Bayesian Inference of Mixed-effects Ordinary Differential Equations Models Using Heavy-tailed Distributions

Baisen Liu[a], Liangliang Wang[b], Yunlong Nie[b], Jiguo Cao[b,*]

[a]*School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, China*
[b]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada*

Abstract

A mixed-effects ordinary differential equation (ODE) model is proposed to describe complex dynamical systems. In order to make the inference of ODE parameters robust against the outlying observations and subjects, a class of heavy-tailed distributions is applied to model the random effects of ODE parameters and measurement errors in the data. The heavy-tailed distributions are so flexible that they include the conventional normal distribution as a special case. An MCMC method is proposed to make inferences on ODE parameters within a Bayesian hierarchical framework. The proposed method is demonstrated by estimating a pharmacokinetic mixed-effects ODE model. The finite sample performance of the proposed method is evaluated using some simulation studies.

Keywords: Metropolis-Hastings, Outliers, Pharmacokinetics, Scale mixtures of multivariate normal distributions, Smoothing Spline

1. Introduction

Ordinary differential equations are widely used to model complex dynamical systems in many areas of science and technology. For example, ODE models have been used in the study of HIV viral dynamics (Perelson et al., 1996; Perel-

*Corresponding email: jiguo_cao@sfu.ca

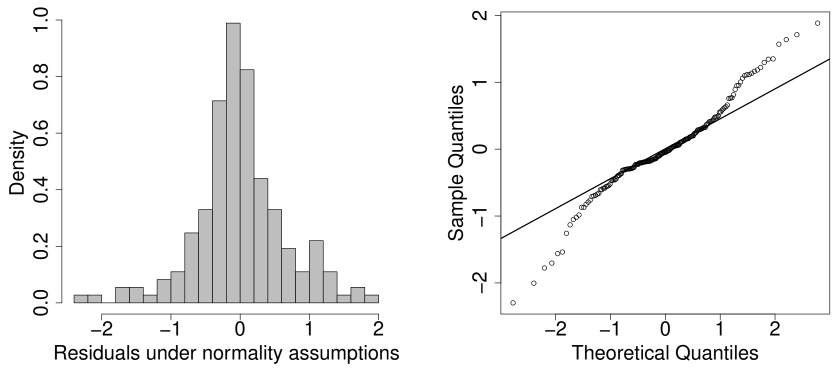
son and Nelson, 1999; Wu and Ding, 1999). Although ODE models are often proposed based on expert knowledge of the dynamical process of interest, the values of the ODE parameters are rarely known. Estimating these parameters from observational (noisy) data is an important but challenging statistical problem because most ODEs have no analytic solutions, and it is often computationally intensive to solve ODEs numerically.

Several methods have been developed for estimating ODE parameters from the noisy data. For instance, Liang and Wu (2008) proposed a two-step method and estimated the derivative using local polynomial regression. Ramsay et al. (2007) and Cao et al. (2008) developed a generalized profiling approach to estimate the ODE parameters. Cao et al. (2011) proposed a robust method for estimating ODE parameters when the data have outliers. Hall and Ma (2014) suggested a class of fast, easy-to-use, genuinely one-step procedures for estimating unknown parameters in dynamical system models. Brunel et al. (2014) developed a gradient matching approach for estimating ODE parameters. Li et al. (2015) considered a regularization estimation issue of the time-varying parameters of an ODE system and developed a modification of the parameter cascade approach (Ramsay et al., 2007). Chen and Wu (2008) and Cao et al. (2012) proposed a local estimation method and a penalized least square method, respectively, for estimating time-varying parameters in the ODE model. With the development of computing technology and MCMC algorithms, Bayesian approaches gain more and more attentions and are applied to estimate ODE models in recent years. For example, Campbell and Steele (2012) proposed a Bayesian smooth functional tempering method for the ODE models. Bhaumik and Ghosal (2015) considered the two-step estimation under the Bayesian framework. Dass et al. (2017) suggested a Laplace approximation method for obtaining the posterior inference of ODE parameters.

Longitudinal dynamical systems, also called mixed-effects ODE models, have been studied by Li et al. (2002); Putter et al. (2002); Huang and Wu (2006); Huang et al. (2006); Guedj et al. (2007). For instance, Huang and Wu (2006) proposed a parametric hierarchical Bayesian approach to model HIV dynamical data and provided an MCMC algorithm to sample from the posterior distribution of ODE parameters. Guedj et al. (2007) used the maximum likelihood approach directly to estimate unknown parameters in mixed-effects ODE models. Lahiri (2003) proposed a spline-enhanced population model to study pharmacokinetics using a random time-varying coefficient ODE model. Lately, Fang et al. (2011) proposed a fast two-stage estimating procedure for mixed-effects dynamical systems and applied it study longitudinal HIV virus data. Wang et al. (2014) proposed a semiparametric method to estimate a mixed-effects ODE model for the HIV combination therapy study. A common fundamental assumption of these methods is that the observations for the dynamical process follow a normal distribution, but this assumption may lack robustness and lead to biased inference when outliers exist.

As an illustration, we consider the PK/PD experiment (see Wasmuth et al., 2004) which investigated the pharmacokinetics of antiretroviral drugs in order to understand the widely used protease inhibitor combinations of indinavir (IDV) and ritonavir (RTV) for treating HIV-positive patients. Their study was designed to compare two different combinations of IDV and RTV, and each combination was taken by healthy volunteers twice daily for two weeks before the serum concentrations of IDV and RTV were measured at 13 unequally-spaced time points within twelve hours. Figure 1 displays the histogram and normal Q-Q plot of the obtained residuals by applying the conventional method which assumes the observations and random-effects follow normal distributions. Figure 1 shows that the underlying distribution of serum concentration may not follow the normal distribution. Hence, assuming normal distributions may be too restrictive to accurately model the serum concentration of the IDV in ODE mixed-effects models. Moreover, by performing a Shapiro-Wilk test of normality for the obtained residuals, the p-value is approximately 1.36×10^{-4}, which confirms that the normal distribution assumption is quite doubtful in this PK/PD data set.

To deal with this departure from normality, we propose to model the observations of the dynamical process and random effects of ODE parameters with



(a) Histgram

(b) Normal Q-Q plot

Figure 1: The histogram and the normal Q-Q plot of the obtained residuals assuming normal distributions for observations and random-effects in the PK/PD experiment.

a class of heavy-tailed distributions, called the scale mixture of multivariate normal distributions (SMN) (Andrews and Mallows, 1974), which includes the multivariate normal distribution as a special case. In the literature, this class of heavy-tailed distributions has been applied to regression models (Lange and Sinsheimer, 1993; Liu, 1996), linear mixed-effects models (Choy and Smith, 1997; Rosa et al., 2003, 2004), and nonlinear mixed-effects models (Meza et al., 2012; De la Cruza, 2014), to obtain robust estimates against outlying observations. However, there is little study to apply this class of heavy-tailed distributions on the robust inferences of ODE parameters. This paper will fill this gap and provide a robust inference approach for the ODE models.

