i = (y_{i,1}, ..., y_{i,n_i})'$, $\mathbf{z}_i = (z_{i,0}, z_{i,1}, ..., z_{i,n_i})'$, $\mathbf{Y}^n = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ and $\mathbf{Z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)'$. Following the discussion in [24], we implement a modeling procedure to the joint models by introducing two $n_i \times 1$ random vectors $\boldsymbol{w}_{i,1}$ and $\boldsymbol{w}_{i,2}$ based on the stochastic representation for the SN distribution (see Appendix A in detail). It can be shown that $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ can be hierarchically formulated as follows:

$$\begin{aligned}
&\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{w}_{i,1}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_1^2, \delta_1 \\
&\quad \sim N_{n_i}\left(\boldsymbol{g}_i(\boldsymbol{t}_i, \boldsymbol{\beta}_i) + \delta_1[\boldsymbol{w}_{i,1} - \sqrt{2/\pi}\mathbf{1}_{n_i}], \sigma_1^2 \boldsymbol{I}_{n_i}\right), \\
&\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{w}_{i,2}; \boldsymbol{\alpha}, \sigma_2^2, \delta_2 \\
&\quad \sim N_{n_i}\left(\boldsymbol{z}_i^* + \delta_2[\boldsymbol{w}_{i,2} - \sqrt{2/\pi}\mathbf{1}_{n_i}], \sigma_2^2 \boldsymbol{I}_{n_i}\right), \\
&\boldsymbol{w}_{i,1} \sim N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i}) I(\boldsymbol{w}_{i,1} > \mathbf{0}),
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\boldsymbol{w}_{i,2} \sim N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i}) I(\boldsymbol{w}_{i,2} > \mathbf{0}), \\
&\boldsymbol{b}_i \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}_b),\ \boldsymbol{a}_i \sim N_3(\mathbf{0}, \boldsymbol{\Sigma}_a),
\end{aligned}$$

where $I(\boldsymbol{w} > 0)$ is an indicator function and $\boldsymbol{w}$ follows normal distribution $N_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i})$ truncated in the space $\boldsymbol{w} > 0$; $\boldsymbol{z}_i^* = (z_{i,1}^*, \cdots, z_{in_i}^*)'$, $\boldsymbol{g}_i(\boldsymbol{t}_i, \boldsymbol{\beta}_i) = (g(t_{i1}, \boldsymbol{\beta}_{i1}), \cdots, g(t_{in_i}, \boldsymbol{\beta}_{in_i}))'$ with $g(t_{ij}, \boldsymbol{\beta}_{ij}) = \log_{10}(e^{p_{1i} - \lambda_{1i} t_{ij}} + e^{p_{2i} - \lambda_{2ij} t_{ij}})$.

Our model has many more parameters than typical longitudinal data models. The parameters of interest in general may be partitioned into two subsets, one with prior knowledge, and one without. In this case, a Bayesian analysis on one subset will have benefits from the prior information, while for the other subset a frequentist method is preferred. Thus, a Bayesian-frequentist hybrid method [34] for the inference of all parameters can have advantages over the full frequentist or Bayesian method.

For simplification, we assume that the components of the random-effects $\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4})'$ are independent with each other having distribution $b_{ik} \sim N(0, \sigma_{bk}^2)$ $(k = 1, 2, 3, 4)$. Similarly, we assume $a_{ik} \sim N(0, \sigma_{ak}^2)$ $(k = 1, 2, 3)$. Let $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2$ be the collection of all unknown population parameters in our joint model, where $\boldsymbol{\theta}_1 = \{\alpha_3, \beta_1, \beta_2, \beta_4, \sigma_{a1}^2, \sigma_{a2}^2, \sigma_{a3}^2, \sigma_{b1}^2, \sigma_{b2}^2, \sigma_{b3}^2, \sigma_{b4}^2\}$ and $\boldsymbol{\theta}_2 = \{\alpha_1, \alpha_2, \beta_3, \beta_5, \beta_6, \sigma_1^2, \sigma_2^2, \delta_1, \delta_2\}$. Based on past studies [18, 9, 10, 6], we have prior information on $\boldsymbol{\theta}_1$ summarized in the prior density $\pi(\boldsymbol{\theta}_1)$, but we do not have such information for $\boldsymbol{\theta}_2$, which are parameters of interest and the estimates are supposed to be dominated by data. Thus, we implement the BFH approach to estimate $\boldsymbol{\theta}_1$ using the Bayesian method and $\boldsymbol{\theta}_2$ the frequentist method simultaneously.

Denote observed data $\mathcal{D} = \{\boldsymbol{Y}^n, \boldsymbol{Z}^n, \{t_{ij}\}_{i=1,\ldots,n; j=1,\ldots,n_i}\}$. Let $f(\cdot)$, $f(\cdot|\cdot)$ and $\pi(\cdot)$ be a generic density function, a conditional density function, and a prior density function, respectively. In concise notation, the likelihood for the observed data is

$$f(\mathbf{Y}^n, \mathbf{Z}^n|\boldsymbol{\theta}) = f(\mathbf{Y}^n, \mathbf{Z}^n|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \prod_{i=1}^{n} f(\mathbf{y}_i, \mathbf{z}_i|\boldsymbol{\theta}),$$

where

$$\begin{aligned}
f(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) = &\int\int\int\int f(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{w}_{i,1}; \boldsymbol{\theta}) \\
&\times f(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{w}_{i,2}; \boldsymbol{\theta}) f(\boldsymbol{w}_{i,1}|\boldsymbol{w}_{i,1} > \mathbf{0}) \\
&\times f(\boldsymbol{w}_{i,2}|\boldsymbol{w}_{i,2} > \mathbf{0}) f(\boldsymbol{a}_i) f(\boldsymbol{b}_i) d\boldsymbol{w}_{i,1} d\boldsymbol{w}_{i,2} d\boldsymbol{a}_i d\boldsymbol{b}_i,
\end{aligned} \tag{5}$$

with the density functions in the above integrand given in (4). Let $l(\boldsymbol{\theta}|\mathbf{Y}^n, \mathbf{Z}^n) = \log f(\mathbf{Y}^n, \mathbf{Z}^n|\boldsymbol{\theta})$ denote the log-likelihood. Thus, the joint posterior density of $\boldsymbol{\theta}$ based on the observed data $\mathcal{D}$ can be written as

$$f(\boldsymbol{\theta}|\mathcal{D}) \propto \left\{ \prod_{i=1}^{n} \int\int\int\int f(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{w}_{i,1}; \boldsymbol{\theta}) \right. \tag{6}$$

$$\times f(\boldsymbol{z}_i|\boldsymbol{a}_i,\boldsymbol{w}_{i,2};\boldsymbol{\theta})$$
$$\times f(\boldsymbol{w}_{i,1}|\boldsymbol{w}_{i,1}>\boldsymbol{0})f(\boldsymbol{w}_{i,2}|\boldsymbol{w}_{i,2}>\boldsymbol{0})$$
$$\left.\times f(\boldsymbol{a}_i)f(\boldsymbol{b}_i)d\boldsymbol{w}_{i,1}d\boldsymbol{w}_{i,2}d\boldsymbol{a}_i d\boldsymbol{b}_i\right\}\pi(\boldsymbol{\theta}_1).$$

Recall that the frequentist estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_n = \arg\sup_{\boldsymbol{\theta}} f(\mathbf{Y}^n,\mathbf{Z}^n|\boldsymbol{\theta}) = \arg\sup_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}^n,\mathbf{Z}^n).$$

Given a loss function $L(\boldsymbol{\theta},\boldsymbol{d})$ and a prior $\pi(\boldsymbol{\theta})$, the generalized Bayes estimate $\check{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is

$$\check{\boldsymbol{\theta}}_n = \arg\inf_{\boldsymbol{d}} \int L(\boldsymbol{\theta},\boldsymbol{d})f(\mathbf{Y}^n,\mathbf{Z}^n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Then given a loss $L(\boldsymbol{\theta}_1,\boldsymbol{d})$ and a prior $\pi(\boldsymbol{\theta}_1)$ only on $\boldsymbol{\theta}_1$, for general model with independently and identically distributed observations, the hybrid estimate (HE) $\boldsymbol{\theta}_n = (\check{\boldsymbol{\theta}}_{1,n},\hat{\boldsymbol{\theta}}_{2,n})'$ originally discussed by [34] is

$$\begin{aligned}\boldsymbol{\theta}_n &= (\check{\boldsymbol{\theta}}_{1,n},\hat{\boldsymbol{\theta}}_{2,n})'\\ &= \arg(\inf_{\boldsymbol{d}},\sup_{\boldsymbol{\theta}_2})\int L(\boldsymbol{\theta}_1,\boldsymbol{d})f(\mathbf{Y}^n,\mathbf{Z}^n|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1.\end{aligned}$$

The estimation regarding $\boldsymbol{\theta}_n$ for case-model control genetic data under the commonly used squared error loss, absolute error loss and the 0-1 loss were discussed in [36]. With the 0-1 loss, the HE is computationally the simplest and parallel to that for MLE. Here, for the hierarchical joint model (4), we focus on the 0-1 loss on $\boldsymbol{\theta}_1$. With this loss, the HE of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)'$ is given by

$$\begin{aligned}(7)\quad \boldsymbol{\theta}_n &= (\check{\boldsymbol{\theta}}_{1,n},\hat{\boldsymbol{\theta}}_{2,n})'\\ &= \arg\sup_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}\left\{l(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{Y}^n,\mathbf{Z}^n)+\log\pi(\boldsymbol{\theta}_1)\right\}.\end{aligned}$$

Note that the BFH inference for HE considered here is an extension of [34] who considered the HE of the parameter $\boldsymbol{\theta}$ for independent response variable-based models. We consider an HE of the parameter $\boldsymbol{\theta}$ for longitudinal response and covariate variables-based hierarchical joint models.