To make robust inference on the ODE parameters, one possible approach is to implement a maximum likelihood estimation (MLE) method. However, due to the complexity of dynamic systems, the solutions of ODEs generally have no explicit expressions, which makes it difficult to maximize the likelihood function. In contrast, the Bayesian methods are widely welcomed due to the convenient and efficient implementations.

This article has four main contributions. (i) We propose a mixed-effects ODE model, which considers the within-subject and between-subject variations simultaneously and makes statistical inference by borrowing information from all subjects. (ii) Our method uses a class of heavy-tailed distributions for random-effects and observations for the dynamical process, which is robust against the outlying subjects and the outlying observations within individual subjects. (iii) Our method can detect the subjects which are outliers or have outlying observations by estimating latent variables in the model. (iv) We develop a highly efficient MCMC sampling scheme which allows to estimate complex dynamic models using the hierarchical structure of the proposed approach.

The remainder of this article is organized as follows. Section 2 briefly reviews the scale mixture of multivariate normal distributions. Section 3 introduces our proposed Bayesian estimation method for the mixed-effect ODE models. Section 4 demonstrates our proposed method in comparison with conventional methods by analyzing a real pharmacokinetics application. Section 5 evaluates the finite sample performance of our proposed method using some simulation studies. We end this article with conclusions and some discussions in Section 6. The Matlab codes for our simulation studies can be downloaded at `https://github.com/caojiguo/ODEHeavyTail`.

2. A Brief Review of the Scale Mixture of Multivariate Normal Distributions

In this section, we provide a brief review of the scale mixture of multivariate normal (SMN) distributions that will be applied in our hierarchical models.

An m-dimensional random vector $\mathbf{Y}$ is said to follow a *scale mixture of multivariate normal distribution* with parameters $\boldsymbol{\mu} \in R^m$, an $m \times m$ positive definite symmetric matrix $\boldsymbol{\Sigma}$, and a univariate probability distribution function $H(\cdot; \boldsymbol{\nu})$ with $H(0; \boldsymbol{\nu}) = 0$, if the probability density function of $\mathbf{Y}$ is given by

$$p(\mathbf{y}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \int_0^\infty u^{m/2} \exp(-\frac{uD^2(\mathbf{y})}{2}) \mathrm{d}H(u; \boldsymbol{\nu}), \tag{1}$$

where $D^2(\mathbf{y}) = (\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})$. We use the notation $\mathbf{Y} \sim SMN_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$ to indicate that $\mathbf{Y}$ has the density (1). When the mixture distribution function

H is degenerate, $SMN_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$ reduces to the usual multivariate normal distribution $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Azzalini and Capitanio (2014) provided a convenient stochastic representation for the SMN distributions

$$\mathbf{Y} = \boldsymbol{\xi} + U^{-1/2}\mathbf{Z}, \tag{2}$$

where $\mathbf{Z} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$ is independent of the mixture variable $U \sim H(\cdot; \boldsymbol{\nu})$, and $\boldsymbol{\nu}$ is a scalar or vector valued parameter. Another convenient form is to use the following hierarchical representation

$$\mathbf{Y}|U \sim N_m(\boldsymbol{\mu}, U^{-1}\boldsymbol{\Sigma}), \quad U \sim H(\cdot; \boldsymbol{\nu}). \tag{3}$$

From (3), the mean and covariance of $\mathbf{Y}$ are given, respectively, by

$$\mathrm{E}(\mathbf{Y}) = \mathrm{E}[\mathrm{E}(\mathbf{Y}|U)] = \boldsymbol{\mu},$$

and

$$\mathrm{Cov}(\mathbf{Y}) = \mathrm{E}(\mathrm{Cov}(\mathbf{Y}|U)) + \mathrm{Cov}(\mathrm{E}(\mathbf{Y}|U)) = \mathrm{E}(U^{-1})\boldsymbol{\Sigma}.$$

Obviously, if $\mathrm{E}(U^{-1}) < \infty$, then $\mathbf{Y}$ has a finite positive definite covariance matrix.

The class of SMN distributions provides a group of heavy-tailed distributions that are often useful for robust inference. A special distribution of the SMN class is the Student's t distribution (Lange et al., 1989) that has been extensively applied in robust regressions, which can be obtained by assuming a Gamma distribution with shape parameter $\nu/2$ and rate parameter $\nu/2$ for U, i.e., $U \sim Ga(\nu/2, \nu/2)$, which has the following density

$$p(x) = \frac{(\nu/2)^{\nu/2} x^{\nu/2-1}}{\Gamma(\nu/2)} \exp\left(-\frac{1}{2}\nu x\right), \quad x, \nu > 0,$$

where the parameter ν corresponds to the degrees of freedom of the Student's t distribution. If letting $\nu \to \infty$, the Gaussian distribution is recovered.

3. Estimating Mixed-effects ODEs

3.1. Bayesian Framework

Suppose that the dynamical process $X_i(t)$, $i = 1, ..., n$, for the i-th subject is defined as

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = f(X_i(t)|\boldsymbol{\theta}_i), \tag{4}$$

where t is continuous in some interval $[0, T]$, f is a known parametric function, and $\boldsymbol{\theta}_i$ is a q-dimensional vector of ODE parameters for individual subjects. Without loss of generality, we assume that $X_i(t)$ is one-dimensional dynamical curve in this article. Let $\mathbf{X}_i = (X_i(t_{i1}), \cdots, X_i(t_{in_i}))^T$ with $X_i(t)$ being the solution of the ODE (4) given the initial condition $X_i(0)$ and the ODE parameters $\boldsymbol{\theta}_i$. Generally, the ODE solution $X_i(t)$ is often observed with noise in practice. Moreover, the initial condition $X_i(0)$ is always unknown and needed to be estimated. In this article, we incorporate the unknown condition $X_i(0)$ into $\boldsymbol{\theta}_i$ and treat the initial condition $X_i(0)$ as part of the unknown parameters $\boldsymbol{\theta}_i$. In other words, the first element of $\boldsymbol{\theta}_i$ denotes the unknown initial condition $X_i(0)$ and the rest of $\boldsymbol{\theta}_i$ are the ODE parameters.