3.2 Asymptotic properties

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{1,0},\boldsymbol{\theta}_{2,0})$ be the true parameter generating the observed data, $(\boldsymbol{y},\boldsymbol{z})$ be a generic independent copy of the $(\boldsymbol{y}_i,\boldsymbol{z}_i)$'s, and it can have dimensions n_i $(i=1,...,n)$. Let $l^{(2)}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z},n_i) = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z},n_i)/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$ for $\mathbf{y}$ and $\mathbf{z}$ with dimension n_i, and $I(\boldsymbol{\theta}|n_i) = -E_{\boldsymbol{\theta}}l^{(2)}(\boldsymbol{\theta}|\mathbf{y},\mathbf{z},n_i)$ be the Fisher information about $\boldsymbol{\theta}$ when $\mathbf{y}$ and $\mathbf{z}$ are of dimension n_i. We assume the following conditions:
(C0). $\lim_{n\to\infty}\max_{1\le i\le n} n_i < \infty$.
(C1). There is a convex set A such that $\inf_{\boldsymbol{\theta}\in A}|I(\boldsymbol{\theta})| > 0$, where $I(\boldsymbol{\theta})$ is given in the Proposition, that $(\boldsymbol{\theta}_{1,0},\boldsymbol{\theta}_{2,0})\in A$ and that $(\check{\boldsymbol{\theta}}_{1,n},\hat{\boldsymbol{\theta}}_{2,n})\in A$ for all large n.
(C2). On A, $0<\pi(\cdot)<\infty$.
(C3). The first and second derivatives, $\pi^{(k)}(\cdot)$ $(k=1,2)$, are bounded and away from zero on A,
(C4). $I(\cdot)$ is continuous at $\boldsymbol{\theta}_0$.
(C5). In a neighborhood of $\boldsymbol{\theta}_0$, $\partial\int\int\int f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})d\mathbf{y}d\mathbf{z}/\partial\boldsymbol{\theta} = \int\int\partial f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}d\mathbf{y}d\mathbf{z}$.

The following property of $\boldsymbol{\theta}_n$ for the hierarchical joint model (4) is comparable with that in [36] for case-control genetics data model (see proof in Appendix C).

Proposition 1. *Assume conditions (C0)-(C5). Then under the 0-1 loss, we have*

$$(i)\qquad (\check{\boldsymbol{\theta}}_{1,n},\hat{\boldsymbol{\theta}}_{2,n})\to(\boldsymbol{\theta}_{1,0},\boldsymbol{\theta}_{2,0})(a.s.), \text{ and}$$
$$(ii)\qquad \sqrt{n}(\check{\boldsymbol{\theta}}_{1,n}-\boldsymbol{\theta}_{1,0},\hat{\boldsymbol{\theta}}_{2,n}-\boldsymbol{\theta}_{2,0})\xrightarrow{D}N(0,I^{-1}(\boldsymbol{\theta}_0)),$$

where $I(\boldsymbol{\theta}_0) = -\sum_{j=1}^r E_{\boldsymbol{\theta}_0}l^{(2)}(\boldsymbol{\theta}_0|\mathbf{y},\mathbf{z},k_j)p_j$, $k_1,...,k_r$ *are all the different numbers the* n_i*'s take, and the* p_j*'s are the corresponding frequencies.*

In the above the notation $(\check{\boldsymbol{\theta}}_{1,n}-\boldsymbol{\theta}_{1,0},\hat{\boldsymbol{\theta}}_{2,n}-\boldsymbol{\theta}_{2,0})$ is understood in the column vector sense (not an array of two column vectors). When the prior is fixed, the effect of the prior will be asymptotically negligible, such hybrid estimator is asymptotically first order equivalent to the maximum likelihood estimate (MLE) and efficient, but their finite sample properties are different, the hybrid estimator can outperform the MLE due to the helpful prior information. Below we give an example on how to use Proposition 1. We only need to illustrate the computation of $I(\boldsymbol{\theta}_0)$.

Example. Assume that the observed data are $(\boldsymbol{y}_i,\boldsymbol{z}_i)$ with $dim(\boldsymbol{y}_i) = dim(\boldsymbol{z}_i) = n_i$, $(i=1,...,1000)$. For $n=1000$, the n_i's only take $r=10$ different numbers such that $(k_1,\ldots,k_{10}) = (3,5,6,7,8,9,10,11,12,15)$, with multiplicities $(m_1,...,m_{10})$, and $\sum_{j=1}^r m_j = n$. Let $(p_1,...,p_{10}) = (m_1,...,m_{10})/n$ be the corresponding frequencies. For $(\boldsymbol{y}_i,\boldsymbol{z}_i)$'s with $n_i = k_j$, their density function is $f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta},k_j)$, with log-likelihood $l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z},k_j)$ and the corresponding Hessen matrix $l^{(2)}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z},k_j)$ for any $i\in\{1,\ldots,1000\}$ and $j\in\{1,\ldots,r\}$. Thus, in this case

$$I(\boldsymbol{\theta}_0) = -\sum_{j=1}^r E_{\boldsymbol{\theta}_0}l^{(2)}(\boldsymbol{\theta}_0|\mathbf{y},\mathbf{z},k_j)p_j,$$

which is well approximated, for large n, by

$$(8)\qquad \hat{I}(\hat{\boldsymbol{\theta}}) = -\sum_{j=1}^r\frac{1}{m_j}\sum_{n_i=k_j}l^{(2)}(\boldsymbol{\theta}_0|\mathbf{y}_i,\mathbf{z}_i,k_j)\frac{m_j}{n}.$$

3.3 Asymptotically informative prior

In the full Bayesian setting, when the prior is collected from prior data of size m with the ratio of sample sizes $m/n\to c$ for some $c\in(0,\infty)$, the prior can be asymptotically informative, and the effect of the prior will be asymptotically non-negligible [35]. Here we propose an asymptotically informative prior for the case of BFH model with

the 0-1 loss. Below we give a review of the notion of asymptotically informative prior in the full Bayesian setting.

In classical Bayes the prior is fixed (or treated as fixed). So it is washed out as the current data sample increases to infinity. In the asymptotically informative prior (AIP), the prior information grows together with the data sample size so that AIP can remain effective in the limit. That is why the Bayes estimate based on AIP can be more efficient than the classical Bayes. In some cases the prior itself is AIP, but the classical Bayes approach does not recognize this, still treats it as fixed, and thus ignores such information in the asymptotic results. In this situation, the classical Bayes and our method with the AIP should be the same, but the former does not acknowledge the nature of the prior. For example, the prior information is summarized from some previous similar study with data sample size 2000, and we have current data with sample size 1000. The classical Bayes method regards the prior from the results of 2000 data as being fixed and computes the asymptotic result, while data with sample size 1000 is used to approximately interpret the asymptotic result. Here the prior from 2000 data is ignored, while the size of 1000 is treated as "infinity", and the prior effect is washed out so that the analysis does not take advantage of the prior efficiency. With the AIP method proposed in [35], we attempt to recover such efficiency. So it is not really our claim that the proposed method is "better" than the classical Bayes regarding the prior. Also, prior information from previous studies with large sample size may not be formulated in a proper way to get an AIP by classical Bayes, thus the classical Bayes method may not be able to recover such prior information in the asymptotics.

We let $\pi_m(\cdot)$ denote the informative prior density. We index the prior with an integer m to represent the sample size from prior similar studies about $\boldsymbol{\theta}$. In practice, $\pi_m(\cdot)$ is constructed using existing inferential results based on datasets generated from the parameter, but not on the current observed data. Let $h_m(\boldsymbol{\theta}_1) = \log \pi_m(\boldsymbol{\theta}_1)$. As in [35], we define the asymptotically informative prior on $\boldsymbol{\theta}_1$ as below.

Definition. $\pi_m(\boldsymbol{\theta}_1)$ is an *asymptotically informative prior* (AIP), if $h_m^{(2)}(\cdot)$ exists for all m, and as $m \to \infty$,

$$\frac{1}{m} h_m^{(1)}(\boldsymbol{\theta}_{1,0}) \xrightarrow{a.s.} \mathbf{0},$$
$$m^{-1/2} h_m^{(1)}(\boldsymbol{\theta}_{1,0}) \xrightarrow{D} N(\mathbf{0}, J(\boldsymbol{\theta}_{1,0})),$$
and
$$\frac{1}{m} h_m^{(2)}(\boldsymbol{\theta}_1) \xrightarrow{a.s.} -J(\boldsymbol{\theta}_1)$$

on any compact set, for some $d \times d$ matrix $J(\boldsymbol{\theta}_1)$, nonnegative definite and componentwise continuous on that set, with $d = dim(\boldsymbol{\theta}_1)$.

Two remarks for this definition are worth noting. One, in some cases, the AIP can be constructed from existing independent parameter estimates by a general density estimator that may not be explicitly associated with some integer m. In this case we can simply modify the above definition as: let $h(\boldsymbol{\theta}_1) = \log \pi(\boldsymbol{\theta}_1)$. $\pi(\cdot)$ is an AIP, if

$$\frac{1}{n} h^{(1)}(\boldsymbol{\theta}_{1,0}) \xrightarrow{a.s.} \mathbf{0},$$
$$n^{-1/2} h^{(1)}(\boldsymbol{\theta}_{1,0}) \xrightarrow{D} N(\mathbf{0}, cJ(\boldsymbol{\theta}_{1,0})),$$
and
$$\frac{1}{n} h^{(2)}(\boldsymbol{\theta}_1) \xrightarrow{a.s.} -cJ(\boldsymbol{\theta}_1)$$

for some $0 \le c < \infty$ and some matrix $J(\boldsymbol{\theta})$ which is nonnegative definite and componentwise continuous on some compact set. This second definition includes the first one by setting $c = \lim_n m/n$, and including any fixed prior by $c = 0$. But we are mainly interested in the case $c \neq 0$.

Two, $\pi_m(\cdot)$ can sometimes be formulated as a multivariate exponential family; i.e., $\pi_m(\boldsymbol{\theta}) = \exp\{m[\bar{\boldsymbol{\theta}}_m' T(\boldsymbol{\theta}) + B(\boldsymbol{\theta}) + C(\bar{\boldsymbol{\theta}}_m)]\}$, for some known differentiable functions $T(\cdot)$, $B(\cdot)$ and some known function $C(\cdot)$, where $\bar{\boldsymbol{\theta}}_m$ is a consistent estimator of $\boldsymbol{\theta}_0$ constructed from past estimators and is asymptotically normal, i.e. $\sqrt{m}(\bar{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, J^{-1}(\boldsymbol{\theta}_0))$, with $T(\cdot)$ and $B(\cdot)$ satisfying $T^{(1)}(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0) + o(1)$ and $B^{(1)}(\boldsymbol{\theta}_0) = -\boldsymbol{\theta}_0' T^{(1)}(\boldsymbol{\theta}_0) + o(1)$. Here $\bar{\boldsymbol{\theta}}_m$ can be viewed as a hyperparameter. For example, if $\bar{\boldsymbol{\theta}}_m$ is a consistent and asymptotically normal estimator of $\boldsymbol{\theta}_0$ constructed from existing results, with asymptotic variance matrix $J^{-1}(\boldsymbol{\theta}_0)$, then

$$\pi_m(\boldsymbol{\theta}) = (2\pi)^{-d/2} m^{d/2} |J(\bar{\boldsymbol{\theta}}_m)|^{1/2} \times \exp\{-\frac{m}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_m)' J(\bar{\boldsymbol{\theta}}_m)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_m)\}$$

is an AIP and an exponential family with $T(\boldsymbol{\theta}) = J(\bar{\boldsymbol{\theta}}_m)\boldsymbol{\theta}$, $B(\boldsymbol{\theta}) = -\boldsymbol{\theta}' J(\bar{\boldsymbol{\theta}}_m)\boldsymbol{\theta}/2$ and $C(\bar{\boldsymbol{\theta}}_m) = -\bar{\boldsymbol{\theta}}_m' J(\bar{\boldsymbol{\theta}}_m)\bar{\boldsymbol{\theta}}_m/2 + \frac{d}{2m}\log\frac{m}{2\pi} + \frac{1}{2m}\log|J(\bar{\boldsymbol{\theta}}_m)|$.

With the above notion of AIP $\pi_m(\boldsymbol{\theta}_1)$ on $\boldsymbol{\theta}_1$ only, we construct the hybrid estimator for our hierarchical joint model. With 0-1 loss and $\pi(\boldsymbol{\theta}_1)$ replaced by $\pi_m(\boldsymbol{\theta}_1)$, formula (7) can be written as

$$\begin{aligned}(9)\qquad \boldsymbol{\theta}_n &= (\check{\boldsymbol{\theta}}_{1,n}, \hat{\boldsymbol{\theta}}_{2,n})' \\ &= \arg \sup_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \{l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{Y}^n, \mathbf{Z}^n) + \log \pi_m(\boldsymbol{\theta}_1)\}.\end{aligned}$$

Here $\pi_m(\cdot)$ differs from the prior in the classical Bayesian setting where it changes along with m, and the latter can be viewed as a special case of the former in which the rate is zero at which $\pi_m(\cdot)$ concentrates toward $\boldsymbol{\theta}_{1,0}$. Note that when $\pi_m(\boldsymbol{\theta}_1)$ is an AIP only for $\boldsymbol{\theta}_1$, the corresponding $J(\boldsymbol{\theta}_1)$ is of dimension $dim(\boldsymbol{\theta}_1)$, we embed it into the matrix $J(\boldsymbol{\theta})$ with dimension $dim(\boldsymbol{\theta})$ as its upper-left block, and with other entries be zeros. Thus, $J(\boldsymbol{\theta})$ is of the form $J(\boldsymbol{\theta}) = \mathrm{diag}(J(\boldsymbol{\theta}_1), \mathbf{0})$. To study the asymptotic behavior of $\boldsymbol{\theta}_n$ in this case, we need the following condition.

(C6). $0 \le c = \lim_n m/n < \infty$.

Proposition 2. *Assume (C0)-(C6). Then under the 0–1 loss for $\boldsymbol{\theta}_1$, we have*

$$(i) \qquad \boldsymbol{\theta}_n \to \boldsymbol{\theta}_0 \quad (a.s.)$$
$$(ii) \qquad \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, [I(\boldsymbol{\theta}_0) + cJ(\boldsymbol{\theta}_0)]^{-1}),$$

where $I(\boldsymbol{\theta}_0)$ is given in Proposition 1.

The above asymptotic normality result includes the classical Bayes estimator in which $c = 0$ and includes the results of Proposition 1. When $c > 0$, since $J(\cdot)$ is non-negative definite, $[I(\boldsymbol{\theta}_0) + cJ(\boldsymbol{\theta}_0)]^{-1} \leq I^{-1}(\boldsymbol{\theta}_0)$ in the matrix non-negative definite sense, and $I^{-1}(\boldsymbol{\theta}_0)$ is the asymptotic variance matrix for the classical Bayesian estimator and MLE based on the likelihood only. As a result, the hybrid estimate with asymptotically informative prior can be more efficient than the classical Bayes estimate, as well as the MLE based only on $l(\cdot|\mathbf{Y}^n, \mathbf{Z}^n)$.