Let $\mathbf{Y}_i = (y_{i1}, \cdots, y_{in_i})^T$ denote the vector of observations or measurements for the i-th subject at the observation time $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})^T$. The following hierarchical regression model is used:

$$\text{Within} - \text{subject variation}: \qquad \mathbf{Y}_i = h(\mathbf{X}_i|\boldsymbol{\theta}_i) + \boldsymbol{\epsilon}_i, \tag{5}$$

$$\text{Between} - \text{subject variation}: \qquad \boldsymbol{\theta}_i = \boldsymbol{\xi} + \mathbf{b}_i, \tag{6}$$

where $h(\cdot)$ is a known function (e.g., $h(\cdot) = \log(\cdot)$ in many statistical analysis), $\boldsymbol{\epsilon}_i$ are measurement errors, $\boldsymbol{\xi}$ is a q-dimensional fixed effect, and $\mathbf{b}_i$ is a q-dimensional random effect which accounts for the within-subject correlation.

In conventional methods, a common assumption is that the random effect of ODE parameters $\mathbf{b}_i$ and the data errors $\boldsymbol{\epsilon}_i$ both follow the multivariate normal distributions. However, as discussed in Section 1, such normality assumptions are vulnerable in the presence of outlying observations, which can seriously affect the estimation accuracy of the mixed-effects ODE model. Thus, more flexible distributions are necessary to replace the normality assumption. Therefore, we propose to use the scale mixture of multivariate normal distributions for ODE random effects $\mathbf{b}_i$ and within-subject data errors $\boldsymbol{\epsilon}_i$. In other words, we assume that $\mathbf{b}_i \sim SMN_q(\mathbf{0}, \boldsymbol{\Sigma}, H_1)$ and $\boldsymbol{\epsilon}_i \sim SMN_{n_i}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}, H_2)$.

Applying the stochastic representation (3), our proposed mixed-effects ODE model can be written as the following hierarchical structure

$$\begin{array}{ccl} \mathbf{Y}_i|\boldsymbol{\theta}_i, U_i, \sigma_\epsilon^2 & \overset{ind.}{\sim} & N_{n_i}(h(\mathbf{X}_i|\boldsymbol{\theta}_i), U_i^{-1}\sigma_\epsilon^2 \mathbf{I}_{n_i}) \ , \\ \boldsymbol{\theta}_i|W_i, \boldsymbol{\Sigma} & \overset{ind.}{\sim} & N_q(\boldsymbol{\xi}, W_i^{-1}\boldsymbol{\Sigma}) \ , \\ U_i & \overset{ind.}{\sim} & H_1(\kappa) \ , \\ W_i & \overset{ind.}{\sim} & H_2(\nu) \ , \end{array} \tag{7}$$

where U_i and W_i are two latent variables with distributions H_1 and H_2, respectively, and σ_ϵ^2 and $\boldsymbol{\Sigma}$ have pre-specified priors, $\sigma_\epsilon^{-2} \sim Ga(a_0, b_0)$ and $\boldsymbol{\Sigma} \sim IW(\mathbf{S}_0, df)$, respectively, where the Gamma distribution $Ga(a_0, b_0)$ has the shape parameter a_0 and the rate parameter b_0, and the Inverse Wishart distribution $IW(\mathbf{S}_0, df)$ has the scale matrix $\mathbf{S}_0$ and degrees of freedom df. The hyper-parameters a_0, b_0, $\mathbf{S}_0$ and df are pre-specified. One popular choice for H_1 and H_2 is to use the gamma distribution; other possible choices are discussed in Azzalini and Capitanio (2014). When U_i and W_i have degenerate distributions, model (7) reduces to the conventional model with the normal distribution assumption. However, when some U_i^{-1} has a large value, it indicates that the i-th subject may have outlying observations. When some W_i^{-1} has a large value, it indicates that the i-th subject may be an outlying subject with outlying ODE parameters. This outlier detection will be demonstrated in our applications at Section 4. Hence, our proposed model (7) is more flexible than the conventional model with the normal distribution assumption.

The ODE model (4) often has no analytical solutions, and can be obtained numerically after specifying the values of ODE parameters and initial conditions. It is well known that, the ODE solution is very sensitive to the values of ODE parameters, and we have to solve ODEs repeatedly over thousands candidate values of ODE parameters, which leads to intensive computation. Therefore, we propose to estimate the ODE solution $X_i(t)$ with a linear combination of basis functions.

Let $\boldsymbol{\phi}_i(t) = (\phi_1(t), \cdots, \phi_{K_i}(t))^T$ be a vector of basis functions with dimension K_i. We estimate the ODE solution $X_i(t)$ with a linear combination of basis functions, i.e.

$$X_i(t) = \sum_{k=1}^{K_i} c_{ik}\phi_k(t) = \mathbf{c}_i^T \boldsymbol{\phi}_i(t)\,, \tag{8}$$

where $\mathbf{c}_i = (c_{i1}, \cdots, c_{iK_i})^T$ is a vector of basis coefficients which needs to be estimated from the noisy data. We choose cubic B-splines as basis functions, because any B-spline basis function is only positive over a short subinterval and zero elsewhere. To ensure the desired flexibility, a number of basis functions has to be large enough. Our numerical studies show that the proposed approximation obtains similar results when the number of basis functions is large enough.

We measure the fidelity of the nonparametric function $X_i(t)$ to the ODE model by defining a penalty term

$$F(X_i(t)|\boldsymbol{\theta}_i) = \int_0^T [LX_i(t)]^2 \mathrm{d}t, \tag{9}$$

where a differential operator $LX_i(t) = \mathrm{d}X_i(t)/\mathrm{d}t - f(X_i(t)|\boldsymbol{\theta}_i)$. Then, given any values of $\boldsymbol{\theta}_i$, $X_i(t)$ is estimated by minimizing

$$\int_0^T [LX_i(t)]^2 \mathrm{d}t = \int_0^T \left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(t) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(t)|\boldsymbol{\theta}_i)\right]^2 \mathrm{d}t\,, \tag{10}$$

where $\dot{\boldsymbol{\phi}}_i(t)$ denotes the derivative $\mathrm{d}\boldsymbol{\phi}_i(t)/\mathrm{d}t$. This idea was first proposed by Ramsay et al. (2007), who showed that using this approximated ODE solution made the optimization iterations converge faster than using the numerical ODE solution directly.

The integration in (10) usually does not have a closed-form expression and needs to be evaluated using numerical quadrature. We use the composite Simpson's rule (Burden and Douglas, 2000), which provides a good approximation to the exact integral. Let Q be an even integer. The interval $[0, T]$ is partitioned by equally-spaced quadrature points $0 = s_0 < s_1 < ... < s_Q = T$. Then, by the composite Simpson's rule, we have

$$\begin{aligned}
&\int_0^T \left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(t) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(t)|\boldsymbol{\theta}_i)\right]^2 \mathrm{d}t \\
\approx\ & \frac{T}{3Q}\Bigg\{\left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(s_0) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(s_0)|\boldsymbol{\theta}_i)\right]^2 + 2\sum_{q=1}^{Q/2-1}\left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(s_{2q}) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(s_{2q})|\boldsymbol{\theta}_i)\right]^2 \\
&+4\sum_{q=1}^{Q/2}\left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(s_{2q-1}) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(s_{2q-1})|\boldsymbol{\theta}_i)\right]^2 + \left[\mathbf{c}_i^T \dot{\boldsymbol{\phi}}_i(s_Q) - f(\mathbf{c}_i^T \boldsymbol{\phi}_i(s_Q)|\boldsymbol{\theta}_i)\right]^2\Bigg\}.
\end{aligned}$$

To make the approximation accurate, Q needs to be reasonably large, for example, $Q = 10K_i$. The above optimization procedure can be implemented by the Matlab function "**lsqnonlin**" conveniently.