4. MODEL IMPLEMENTATION

We implement the proposed model using the Monte Carlo Expectation-Maximization (EM) algorithm as in [6]. Let

$$\boldsymbol{\gamma} = \{\{\boldsymbol{a}_i\}, \{\boldsymbol{b}_i\}, \{\boldsymbol{w}_{1,i}\}, \{\boldsymbol{w}_{2,i}\}\}_{i=1,\ldots,n}$$

denote all random effects. With the observations $\mathcal{D}$ and model parameters $\boldsymbol{\theta}$, the Bayesian density can be derived as

$$f(\mathcal{D}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}_1) = \frac{f(\mathcal{D}, \boldsymbol{\gamma}|\boldsymbol{\theta})}{f(\boldsymbol{\gamma}|\mathcal{D}, \boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}_1). \tag{10}$$

Note that this conditional density expression could be maximized using the EM algorithm. For example, the Bayesian density in [6] has the same format as (10) except for a different likelihood function and random effects. We maximize the Bayesian density by taking expectation with respect to $[\boldsymbol{\gamma}|\mathcal{D}, \boldsymbol{\theta}^{(k)}]$ on both sides of (10) and then maximizing

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}, \mathcal{D}) = \log \pi(\boldsymbol{\theta}_1) + \int \log(f(\mathcal{D}, \boldsymbol{\gamma}|\boldsymbol{\theta})) \times f(\boldsymbol{\gamma}|\mathcal{D}, \boldsymbol{\theta}^{(k)}) d\boldsymbol{\gamma}, \tag{11}$$

where $\boldsymbol{\theta}^{(k)}$ is the set of the values of all model parameters at one iteration in the EM algorithm. If we generate N_{sim} Monte Carlo realizations of $\boldsymbol{\gamma}$ given parameters $\boldsymbol{\theta}^{(k)}$, then the density (11) can be computed using Monte Carlo samples of $\boldsymbol{\gamma}$ conditional on $\boldsymbol{\theta}^{(k)}$

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}, \mathcal{D}) &= E_{[\boldsymbol{\gamma}|\mathcal{D}, \boldsymbol{\theta}^{(k)}]} (\log(f(\mathcal{D}|\boldsymbol{\gamma}, \boldsymbol{\theta})) + \log(f(\boldsymbol{\gamma}|\boldsymbol{\theta}))) \\ &\quad + \log \pi(\boldsymbol{\theta}_1) \\ &= \frac{1}{N_{sim}} \sum_{q=1}^{N_{sim}} \left[\sum_{i=1}^{n} \sum_{j=1}^{n_i} \{\log f(y_{ij}|\boldsymbol{z}_i, \boldsymbol{\gamma}_{i,q}; \boldsymbol{\theta}) \right. \\ &\quad \left. + \log f(z_{ij}|\boldsymbol{\gamma}_{i,q}; \boldsymbol{\theta})\} \right] \\ &\quad + \frac{1}{N_{sim}} \sum_{q=1}^{N_{sim}} \left[\sum_{i=1}^{n} f(\boldsymbol{\gamma}_{i,q}|\boldsymbol{\theta}) \right] + \log \pi(\boldsymbol{\theta}_1), \end{aligned} \tag{12}$$

where $\boldsymbol{\gamma}_{i,q} = \{\boldsymbol{a}_{i,q}, \boldsymbol{b}_{i,q}, \boldsymbol{w}_{1i,q}, \boldsymbol{w}_{2i,q}\}$ is the qth realization of the random effects conditional on $\boldsymbol{\theta}^{(k)}$. The density functions in (12) can be calculated as

- $f(y_{ij}|\boldsymbol{z}_i, \boldsymbol{\gamma}_{i,q}; \boldsymbol{\theta}) = f(y_{ij}|\boldsymbol{a}_{i,q}, \boldsymbol{b}_{i,q}, \boldsymbol{w}_{1i,q}; \boldsymbol{\theta})$ is the normal distribution density with mean $\log_{10}[\exp\{\beta_1 + \boldsymbol{b}_{i,q}(1) - t_{ij}(\beta_2 + \beta_3 z_{i0} + \boldsymbol{b}_{i,q}(2))\} + \exp\{\beta_4 + \boldsymbol{b}_{i,q}(3) - t_{ij}(\beta_5 + \beta_6 z^{\star}_{ij,q} + \boldsymbol{b}_{i,q}(4))\}] + \delta_1 \left[\boldsymbol{w}_{1i,q}(j) - \sqrt{2/\pi}\right]$, where $z^{\star}_{ij,q} = (\alpha_1 + \boldsymbol{a}_{i,q}(1)) + (\alpha_2 + \boldsymbol{a}_{i,q}(2))t_{ij} + (\alpha_3 + \boldsymbol{a}_{i,q}(3))t^2_{ij}$, and variance σ^2_1;
- $f(z_{ij}|\boldsymbol{\gamma}_{i,q}; \boldsymbol{\theta}) = f(z_{ij}|\boldsymbol{a}_{i,q}, \boldsymbol{w}_{2i,q}; \boldsymbol{\theta})$ is the normal distribution density with mean $z^{\star}_{ij,q} + \delta_2 \left[\boldsymbol{w}_{2i,q}(j) - \sqrt{2/\pi}\right]$ and variance σ^2_2;
- $f(\boldsymbol{\gamma}_{i,q}|\boldsymbol{\theta}) = f(\boldsymbol{a}_{i,q}|\boldsymbol{\theta}) f(\boldsymbol{b}_{i,q}|\boldsymbol{\theta}) f(\boldsymbol{w}_{1i,q}|\boldsymbol{\theta}) f(\boldsymbol{w}_{2i,q}|\boldsymbol{\theta})$, where $f(\boldsymbol{a}_{i,q}(k)|\boldsymbol{\theta})$ is the normal density with mean 0 and variance σ^2_{ak} for $k = 1, 2, 3$; $f(\boldsymbol{b}_{i,q}(k)|\boldsymbol{\theta})$ is the normal density with mean 0 and variance σ^2_{bk} for $k = 1, \ldots, 4$; each element in $\boldsymbol{w}_{1i,q}$ and $\boldsymbol{w}_{2i,q}$ follows standard normal distribution truncated to take non-negative values only.

The prior distribution $\pi(\boldsymbol{\theta}_1)$ and the initial parameter values shall be specified to start the first iteration of the algorithm. Some details about priors and initial parameters in the AIDS clinical data example are provided in Section 5. We run the following steps at each iteration.

1. Generate N_{sim} random effects $\boldsymbol{\gamma}_i$ for each $i \in \{1, \ldots, n\}$ using the model parameters at the current iteration $\boldsymbol{\theta}^{(k)}$;
2. E-step: compute (12);
3. M-step: update $\boldsymbol{\theta}$ value by maximizing (12).

Same as [6], we complete N_{em} iterations to achieve the final estimate $\boldsymbol{\theta}^{(N_{em})}$. We use equation (8) to estimate the Fisher information matrix and the formula in [21] to approximate the variance in the estimate. In brief, for a generic parameter θ, the estimated variance of θ can be written as

$$\widehat{Var}(\theta) = \left[E(l'') + E(l'^2) + (E(l'))^2\right]^{-1}, \tag{13}$$

where l is the log-likelihood, l' and l'' are the first and second order derivatives, respectively. We approximate the expectations in (13) using the simulated parameters $\boldsymbol{\theta}^{(k)}$ from the EM iterations.

5. ANALYSIS OF THE AIDS CLINICAL DATA

As described in Section 2, 48 patients were included in the analysis. Among them 1 patient had 7 longitudinal mea-

Table 1. Parameter estimates from the frequentist and BFH models based on the EM at the 50-th iteration. Numbers in the parentheses are estimated standard deviation for the estimates of $(\beta_1, \beta_2, \ldots, \beta_6)$.

Parameter	Freq. estimate (STD)	BFH estimate (STD)
$\widehat{\beta}_1$	8.785 (0.018)	9.217 (0.028)
$\widehat{\beta}_2$	8.956 (0.085)	24.821 (0.149)
$\widehat{\beta}_3$	3.365 (0.062)	10.766 (0.085)
$\widehat{\beta}_4$	3.505 (0.027)	3.812 (0.022)
$\widehat{\beta}_5$	-1.008 (0.053)	-2.724 (0.038)
$\widehat{\beta}_6$	1.008 (0.038)	1.001 (0.046)
$\widehat{\alpha}_1$	-0.293	-0.303
$\widehat{\alpha}_2$	3.190	3.262
$\widehat{\alpha}_3$	-2.599	-2.734
$\widehat{\sigma}_1^2$	1.443	1.288
$\widehat{\sigma}_2^2$	1.650	1.196
$\widehat{\sigma}_{a,1}^2$	0.615	0.250
$\widehat{\sigma}_{a,2}^2$	0.780	0.261
$\widehat{\sigma}_{a,3}^2$	0.580	0.208
$\widehat{\sigma}_{b,1}^2$	1.239	0.596
$\widehat{\sigma}_{b,2}^2$	5.640	3.351
$\widehat{\sigma}_{b,3}^2$	3.401	2.072
$\widehat{\sigma}_{b,4}^2$	25.662	20.505
$\widehat{\delta}_1$	-0.0051	0.0033
$\widehat{\delta}_2$	-0.0041	0.0056

surements, 6 patients had 9 measurements, 12 patients had 10 measurements, 27 patients had 11 measurements, and 2 patients had 12 measurements. So $r = 5$, $(k_1, \ldots, k_5) = (7, 9, 10, 11, 12)$, and $(m_1, \ldots, m_5) = (1, 6, 12, 27, 2)$ according to the notation in (8). In the analysis of this AIDS clinical data set, we choose the 0-1 loss for the BFH model because 1) it is the most commonly used loss function in the hybrid Bayesian literature [34, 6, 36], and 2) with the 0-1 loss, the algorithm complexity of BFH is equivalent to that in the frequentist analysis. We use the priors in [6] and [13] to construct the prior of $\pi(\boldsymbol{\theta}_1)$, where [6] constructed a Bayesian-frequentist hybrid model for $\boldsymbol{Y}^n$ with random effects $\{\boldsymbol{b}_i\}_{i=1,\ldots,n}$ only and [13] implemented a survival and longitudinal joint model involving the model parameters in (4). These priors are essentially AIP because they were solicited from existing studies. Specifically,

- $\beta_1 \sim N(8.4, 0.3^2)$, $\beta_2 \sim N(0.3^2)$, $\beta_4 \sim N(3.8, 0.3^2)$;
- $\alpha_3 \sim N(-2.91, 0.53)$;
- $\sigma_{b1}^2, \ldots, \sigma_{b4}^2$ follow inverse gamma distributions with means 1.43, 5.7, 3.89, and 24.54, respectively; $\sigma_{a1}^2, \ldots, \sigma_{a3}^2$ follow inverse gamma distributions with means 0.674, 0.658, and 0.553, respectively. All the variances are three times of the mean values.

Initial parameter values in our EM algorithm are the same as from the EM algorithm in [6] for $(\beta_1, \ldots, \beta_6$, and $\sigma_{b1}^2, \ldots, \sigma_{b4}^2)$, and estimates in [13] for all other model parameters. We set the number of simulation at each iteration (N_{sim}) to be 200, and the number of EM iteration (N_{em}) is set to 50.

Parameter estimates from the frequentist and BFH models are listed in Table 1. For the parameters describing the viral load $(\beta_1, \ldots, \beta_6)$, the estimated values from the BFH model are comparable with existing estimates in the literature; i.e., [6, 13]. Some frequentist estimates, however, converged to unreasonable values in the parameter space (e.g., β_2 and β_3). Most of the estimates of other parameters from the two models are similar except for $(\sigma_{b,1}^2, \ldots, \sigma_{b,4}^2)$ and (δ_1, δ_2). This can be due to the large number of model parameters and relatively small sample size. Figures 3 and 4 show the trace plots of $(\beta_1, \ldots, \beta_6)$ from the frequentist and BFH methods, respectively. We can see that the estimates converge to different values from the two methods. The Bayesian hybrid inference leads to more stable and reliable estimates even with a moderate number (i.e., 10) of model parameters. Similar to [6], we estimate the standard deviation of $(\beta_1, \ldots, \beta_6)$ from both models in the parentheses of Table 1. We can see that the estimated standard deviations are similar between the two models if the estimates are close to each other, but the estimated standard deviation from the BFH model can be smaller when the frequentist model has an unreasonable estimate. For example, the estimated β_4 and β_5 from the frequentist model are bigger than those in the BFH model. This is consistent with the conclusion of Proposition 2, and the findings in [6] that the Bayesian hybrid inference may not correspond to greater estimation uncertainty.

To compare the convergence of the two models, we make a trace plot of the objective function in Figure 5, where the objective function is the negative log likelihood in the

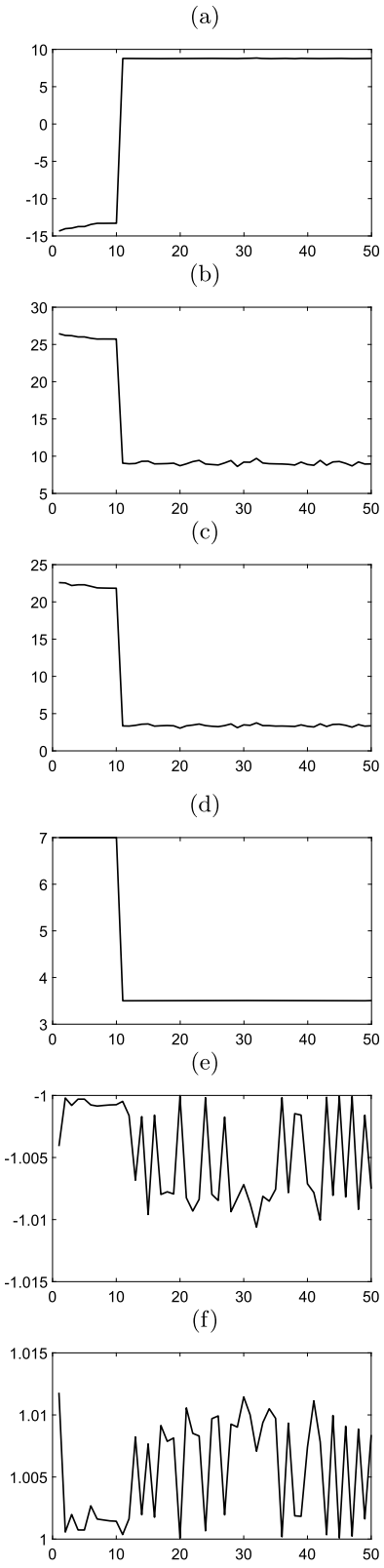

Figure 3. Trace plots of the estimates (a) β_1, (b) β_2, (c) β_3, (d) β_4, (e) β_5, (f) β_6 from the **frequentist model** *through 50 iterations.*

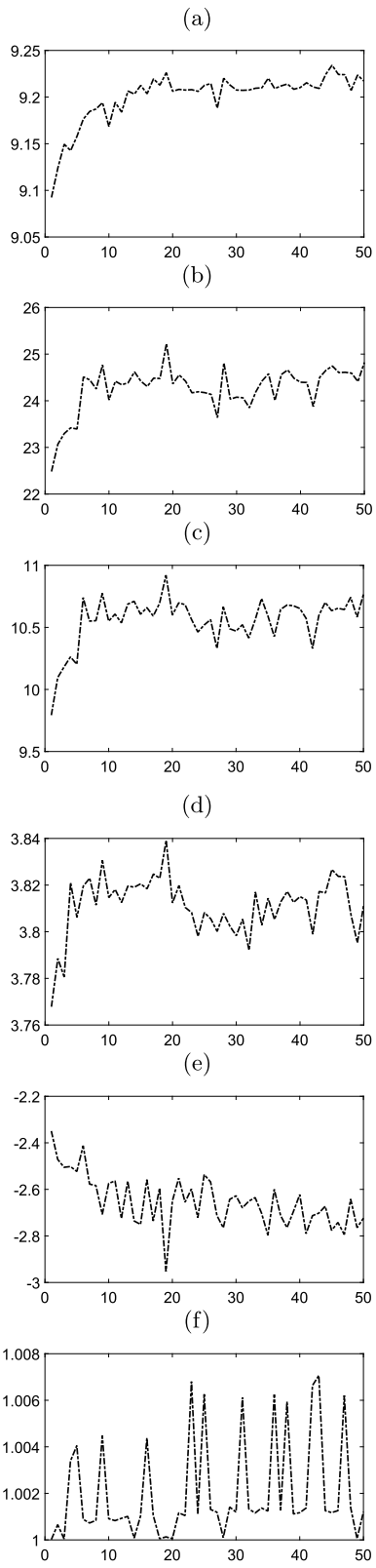

Figure 4. Trace plots of the estimates (a) β_1, (b) β_2, (c) β_3, (d) β_4, (e) β_5, (f) β_6 from the **BFH model** *through 50 iterations.*

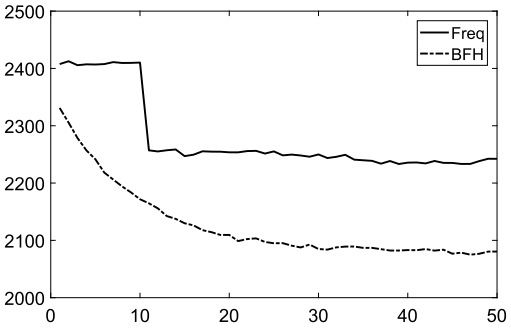

Figure 5. Trace plot of the objective function (to be minimized in the EM algorithm) from the frequentist model (solid curve) and BFH model (dashed dots) through the 50 iterations.

frequentist model, and the sum of negative log likelihood and negative prior density in the BFH model. The Bayesian hybrid inference can smoothly converge to the final value after about 10 iterations, while the objective function in the frequentist analysis had a drop at the 10th iteration, which is less stable than the iterations in the proposed BFH model. In this example, the BFH model had superior convergence properties to the frequentist alternative.