Denote $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_n^T)^T$. Let $\mathbf{U} = (U_1, ..., U_n)^T$ and $\mathbf{W} = (W_1, ..., W_n)^T$ be the latent variables. Then the joint likelihood can be expressed explicitly as

$$\begin{aligned}
& L(\mathbf{Y}, \boldsymbol{\Theta}, \mathbf{U}, \mathbf{W}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \sigma_\epsilon^2, \kappa, \nu) \\
= & \prod_{i=1}^n L_i(\mathbf{Y}_i, \boldsymbol{\theta}_i, U_i, W_i|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \sigma_\epsilon^2, \kappa, \nu),
\end{aligned}$$

where $L_i(\cdot|\cdot)$ is the likelihood function of the i-th subject, which is given by

$$\begin{aligned}
& L_i(\mathbf{Y}_i, \boldsymbol{\theta}_i, U_i, W_i|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \sigma_\epsilon^2, \kappa, \nu) \\
= & L_i(\mathbf{Y}_i, U_i|\boldsymbol{\theta}_i, \sigma_\epsilon^2, \kappa) L_i(\boldsymbol{\theta}_i, W_i|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \nu),
\end{aligned}$$

with

$$L_i(\mathbf{Y}_i, U_i|\boldsymbol{\theta}_i, \sigma_\epsilon^2, \kappa) = p(\mathbf{Y}_i|U_i, \boldsymbol{\theta}_i, \sigma_\epsilon^2) H_1(U_i|\kappa),$$

and

$$L_i(\boldsymbol{\theta}_i, W_i|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \nu) = p(\boldsymbol{\theta}_i|\boldsymbol{\xi}, \boldsymbol{\Sigma}, W_i) H_2(W_i|\nu).$$

To complete the Bayesian specification of the proposed model, the following prior distribution is assigned on the fixed-effects: $\boldsymbol{\xi} \sim N_q(\boldsymbol{\xi}_0, \boldsymbol{\Omega}_0)$, where the hyper-parameters $\boldsymbol{\xi}_0$ and $\boldsymbol{\Omega}_0$ are pre-specified. Following the recommendations of Massuia et al. (2017), the prior distributions for κ and ν are chosen as an exponential distribution with the hyperparameter λ_κ and λ_ν, respectively. Furthermore, we assign a restriction of $(2.0, \infty)$ on both κ and ν, because the values of κ and ν must be greater than 2.0 to ensure $\mathrm{E}(U^{-1}) < \infty$ and $\mathrm{E}(W^{-1}) < \infty$ which further lead to both $\mathbf{Y}_i$ and $\boldsymbol{\theta}_i$ have finite positive definite covariance matrices. The hyper-priors for λ_κ and λ_ν are set as the Uniform distributions $U(c, d)$ given the values of c and d.

The joint posterior distribution of the parameters of the model conditional on the data is obtained by combining the joint likelihood and the prior distributions using the Bayes' theorem. The full conditional posterior distributions are presented in the appendix. They are sampled using the Monte Carlo methods.

3.2. Model Comparison

To compare the candidate models, in this article, we apply the following measures of model adequacy: the *conditional predictive ordinate* (CPO; Chen et al., 2000), the *deviance information criterion* (DIC; Spiegelhalter et al., 2002) and the *Widely Applicable Information Criterion* (WAIC; Watanabe, 2010). In this section, we briefly review the theory of these model selection criteria under the general Bayesian hierarchical framework.

Assume that we have a sample $\mathbf{y} = (y_1, ..., y_n)^T$. Let $\mathbf{y}_{-i} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)^T$ be the $(n-1) \times 1$ vector, with y_i omitted. Let $f(y_i|\boldsymbol{\vartheta})$ denote the density function of y_i that depends on some unknown parameters $\boldsymbol{\vartheta}$. Then, the conditional predictive distribution for y_i is defined by

$$\mathrm{CPO}_i = f(y_i|\mathbf{y}_{-i}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{-i})} = \int f(y_i|\boldsymbol{\vartheta}, \mathbf{y}_{-i}) p(\boldsymbol{\vartheta}|\mathbf{y}_{-i}) \mathrm{d}\boldsymbol{\vartheta},$$

which gives the likelihood of each data point conditional on the remainder of the data. We estimate CPO_i based on the MCMC samples of $\boldsymbol{\vartheta}$ (Carlin and Louis, 2008). Let $\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_M$ be the posterior samples from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$ with the size M after the burn-in. A Monte Carlo estimate of CPO_i is given by

$$\widehat{\mathrm{CPO}}_i = \left\{ \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{f(y_i|\boldsymbol{\vartheta}_\ell)} \right\}^{-1},$$

where $\{\boldsymbol{\vartheta}_\ell\}_{\ell=1}^M$ are the posterior samples of $\boldsymbol{\vartheta}$ (De la Cruza, 2014). Finally, the common summary statistic of CPO_i's is defined as $\mathrm{LCPO} = \sum_{i=1}^n \log(\widehat{\mathrm{CPO}}_i)$, which is often called the logarithm of the pseudo Bayes factor. A larger value of LCPO indicates a better model.

The DIC statistic measures the fit and the complexity of the model considered. Define the deviance

$$D(\boldsymbol{\vartheta}) = -2\log f(\mathbf{y}|\boldsymbol{\vartheta}) + 2\log g(\mathbf{y}),$$

where $f(\mathbf{y}|\boldsymbol{\vartheta})$ is the likelihood function of $\mathbf{y}$ and $g(\mathbf{y})$ is the normalized constant. Then the DIC statistic is defined as

$$\mathrm{DIC} = \overline{D(\boldsymbol{\vartheta})} + p_D = 2\,\overline{D(\boldsymbol{\vartheta})} - D(\bar{\boldsymbol{\vartheta}}),$$

where $\overline{D(\boldsymbol{\vartheta})} = \mathrm{E}_{\boldsymbol{\vartheta}|\mathbf{y}}[D(\boldsymbol{\vartheta})] = E_{\boldsymbol{\vartheta}|\mathbf{y}}[-2\log f(\mathbf{y}|\boldsymbol{\vartheta})]$ is the posterior expectation of the deviance, $p_D = \overline{D(\boldsymbol{\vartheta})} - D(\bar{\boldsymbol{\vartheta}})$ is the effective number of parameters, and $\bar{\boldsymbol{\vartheta}}$ is the posterior mean of $\boldsymbol{\vartheta}$. A smaller DIC value indicates a better model.