6. A SIMULATION STUDY COMPARING BFH AND FREQUENTIST METHODS

We further compare BFH and the frequentist methods, as well as the models with and without skewness in the measurement error. We generate a simulation data set for 48 patients at observation times 0, 1, 2, 3, 4, 5, 6, 7, 8, 12, 16, 20, and 24 (assumably in month). Values of the model parameters used to simulate the data are close to the BFH estimates from the real example in Table 1, except that $\sigma^2_{b,1}, \ldots, \sigma^2_{b,4}$ are set to a small value 0.001, and $\delta_2 = 0$. Same as the simulation studies in [2] and [19], we generate each random error e_{ij} in (1) as $\epsilon_{ij} - 2$, where ϵ_{ij} is an independent realization from gamma distribution with parameters 2 and 1. The initial values of the parameters are close to the real values in the simulation. Prior distributions are the same as in the real example in Section 5. Such simplified setting can facilitate the result interpretation and guarantee all methods converge in a few iterations typically less than 10.

Our simulation study compares four scenarios: A, Frequentist method with skewness parameter δ_1 set to 0, B, BFH with δ_1 set to 0; C, Frequentist with δ_1 to be estimated; D, BFH with δ_1 to be estimated. For C and D, we let the lower bound of δ_1 in the iteration to be 0.9 and the initial value to be 1 to guarantee some degree of skewness. In each scenario, the numerical integration/maximization is based on 200 Monte Carlo simulations, and the number of iteration is 10. We present the results as the estimates of $\beta_1, \ldots, \beta_6$ in Table 2 with the true and initial values. We can

Table 2. Estimates of $\beta_1, \ldots, \beta_6$ from Frequentist, no skewness (F. NS); BFH, no skewness (BFH NS); Frequentist, skewness (F. S); BFH, skewness (BFH S), and the true and initial values of the parameters.

	F. NS	BFH NS	F. S	BFH S	True;Initial
$\widehat{\beta}_1$	21.62	18.12	21.60	18.14	8; 8.2
$\widehat{\beta}_2$	54.17	23.16	53.92	23.14	25; 26.5
$\widehat{\beta}_3$	22.25	11.95	22.18	11.96	8.68; 8.5
$\widehat{\beta}_4$	8.78	9.02	8.78	8.99	3.8; 3.8
$\widehat{\beta}_5$	-5.56	-3.42	-5.56	-3.40	-2.34; -2.5
$\widehat{\beta}_6$	0.03	-1.51	0.01	-1.51	-0.073; 0
$\widehat{\delta}_1$	NA	NA	0.90	0.95	gamma(2,1); 1

see that in the estimation of $\beta_1, \ldots, \beta_3$, the two FBH scenarios are significantly more accurate than those from the frequentist method, while the models with or without the skewness are comparable. For β_4, all estimates have similar bias. In the estimation of β_5, BFH with the skewness parameter is better than that without, and the BFH estimates are both more accurate than the frequentist estimates. For β_6, the frequentist estimates have smaller predictive errors than the FBH estimates, but FBH estimates were able to maintain the correct negative slop. The relatively big bias in the estimates of β_1 and β_4 in all scenarios can be due to the systematic discrepancy between the simulation process and true model formula. In the estimation of δ_1 in C and D, we can see that the frequentist estimate stopped at the lower bound 0.9, while the FHB estimate converged to 0.95. After changing the lower bound to 0 for another run, the frequentist estimate of δ_1 is then close to 0. In summary, the FHB method can lead to more accurate and reliable estimates than the frequentist counterpart when the number of parameters is large and data sample size is relatively small. The inclusion of the skewness parameter δ_1 can lead to more accurate estimates.

7. CONCLUDING REMARKS

The HIV viral dynamics models have been developed extensively in the statistical science community in the past 15 years. But the challenges still remain on developing proper statistical methods incorporating a large number (20 or more) of parameters with some prior knowledge and the skewed measurement error in the joint nonlinear mixed-effects models. In this article we develop a hybrid Bayesian inference method for the viral load modeling with skewed measurement errors. We are able to incorporate available informative priors on any parameters while leaving other parameters as frequentist parameters. Our rationale of choosing the frequentist and Bayesian parameters was to estimate the key parameters from the scientific perspective using only the data information (frequentist method) and to incorporate the prior on less important parameters (Bayesian inference). In addition, we consider multivariate SN distribution

introduced by [24], which is suitable for a BFH computation as briefly discussed in Appendix A. We propose the use of asymptotically informative prior in the longitudinal joint models as originally introduced in [35].

Regarding the real example illustrated in Figures 1 and 2, the results presented in the Table 1 based on the proposed BFH approach indicate that the first and second population decay rates could be approximated by $\widehat{\lambda}_1 = 24.75 + 10.68 z_0$ and $\widehat{\lambda}_2(t) = -2.34 + 0.07 \times (0.31 - 3.34t + 2.84t^2)$, respectively, where z_0 is the standardized CD4 value at the baseline. The population viral load value is estimated by $\widehat{V}(t) = \exp\{9.22 - \lambda_1 \times t\} + \exp\{3.79 - \lambda_2(t) \times t\}$. We can see that the first phase decay rate (λ_1) is significantly associated with the baseline CD4 value because of the significant estimate of β_3, and the second phase decay rate ($\lambda_2(t)$) is associated with the unknown true CD4 values that increase significantly with time t because of the significant estimates of a_3. Although the true association between the viral load and CD4 cell count at various time points can be more complicated than the aforementioned interpretation, our analysis provides a simple approximation that may facilitate the medical researchers to generate new scientific hypotheses and point to further research of the HIV viral load dynamics.

To the best of our knowledge, this is the first attempt to investigate the FHB inference approach based on the NLME joint model with SN distribution for longitudinal data. This kind of modeling approach by assuming the model errors with an SN distribution is important in many statistical applications areas not only HIV dynamic study, allowing accurate inference of parameters while adjusting for the data with skewness. In the real example and the simulation study, we have shown that the BFH inference can outperform the frequentist inference. In summary, we recommend using the BFH instead of frequrntist inference or full Bayesian inference with non-informative priors in the analysis of HIV viral load data when part of the model parameters have prior knowledge and the sample size is relatively small. The software for running the real example in Section 5 is available upon request. Future research on joint modeling incorporating time-to-event outcomes into the current model (1) is underway. Another theoretical challenge would be the hypothesis testing incorporating Bayesian and frequentist parameters. Finally, as one of referees mentioned, the BFH inference and full Bayesian inference should be further compared not only limited to the 0-1 loss (because with the 0-1 loss, the two approchaes are computationally equivalent if the frequentist parameters had constant prior in the full Bayesian analysis). Under a general loss function (e.g., squared error loss and absolute loss), the comparison would involve intensive computations and require significantly additional efforts given the proposed nonlinear mixed-effects joint model. These complicated problems are beyond the focus of this article, but investigation of this important issue is underway.

APPENDIX A. MULTIVARIATE SKEW-NORMAL DISTRIBUTION

Different versions of multivariate skew distributions have been proposed and used in the literature [24, 2, 3, 4, 15]. A new class of distributions by introducing skewness in multivariate elliptically distributions were developed in publication [24]. The class, which is obtained by using transformation and conditioning, which contains many standard families including the multivariate skew-normal (SN) distribution as special case. A k-dimensional random vector $\boldsymbol{Y}$ follows an k-variate skew-elliptical (SE) distribution if its probability density function (pdf) is given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}; m_\nu^{(k)}) = 2^k f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{A}; m_\nu^{(k)}) P(\boldsymbol{V} > \boldsymbol{0}), \tag{A.1}$$

where $\boldsymbol{A} = \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2$, $\boldsymbol{\mu}$ is a location parameter vector, $\boldsymbol{\Sigma}$ is a $k \times k$ positive (diagonal) covariance matrix, $\boldsymbol{\Delta} = \text{diag}(\delta_1, \delta_2, \ldots, \delta_k)$ is a $k \times k$ skewness matrix with the skewness parameter vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)^T$; $\boldsymbol{V}$ follows the elliptical distribution $El\left(\boldsymbol{\Delta}\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \boldsymbol{I}_k - \boldsymbol{\Delta}\boldsymbol{A}^{-1}\boldsymbol{\Delta}; m_\nu^{(k)}\right)$ and the density generator function $m_\nu^{(k)}(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{m_\nu(u)}{\int_0^\infty r^{k/2-1} m_\nu(u) dr}$, with $m_\nu(u)$ being a function such that $\int_0^\infty r^{k/2-1} m_\nu(u) dr$ exists. The function $m_\nu(u)$ provides the kernel of the original elliptical density and may depend on the parameter ν. This SE distribution is denoted by $SE(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}; m^{(k)})$. One example of $m_\nu(u)$, leading to an important special case used throughout the paper, is $m_\nu(u) = \exp(-u/2)$. This expression leads to the multivariate SN distribution.

As we know, a normal distribution is a special case of an SN distribution when the skewness parameter is zero. For completeness, this Appendix briefly summarizes the multivariate SN distribution introduced by [24] to be suitable for a Bayesian inference since it is built using the conditional method. See [24] for detailed discussion on properties of SN distribution. Assume a k-dimensional random vector $\boldsymbol{Y}$ follows a k variate SN distribution with location vector $\boldsymbol{\mu}$, $k \times k$ positive (diagonal) covariance matrix $\boldsymbol{\Sigma}$ and $k \times k$ skewness matrix $\boldsymbol{\Delta} = \text{diag}(\delta_1, \delta_2, \ldots, \delta_k)$.

A k-dimensional random vector $\boldsymbol{Y}$ follows a k-variate SN distribution, if its pdf is given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}) = 2^k |\boldsymbol{A}|^{-1/2} \phi_k\{\boldsymbol{A}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})\} P(\boldsymbol{V} > \boldsymbol{0}), \tag{A.2}$$

where $\boldsymbol{V} \sim N_k\{\boldsymbol{\Delta}\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \boldsymbol{I}_k - \boldsymbol{\Delta}\boldsymbol{A}^{-1}\boldsymbol{\Delta}\}$, and $\phi_k(\cdot)$ is the pdf of $N_k(\boldsymbol{0}, \boldsymbol{I}_k)$. We denote the above distribution by $SN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$. An appealing feature of equation (A.2) is that it gives independent marginal when $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$. The pdf (A.2) thus simplifies to

(A.3)

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. The mean and covariance matrix are given by $E(\boldsymbol{Y}) = \boldsymbol{\mu}+\sqrt{2/\pi}\boldsymbol{\delta}$, $cov(\boldsymbol{Y}) = \boldsymbol{\Sigma}+(1-2/\pi)\boldsymbol{\Delta}^2$. It is noted that when $\boldsymbol{\delta} = \boldsymbol{0}$, the SN distribution reduces to usual normal distribution. In order to have a zero mean vector, we should assume the location parameter $\boldsymbol{\mu} = -\sqrt{2/\pi}\boldsymbol{\delta}$.

According to the study by [24], if $\boldsymbol{Y}$ follows $SN_k(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Delta})$, it can be expressed by a convenient stochastic representation as follows.

(A.4) $$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta}|\boldsymbol{X}_0| + \boldsymbol{\Sigma}^{1/2}\boldsymbol{X}_1,$$

where $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ are two independent $N_k(\boldsymbol{0},\boldsymbol{I}_k)$ random vectors. Let $\boldsymbol{w} = |\boldsymbol{X}_0|$; then, $\boldsymbol{w}$ follows an k-dimensional standard normal distribution $N_k(\boldsymbol{0},\boldsymbol{I}_k)$ truncated in the space $\boldsymbol{w} > \boldsymbol{0}$. Thus, a two-level hierarchical representation of (A.4) is given by

(A.5) $$\boldsymbol{Y}|\boldsymbol{w} \sim N_k(\boldsymbol{\mu}+\boldsymbol{\Delta}\boldsymbol{w},\boldsymbol{\Sigma}),\ \boldsymbol{w} \sim N_k(\boldsymbol{0},\boldsymbol{I}_k)\boldsymbol{I}(\boldsymbol{w} > \boldsymbol{0}).$$

APPENDIX B. HIV DYNAMIC MODELS

Viral dynamic models can be formulated through a system of ordinary differential equations (ODE) [5, 12, 22, 29]. Following notation in [12] and [29], a mathematical ODE model for HIV dynamics can be written as follows by considering an infected cell compartment–productively infected cells (T_p).

(B.1) $$\begin{aligned}\frac{d}{dt}T_p &= kTV_I - d_pT_p,\\ \frac{d}{dt}V_I &= (1-\eta)d - cV_I,\\ \frac{d}{dt}V_{NI} &= \eta d + Nd_pT_p - cV_{NI},\end{aligned}$$

where V_I and V_{IN} are the concentrations of infectious virus and non-infectious virus, respectively, and T denotes the number of uninfected target cells for HIV, which can be assumed to be a constant at the early stage of HIV treatment. To account for compartment where the protease inhibitor drugs cannot completely block the production, we consider an additional virus production term with a constant (average) rate d in the model. The parameters d_p and c are the death rates of productively infected cells and free virus, respectively. Under some reasonable assumptions and simplifications, an analytic solution for equation (B.1) can be obtained. More details on the notation and simplifications can be found in Wu and Ding (1999). Thus, one useful approximate solution, which can be used to capture virus decay, has been proposed as follows.

(B.2) $$V(t) = \exp(p_1 - \lambda_1 t) + \exp(p_2 - \lambda_2 t)$$

where $V(t) = V_I(t) + V_{NI}(t)$ is the total number of HIV-1 RNA copies per mL of plasma, λ_1 and λ_2 are the first- and second-phase viral decay rates, respectively [22], $\exp(p_i)$ (i=0,1,2) reflect the baseline viral load at time $t = 0$. It is generally assumed that $\lambda_1 > \lambda_2$, which assures that the model is identifiable and is appropriate for empirical studies (Wu and Ding, 1999). It is of particular interest to estimate these viral decay rates because they quantify the antiviral effect, and hence, can be used to assess the efficacy of the antiviral treatments. In estimating these decay rates, only the early segment of the viral load trajectory data before rebound can be used [22, 29, 30, 31].

Nonlinear mixed-effects models based on two-compartment model with two phase decay rates (B.2) are powerful tools for modeling HIV viral dynamics. [30] showed that they are approximately equal to capture the viral load trajectory reasonably well within some time period. Although equation (B.2) is widely used in HIV dynamic studies, it is only applied to the early segment of the viral load data since the viral load trajectory may change to a different shape in later stage; see Figure 1(b). Thus, it may not be reasonable to assume that the viral decay rate is a constant during long-term treatment such as 48 weeks in the study to be considered in this paper. In other words, equation (B.2) is only short-term HIV dynamic model. To model the long-term HIV dynamics, a natural extension is to assume that the viral decay rates change over time, which may be a function of time-varying covariates such as CD4 cell count. Thus, we introduce an extended function as follows.

(B.3) $$V(t) = \exp\{p_1 - \lambda_1 t\} + \exp\{p_2 - \lambda_2(t)t\}$$

where the decay rate $\lambda_2(t)$ is a time-varying function. Intuitively, this equation is more reasonable because it assumes that the the second viral decay rate can vary with time as a result of drug resistance, medication adherence and other relevant clinical factors likely to affect changes in the viral load during the late period of treatment. Therefore, all data obtained during whole study period can be used by fitting this model. We also assume that $\lambda_1 > \lambda_2(t)$, for all time t in order to guarantee that there is the first phase of curve decay.

APPENDIX C. PROOF OF PROPOSITIONS

Proof of Proposition 1. (i) It will be convenient to view the number of repeated observations n_i $(i = 1,...,n)$ as integer valued random variables *iid* with N, with $P(N = k_j) = p_j$ $(j = 1,...,r)$. Thus the $\mathbf{y}_i$'s are iid and the $\mathbf{z}_i$'s

are iid. Let $l_+(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{Y}^n, \mathbf{Z}^n) + \log \pi(\boldsymbol{\theta}_1)$. By definition of $\boldsymbol{\theta}_n = (\check{\boldsymbol{\theta}}_{1,n}, \hat{\boldsymbol{\theta}}_{2,n})$, we have $l_+^{(1)}(\boldsymbol{\theta}_n) = \mathbf{0}$, thus

$$-l_+^{(1)}(\boldsymbol{\theta}_0) = l_+^{(1)}(\boldsymbol{\theta}_n) - l_+^{(1)}(\boldsymbol{\theta}_0) = l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_n - \theta_0),$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_0$. By assumption of set A, $\boldsymbol{\theta}_n \in A$, and for all large n, $-n^{-1}l_+^{(2)}(\boldsymbol{\theta}_n) \geq \inf_{\boldsymbol{\theta}\in A}[-n^{-1}l_+^{(2)}(\theta)] \geq (1/2)\inf_{\boldsymbol{\theta}\in A} I_+(\boldsymbol{\theta}) > 0$, i.e. there is an n_0 such that, $\min_{n\geq n_0} |n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)| > 0$. Thus for large n,

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = \Big(- n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)\Big)^{-1} n^{-1}l_+^{(1)}(\boldsymbol{\theta}_0) \to \mathbf{0}, \quad (a.s.)$$

since $n^{-1}l_+^{(1)}(\boldsymbol{\theta}_0) \overset{a.s.}{\to} E_{\boldsymbol{\theta}_0}[\partial \log f^{(1)}(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}] = \mathbf{0}$.