The third comparison criterion is to use the Widely Applicable or Watanabe-Akaike Information Criterion (WAIC) which was first proposed by Watanabe (2010). In Bayesian models, the WAIC can be viewed as an improvement on the DIC and it is asymptotically equal to Bayesian cross-validation. Define the log point-wise predictive density (LPPD)

$$\mathrm{LPPD} = \sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\vartheta}) p_{post}(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta}.$$

Then the WAIC is given by (Gelman et al. 2014)

$$\mathrm{WAIC} = -2\mathrm{LPPD} + 2p_{\mathrm{WAIC}},$$

where the penalty term, p_{WAIC}, is used to correct the effective number of parameters. There are two different approaches to calculate this correction. Here, following the suggestion of Gelman et al. (2014), we use the variance version,

$$p_{\mathrm{WAIC}} = \sum_{i=1}^{n} \mathrm{var}_{post}(\log p(y_i|\boldsymbol{\vartheta})),$$

which can be estimated by

$$\hat{p}_{\mathrm{WAIC}} = \sum_{i=1}^{n} V_{\ell=1}^{M}(\log p(y_i|\boldsymbol{\vartheta}_\ell)),$$

where $\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_M$ are the posterior MCMC sample of $\boldsymbol{\vartheta}$ and $V_{\ell=1}^M a_\ell = \frac{1}{M-1}\sum_{\ell=1}^M (a_\ell - \bar{a})^2$ with $\bar{a} = \frac{1}{M}\sum_{\ell=1}^M a_\ell$. Moreover, the log pointwise predictive density, LPPD, is calculated by

$$\widehat{\text{LPPD}} = \sum_{i=1}^n \log\left(\frac{1}{M}\sum_{\ell=1}^M p(y_i|\boldsymbol{\vartheta}_\ell)\right).$$

Finally, the estimated WAIC criterion is given by

$$\widehat{\text{WAIC}} = -2\widehat{\text{LPPD}} + 2\widehat{p}_{\text{WAIC}}.$$

A smaller WAIC value indicates a better model.

3.3. Bayesian Case Influence Diagnostics

Our proposed hierarchical models may be sensitive to the underlying model assumptions, so it is of interest to determine which subjects/observations may be influential for the analysis. Let $\mathcal{D}$ be the full data and $\mathcal{D}^{(-i)}$ be the data with the ith subject deleted. Let P denote the posterior distribution of $\boldsymbol{\vartheta}$ based on full data and $P_{(-i)}$ denote the posterior distribution of $\boldsymbol{\vartheta}$ based on the data $\mathcal{D}^{(-i)}$. Define the K-L divergence between P and $P_{(-i)}$ by $K\{P, P_{(-i)}\} = \int p(\boldsymbol{\vartheta}|\mathcal{D}) \log\{\frac{p(\boldsymbol{\vartheta}|\mathcal{D})}{p(\boldsymbol{\vartheta}|\mathcal{D}^{(-i)})}\}\mathrm{d}\boldsymbol{\vartheta}$. Following the work of Peng and Dey (1995), $K\{P, P_{(-i)}\}$ can be expressed as $\log \mathrm{E}_{\boldsymbol{\vartheta}|\mathcal{D}}[\{f(\mathbf{y}_i|\boldsymbol{\vartheta})\}^{-1}] + \mathrm{E}_{\boldsymbol{\vartheta}|\mathcal{D}}[\log\{f(\mathbf{y}_i|\boldsymbol{\vartheta})\}] = -\log(\text{CPO}_i) + \mathrm{E}_{\boldsymbol{\vartheta}|\mathcal{D}}[\log\{f(\mathbf{y}_i|\boldsymbol{\vartheta})\}]$, where $\mathrm{E}_{\boldsymbol{\vartheta}|\mathcal{D}}(\cdot)$ denotes the expectation with respect to the joint posterior $p(\boldsymbol{\vartheta}|\mathcal{D})$. A Monte Carlo estimate of $K\{P, P_{(-i)}\}$ (Cancho et al., 2011; Lachos et al., 2011) is given by

$$K\{\widehat{P, P_{(-i)}}\} = -\log(\widehat{\text{CPO}_i}) + \frac{1}{M}\sum_{\ell=1}^M \log\{f(\mathbf{y}_i|\boldsymbol{\vartheta}_\ell)\}, i = 1, ..., n.$$

A large value of the K-L divergence indicates that the subject/observation is influential for the analysis.

4. Applications: A Pharmacokinetic Study

In this section, we utilize our proposed approach to revisit the pharmacokinetic study of the HIV combination therapy (Wasmuth et al., 2004). This experiment follows a crossover design with subjects randomized to two treatments with different combinations of IDV and RTV. For illustration, we only consider the data collected for one treatment with the combination of 600mg IDV and 100mg RTV. In this data set, the serum concentration of IDV was measured at 0, 0.5, 1.0, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0, 8.0, 10.0 and 12.0 hours for 14 healthy volunteers after they took the combination of IDV and RTV twice daily for two weeks. The following PK/PD dynamical model has been extensively considered (Wasmuth et al., 2004; Wang et al., 2014),

$$\frac{\mathrm{d}C_i(t)}{\mathrm{d}t} = -Ke_iC_i(t) + \frac{D_iKe_iKa_i}{Cl_i}\exp(-Ka_it), i = 1, ..., n, \tag{11}$$

where D_i denotes the known cumulative amount of unabsorbed drug at $t = 0$ for the i-th subject (in this dataset, $D_i = 600$), Cl_i denotes the rate of the total body drug clearance, and Ka_i and Ke_i denote the drug absorption and elimination rates, respectively.