(ii) Now we also have $\tilde{\boldsymbol{\theta}}_n \to \theta_0$ (a.s.), so by the extended continuous mapping Theorem [26],

$$-n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n) = -[n^{-1}l^{(2)}(\theta_n) + o(1)] \overset{P}{\to} I(\boldsymbol{\theta}_0).$$

Also, $n^{-1/2}l_+^{(1)}(\boldsymbol{\theta}_0) = n^{-1/2}l^{(1)}(\boldsymbol{\theta}_0) + o(1)$, so by central limit theorem

$$n^{-1/2}l_+^{(1)}(\boldsymbol{\theta}_0) \overset{D}{\to} N(\mathbf{0}, I(\boldsymbol{\theta}_0))$$

and so by Slutsky's theorem,

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$$
$$= \Big(- n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)\Big)^{-1} n^{-1/2}l_+^{(1)}(\boldsymbol{\theta}_0) \overset{D}{\to} N(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0)).$$

We point out that in this case the Fisher information matrix is

$$I(\boldsymbol{\theta}) = -\sum_{j=1}^{r} \int\int l^{(2)}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, k_j) p_j d\mathbf{y} d\mathbf{z}.$$

Proof of Proposition 2. In this case, $l_+(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) + h_m(\boldsymbol{\theta}_1)$,

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = \Big(- n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)\Big)^{-1} n^{-1}l_+^{(1)}(\boldsymbol{\theta}_0) \to \mathbf{0}, \quad (a.s.)$$

since $n^{-1}l_+^{(1)}(\boldsymbol{\theta}_0) \overset{a.s.}{\to} E_{\boldsymbol{\theta}_0}[\partial \log f_+^{(1)}(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}] = \mathbf{0}$, with $f_+(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) = f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) + f_p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}_1)$, $f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ be the density of the observed current data and $f_p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}_1)$ be the density function of the prior data.

Note $l^{(1)}(\boldsymbol{\theta}_0)$ and $h_m^{(1)}(\boldsymbol{\theta}_{1,0})$ are independent, so by the definition of AIP,

$$\begin{aligned} n^{-1/2}l_+^{(1)}(\boldsymbol{\theta}_0) &= n^{-1/2}l^{(1)}(\boldsymbol{\theta}_0) + (m/n)^{1/2}m^{-1/2}h_m^{(1)}(\boldsymbol{\theta}_0) \\ &\overset{D}{\to} N(\mathbf{0}, I(\boldsymbol{\theta}_0) + cJ(\boldsymbol{\theta}_0)), \\ n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n) &= n^{-1}l^{(2)}(\tilde{\boldsymbol{\theta}}_n) + (m/n)m^{-1}h_m^{(2)}(\tilde{\boldsymbol{\theta}}_n) \\ &\overset{P}{\to} I(\boldsymbol{\theta}_0) + cJ(\boldsymbol{\theta}_0), \end{aligned}$$

and so

$$\begin{aligned} \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) &= \Big(- n^{-1}l_+^{(2)}(\tilde{\boldsymbol{\theta}}_n)\Big)^{-1} n^{-1/2}l_+^{(1)}(\boldsymbol{\theta}_0) \\ &\overset{D}{\to} N(\mathbf{0}, [I(\boldsymbol{\theta}_0) + cJ(\boldsymbol{\theta}_0)]^{-1}). \end{aligned}$$

ACKNOWLEDGEMENTS

The authors thank two anonymous referees, an Associate Editor and an Editor for their constructive comments and suggestions that led to a significant improvement of the article.

Received 20 December 2016

REFERENCES

[1] Acosta, E. P., Wu, H., Hammer, S. M., Walawander, A., Eron, J., Fichtenbaum, C. J., Pettinelli, C., Yu, S., Neath, D., Ferguson, E., Saah, A. J., Kuritzkes, D. R., Gerber, J. G. and for the Adult ACTG 5055 Protocol Team (2004). Comparison of two indinavir/ritonavir regimens in treatment-experienced HIV-infected individuals. *Journal of Acquired Immune Deficiency Syndromes* **37**, 1358–1366.

[2] Arellano-Valle, R. B., Bolfarine, H. and Lachos, V. H. (2005). Skew-normal linear mixed model. *Journal of Data Science* **3**, 415–438.

[3] 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. MR2410041

[4] Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distributions. *Journal of Royal Statistical Society, Series B* **6**, 579–602. MR1707862

[5] Guedj, J., Thiébaut, R. and Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198–1206. MR2414598

[6] Han, G., Huang, Y., Li, Q., Chen, L. and Zhang, X. (2013). Hybrid Bayesian inference on HIV viral dynamic models. *Journal of Applied Statistics* **40**, 2516–2532. MR3291179

[7] Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Verlag, New York.

[8] Higgins, K. M., Davidian, M. and Giltinan, D. M. (1997). A two-step method 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.

[9] Huang, Y. and Dagne, G. A. (2010). Skew-normal Bayesian Nonlinear Mixed-Effects Models with Application to AIDS Studies. *Statistics in Medicine* **29**, 2384–2398. MR2759954

[10] Huang, Y. and Dagne, G. A. (2012). Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses and measurement errors in covariates. *Biometrics* **68**, 943–953. MR3055199

[11] Huang, Y., Dagne, G. A. and Wu, L. (2011). Bayesian Inference on Joint Models of HIV Dynamics for Time-to-event and Longitudinal Data with Skewness and Covariate Measurement Errors. *Statistics in Medicine* **30**, 2930–2946. MR2844693

[12] 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. MR2227489

[13] Huang, Y., Yan, C., Xing, D., Zhang, N. and Chen, H. (2015). Jointly modeling event time and skewed-longitudinal data with missing response and mismeasured covariate for AIDS studies. *Journal of Biopharmaceutical Statistics* **25**, 670–694.

[14] Hughes, J. P. (1999). Mixed effects models with censored data with applications to HIV RNA levels. *Biometrics* **55**, 625–629.
[15] Jara, A., Quintana, F. and Martin, E. S. (2008). Linear mixed models with skew-elliptical distributions: A Bayesian approach. *Computational Statistics and Data Analysis* **52**, 5033–5045. MR2526212
[16] LeCam, L. M. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York. MR1066869
[17] Lederman, M. M., Connick, E., Landay, A., Kuritzkes, D. R., Spritzler, J., Clair, S. M., Kotzin, B. L., Fox, L., Chiozzi, M. H., Leonard, J. M., Rousseau, F., Wade, M., Roe, J. D., Martinez, A. and Kessler, H. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS Clinical Trials Group Protocol 315. *Journal of Infectious Diseases* **178**, 70–79.
[18] 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. MR2370792
[19] Lu, X. and Huang, Y. (2014). Bayesian Analysis of Nonlinear Mixed-Effects Mixture Models for Longitudinal Data With Heterogeneity and Skewness. *Statistics in Medicine* **33**, 2830–2849. MR3256541
[20] Nowak, M. A. and May, R. M. (2000). *Virus dynamics: mathematical principles of immunology and virology*. Oxford University Press, Oxford. MR2009143
[21] Oakes, D. (1999). Direct calculation of the information matrix via the EM Algorithm. *Journal of the Royal Statistical Society: Series B* **61**, 479–482. MR1680298
[22] Perelson, A. S., 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.
[23] Rao, B. L. S. P. (2000). *Asymptotic Theory of Statistical Inference*. Wiley, New York. MR0874342
[24] 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. MR2016224
[25] Strasser, H. (1981). Consistency of Maximum Likelihood and Bayes Estimates. *Annals of Statistics* **9**, 1107–1113. MR0628766
[26] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York. MR1385671
[27] 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.
[28] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York. MR0036976
[29] 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.
[30] Wu, H. and Wu, L. (2002). Identification of significant host factors for HIV dynamics modeled by nonlinear mixed-effects models. *Statistics in Medicine* **21**, 753–771.
[31] Wu, H. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley, New Jersey. MR2216899
[32] Wu, H., Ding, A. A. and de Gruttola, V. (1998). Estimation of HIV dynamic parameters. *Statistics in Medicine* **17**, 2463–2485.
[33] 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. MR1951254
[34] Yuan, A. (2009). Bayesian Frequentist Hybrid Inference. *The Annals of Statistics* **37**, 2458–2501. MR2543699
[35] Yuan, A. and Gooijer, J. D. (2014). Asymptotically informative prior for Bayesian analysis. *Communications in Statistics: Theory and Methods* **43**, 3080–3094. MR3225047
[36] Yuan, A., Zheng, G., Qin, J. and Li, Q. (2014). Analysis of Case-Control Genetic Association Studies Incorporating Prior Summary Statistics. *Statistics and Its Interface* **7**, 43–50. MR3197568

Gang Han
Department of Epidemiology and Biostatistics
School of Public Health
Texas A&M University
212 Adriance Lab Road
College Station, Texas 77845
USA
E-mail address: ghan@sph.tamhsc.edu

Yangxin Huang
Department of Epidemiology and Biostatistics
College of Public Health
University of South Florida
Tampa, Florida 33612
USA
E-mail address: yhuang@health.usf.edu

Ao Yuan
Department of Biostatistics
Bioinformatics and Biomathematics
Georgetown University
Washington, DC 20057
USA
E-mail address: ay312@georgetown.edu