In order for the ODE parameters $(Ka_i, Ke_i, Cl_i)^T$ to be meaningful, they must be positive. Therefore, we reparameterized them in the logarithmic scales to remove the positivity constraints. The initial condition $C_i(0)$ is also estimated together with the ODE parameters. Let $\boldsymbol{\theta}_i = (\ln(C_i(0)), \ln(Ka_i), \ln(Ke_i), \ln(Cl_i))^T$. We assume that $\boldsymbol{\theta}_i$ follows the scale mixture of multivariate normal distributions $SMN_4(\boldsymbol{\xi}, \boldsymbol{\Sigma}, H_1)$, where $\boldsymbol{\xi}$ is the fixed effect of the ODE model, and the distribution H_1 is chosen as a gamma distribution with the shape parameter $\nu/2$ and rate parameter $\nu/2$. Using the hierarchical representation (3), this is equivalent to assume that $\boldsymbol{\theta}_i|W_i \sim N_4(\boldsymbol{\xi}, W_i^{-1}\boldsymbol{\Sigma})$ with $W_i \sim Ga(\nu/2, \nu/2)$. Let $\mathbf{C}_i = (C_i(t_{i1}), ..., C_i(t_{in_i}))^T$ be the true drug concentrations at observation times $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})^T$ and $\mathbf{Y}_i = (y_{i1}, \cdots, y_{in_i})^T$ be the noisy measurements of $\mathbf{C}_i$ with $n_i = 13$. We assume that the data follow the scale mixture of multivariate normal distributions $\mathbf{Y}_i \sim SMN_{n_i}(\mathbf{C}_i, \sigma_\epsilon^2\mathbf{I}_{n_i}, H_2)$ where the distribution H_2 is chosen as a gamma distribution with the shape parameter $\kappa/2$ and rate parameter $\kappa/2$. This is equivalent to assume a hierarchical representation $\mathbf{Y}_i|U_i \sim N_{n_i}(\mathbf{C}_i, U_i^{-1}\sigma_\epsilon^2\mathbf{I}_{n_i})$ with $U_i \sim Ga(\kappa/2, \kappa/2)$.

We apply the proposed Bayesian method to estimate the mixed-effects ODE

(11) from the data. We use cubic B-splines with 13 equally-spaced knots in $[0, 12]$ to approximate the ODE solution. We set a gamma prior $Ga(a, b)$ for σ_ϵ^{-2}, an Inverse Wishart prior $IW(\mathbf{S}_0, f_0)$ for $\mathbf{\Sigma}$, a multivariate normal prior $N_4(\boldsymbol{\xi}_0, \mathbf{\Omega}_0)$ for $\boldsymbol{\xi}$, and a Uniform prior $U(c, d)$ for λ_κ and λ_ν. Moreover, we choose the following values for the hyper-parameters: $\boldsymbol{\xi}_0 = (0, -0.30, -1.0, 3.0)^T$, $\mathbf{\Omega}_0 = \text{diag}(1000, 1000, 1000, 1000)$, $\mathbf{S}_0 = \text{diag}(0.01, 0.01, 0.01, 0.01)$, $f_0 = 5$, $a = 1$, $b = 0.01$, $c = 0.02$, and $d = 5$.

The proposed MCMC algorithm is run for 20,000 iterations. With the 'burn-in' of the first 10,000 samples, we choose 1,000 equally-spaced samples from the rest of the iterations. We compare our proposed model using the scale mixture of multivariate normal (SMN) distributions with the conventional model which assumes that both the ODE parameter $\boldsymbol{\theta}_i$ and the data $\mathbf{Y}_i$ follow the normal distributions.

Table 1: The logarithm of the pseudo Bayes factor LCPO $= \sum_{i=1}^n \log(\widehat{\text{CPO}}_i)$, the DIC and the WAIC for the pharmacokinetic mixed effects ODE model (11). A larger value of LCPO or a smaller value of DIC/WAIC indicates a better model.

Distribution of Data	Distribution of ODE Random Effects	LCPO	DIC	WAIC
Normal	Normal	-221.02	420.46	421.65
SMN	SMN	**-193.35**	**373.56**	**386.58**

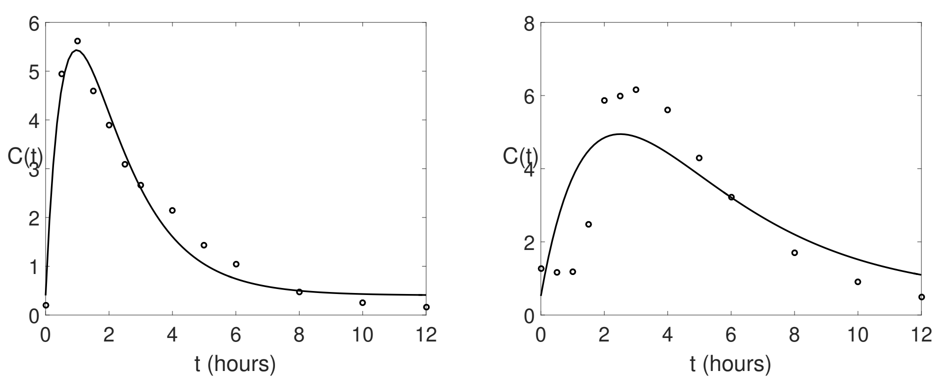
Table 1 shows that our proposed model using the SMN distribution has smaller values of DIC and WAIC and a larger value of LCPO $= \sum_{i=1}^n \log(\widehat{\text{CPO}}_i)$ than the conventional model assuming that the ODE parameters and the data follow the normal distributions; hence our proposed model is better than the conventional method. Table 2 displays the posterior means, the standard errors and the corresponding 95% equal-tail credible intervals for the fixed-effects

Table 2: A summary of the estimated posterior means and posterior standard deviations (STD) of the population ODE parameters $(Ka, Ke, Cl)^T$ in the pharmacokinetic mixed effects ODE model (11) and the corresponding 95% equal-tail credible/conficence intervals when assuming that ODE parameters and noisy data follow the scale mixture of multivariate normal distributions. Here, L_{CI} and R_{CI} denote the left and right side of the 95% credible/confidence intervals.

Parameters	Method	Mean	STD	L_{CI}	R_{CL}
Ka	Bayesian-SMN	0.591	0.051	0.492	0.694
	Bayesian-Normal	0.579	0.042	0.502	0.685
	MLE	0.743	0.235	0.282	1.203
Ke	Bayesian-SMN	0.372	0.027	0.319	0.429
	Bayesian-Normal	0.381	0.034	0.319	0.458
	MLE	0.271	0.022	0.228	0.314
Cl	Bayesian-SMN	20.898	1.836	17.520	24.763
	Bayesian-Normal	19.970	1.893	16.637	23.956
	MLE	16.484	1.290	13.955	19.014

Table 3: The estimated weights in the pharmacokinetic mixed-effect ODE model (11) under the assumption that the ODE parameters and noisy data follow the scale mixture of multivariate normal distributions.

Subject	1	2	3	4	5	6	7
Residual errors ($\widehat{U}_i^{-1}$)	0.647	1.791	0.671	3.955	2.906	2.712	0.830
Random effects ($\widehat{W}_i^{-1}$)	**13.903**	3.655	1.439	2.457	2.235	1.518	2.688
Subject	8	9	10	11	12	13	14
Residual errors ($\widehat{U}_i^{-1}$)	**6.880**	1.946	3.363	0.668	3.207	2.094	0.495
Random effects ($\widehat{W}_i^{-1}$)	1.662	2.637	1.263	1.266	3.826	2.977	3.607



(a)

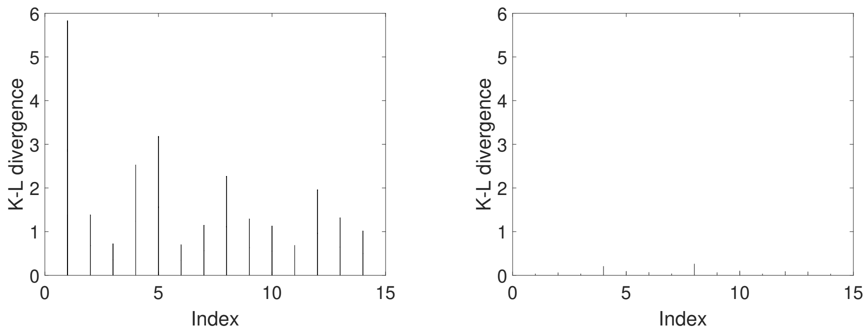
(b)

Figure 2: The numerical solution of the pharmacokinetic mixed-effect ODE model (11) using the estimated ODE parameters and initial conditions for two subjects under the assumption that the ODE parameters and noisy data follow the scale mixture of multivariate normal distributions. The circles are the measured drug concentration. (a) Subject 1; (b) Subject 8.

$(Ka, Ke, Cl)^T$ using our proposed model. As a comparison, an MLE method is implemented on typical PK compartment model of (11) assuming normal distributions and the results are also displayed in Table 1. Compared with Bayesian methods, the maximum likelihood estimates based on normality assumptions have large standard deviations. Our method can also detect the outlying subjects by studying the values of the weights U_i and W_i in our proposed model. Notice that the prior expectations of U_i and W_i are both set to be 1. Hence, the posterior value of U_i substantially below 1 indicates that the i-th subject has outliers. Similarly, the posterior value of W_i substantially below 1 indicates that the i-th subject is an outlying subject. The estimates of U_i^{-1} and W_i^{-1} for our proposed model are displayed in Table 3. Subject 1 has a large value of $\widehat{W}_i^{-1}$, which indicates that subject 1 may be an outlying subject with outlying ODE parameter estimates. However, subject 1 has a small value of $\widehat{U}_i^{-1}$ which indicates that subject 1 has no outlying observations. On the other hand, subjects 8 has a large value of $\widehat{U}_i^{-1}$, which indicates that subject 8 may have outlying observations. Figure 2 displays the estimated serum concentration profiles of these two subjects. Subject 8 has an observed peak drug concentrations higher than the numerical solution of the mixed-effects ODE model using the estimated ODE parameters and the initial condition. Hence, our proposed method has a capability to detect the outlying subject and/or outlying observations.

To determine possible influential observations, we computed the K-L divergence measures for the Normal model and SMN model. The left panel in Figure 3 shows that subject 1, 4, 5, 8 and 12 have much larger $K\{P, P_{(-i)}\}$ in the Normal model in comparison with the SMN model. As expected, the effect of these influential observations on the posterior estimates of ODE parameters were attenuated using the SMN distributions. Hence, our method is robust for estimating mixed-effect ODE models with possible influential observations.

As suggested by the referee, we considered the other prior distributions to study the sensitivity of our method. Gelman (2006) discussed the effects of prior distributions on variance parameters in hierarchical models. Instead of using the inverse-gamma distributions as the “noninformative” priors of vari-



(a) Normal (b) SMN

Figure 3: Index plots of $K\{P, P_{(-i)}\}$ for the IDV600 data set. The left panel is based on Normal distributions and the right panel is based on the SMN distributions.

ance parameters, they suggested to use the half-t family such as half-normal distribution or half-cauchy distribution. Following this idea, we considered a half-normal prior on σ_ϵ. The fitted results were displayed in Table of the supplement file. On the other hand, we also considered an informative priors, called the Penalised Complexity (PC) priors, for κ and ν. The PC priors were first developed by Simpson et al. (2017) which are general enough to be used in realistically complex statistical models and are straightforward enough to be used by general practitioners. The fitted results were displayed in Table S1–S4 of the supplement file, which are similar to the results by assuming an inverse gamma prior on σ_ϵ^2 and gamma priors on κ and ν.

5. Simulation Studies

In this section, we implement some simulation studies to evaluate the finite sample performance of our proposed hierarchical ODE model.

We consider a simple mixed-effects ODE model:

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = -\theta_{i1}X_i(t) + \theta_{i2}, \quad t \in [0, 1]. \tag{12}$$

The true fixed effect is set as $\xi_1 = 3.0$ and $\xi_2 = 10.0$. We generate the individual ODE parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T = (\xi_1, \xi_2)^T + \boldsymbol{\Sigma}^{1/2}(b_{i1}, b_{i2})^T$ where

$\mathbf{\Sigma} = (\mathbf{\Sigma}^{1/2})^2$ and $\mathbf{\Sigma} = (\sigma_{ij})_{2\times 2}$ with $\sigma_{11} = \sigma_{12} = 0.25$ and $\sigma_{22} = 1.0$, and b_{i1}, b_{i2} are independent and identically distributed (i.i.d.) in standardized distribution $F(\cdot)$ for $n = 50$ or 100 subjects. We considered five scenarios for $F(\cdot)$:

(i) The Student's t distribution with the degrees of freedom 4;

(ii) The generalized hyperbolic distribution with *location*=0.0, *scale*=1.0, *skewness*=0.0, *shape*=1.0 and *tail*=5.0;

(iii) The mixture of Student's t distribution, $0.6 \cdot t(3) + 0.4 \cdot t(6)$;

(iv) The inverse Gaussian distribution with *location*=1.0 and *scale*=1.0;

(v) The Birnbaum-Saunders distribution with *shape*=0.5 and *scale*=0.5.

The individual initial condition $X_i(0), i = 1, ..., n$, are independently generated from the same distribution $F(\cdot)$. Then, our simulated data are generated as $Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij}$, where $X_i(t_{ij})$ is the numerical solution of ODE (12) via the fourth-order Runge-Kutta algorithm evaluated at 21 equally-spaced time points on $[0, 1]$, and ϵ_{ij}'s are generated independently from the standardized Student's t distribution with the degrees of freedom 4. We then estimate the mixed-effects ODE (12) by assuming the ODE parameter $\boldsymbol{\theta}_i$ and the measurement error ϵ_{ij} follow the scale mixture of multivariate normal (SMN) distributions. We also compare this proposed model with the conventional model which assumes both $\boldsymbol{\theta}_i$ and ϵ_{ij} follow the normal distributions. With the 'burn-in' of the first 10,000 samples, we obtain 1,000 equally-spaced posterior samples from the rest of the iterations. The above procedure is repeated for 100 simulation replicates.

Due to the limits of space, we only show the simulation results when $F(\cdot)$ is the Student's t distribution at here. The simulation results with respect to other distributions are provided in Tables S5–S6 and Figure S1 of the supplementary file. Table 4 displays the posterior means, standard deviations as well as the mean absolute deviation errors (MADE) for the fixed effect $(\xi_1, \xi_2)^T$. It shows that our proposed model using the SMN distribution has smaller standard deviations and MADEs than the conventional model using the normal distribution, although their posterior means have similar biases. Moreover, the standard deviations and MADEs of fixed effects for both models decrease when the sample size increases from $n = 50$ to $n = 100$. In addition, with simulated data where $n = 50$, we use the LCPO, DIC and WAIC criteria to evaluate the efficiency of model selection when using our method and the conventional methods. To do this, we define

$$\begin{aligned}\Delta_{\text{LCPO}} &= \text{LCPO}_{\text{SMN}} - \text{LCPO}_{\text{Normal}},\\ \Delta_{\text{DIC}} &= \text{DIC}_{\text{SMN}} - \text{DIC}_{\text{Normal}},\\ \Delta_{\text{WAIC}} &= \text{WAIC}_{\text{SMN}} - \text{WAIC}_{\text{Normal}}.\end{aligned}$$

The results are displayed in Figure 4. Remember that a larger value of LCPO or a smaller value of DIC/WAIC indicates a better model. Hence, the proposed method based on the SMN distributions outperforms the conventional method based on the normal distributions.

Table 4: The mean, standard deviation (SD) and mean absolute deviation error (MADE) of estimates for the fixed effects of the mixed-effects ODE model (12) in 100 simulation replicates when assuming the ODE parameters and the data errors follow the scale mixture of multivariate normal (SMN) distributions or the normal distributions. The true values of $(\xi_1, \xi_2)^T$ are $(3.0, 10.0)^T$.

n	Fixed-effects	Distribution assumptions					
		SMN distributions			Normal distributions		
		Mean	SD	MADE	Mean	SD	MADE
50	ξ_1	3.020	0.233	0.182	3.009	0.357	0.283
	ξ_2	10.037	0.621	0.466	10.013	0.948	0.751
100	ξ_1	2.969	0.183	0.148	2.977	0.238	0.187
	ξ_2	9.903	0.499	0.405	9.911	0.636	0.510

As suggested by one reviewer, we also evaluate the prediction accuracy of our method. After obtaining the estimates for ODE parameters and initial

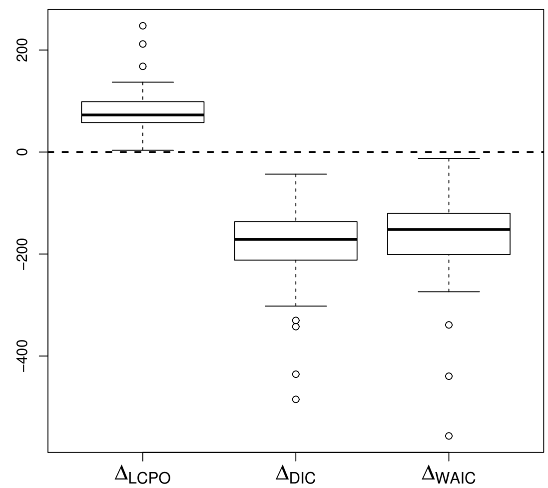


Figure 4: The boxplot of model comparison criteria using the scale mixture of multivariate normal distributions and the traditional normal distributions in Simulation 1, where $\Delta_{\mathrm{LCPO}} = \mathrm{LCPO}_{\mathrm{SMN}} - \mathrm{LCPO}_{\mathrm{Normal}}$, $\Delta_{\mathrm{DIC}} = \mathrm{DIC}_{\mathrm{SMN}} - \mathrm{DIC}_{\mathrm{Normal}}$ and $\Delta_{\mathrm{WAIC}} = \mathrm{WAIC}_{\mathrm{SMN}} - \mathrm{WAIC}_{\mathrm{Normal}}$.

conditions from the simulated data in [0,1], we can solve the ODE numerically in [0,3]. The obtained ODE solution, $\hat{C}_i(t), t \in [1,3]$ can be viewed as the prediction of future observations. Let $C_i(t_j)$ be the true dynamical process at m equally-spaced grid points in [1,3]. The prediction accuracy is quantified with the mean absolute prediction error (MAPE) and the mean squared prediction error (MSPE):

$$\text{MAPE} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}|\hat{C}_i(t_j) - C_i(t_j)|, \quad \text{MSPE} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{C}_i(t_j) - C_i(t_j))^2.$$

We choose $m = 201$ in this simulation study. Table 5 displays the means and standard deviations of MAPE and MSPE for the ODE model (12). It shows that our proposed model using the SMN distribution has smaller prediction errors than the conventional model using the normal distribution.

Table 5: The means and standard deviations (displayed within brackets) of MAPE and MSPE for the mixed-effects ODE model (12) in 100 simulation replicates when assuming the ODE parameters and the data errors follow the scale mixture of multivariate normal (SMN) distributions or the normal distributions.

n	Distribution assumptions	Prediction accuracy criterion	
		MAPE	MSPE
50	SMN	0.219(0.033)	0.374(0.136)
	Normal	0.223(0.034)	0.386(0.140)
100	SMN	0.158(0.015)	0.071(0.034)
	Normal	0.168(0.024)	0.080(0.035)

The number of observed time points plays an important role in modeling ordinary differential equations systems. Further, we considered the simulation of (12) where $n_i = 5, 10, 15$. The simulation results are provided in Tables S7–S8 of the supplement file, which demonstrated that our proposed method works very well. When there are only 3 or 4 time points, our method breaks since that it is impossible to accurately recover the ODE solutions from 3 or 4 observations.

6. Conclusions and Discussions

Ordinary differential equations (ODEs) are elegant and popular models for describing the mechanism of complex dynamical systems. In this paper, we propose a mixed-effects ODE model, which considers the within-subject and between-subject variations simultaneously. We propose to use a class of scale mixture of multivariate normal distributions to model the random effects of ODE parameters and measurement errors in the data to obtain a robust estimation for the ODE parameters when the outlying subjects and the outlying measurement errors exist in the data.

Our proposed model can be framed in a Bayesian hierarchical model by introducing two latent variables. We propose an MCMC algorithm to estimate the ODE parameters. The estimated latent variables enable us to identify outlying subjects and outlying measurement errors. Our proposed method is demonstrated by estimating a mixed-effects ODE model in a pharmacokinetic study. We show that our proposed model using the scale mixture of multivariate normal distribution is preferred in comparison with the conventional model using the normal distribution. Our simulation studies also show that our proposed model can obtain more robust estimation for ODE parameters when using the scale mixture of multivariate normal distributions.

It is common to encounter outlying observations in statistical analysis. To deal with the outlying observations, we consider a class of more flexible distributions like the scale mixtures of normal distributions for data. Another method is to model the distributions with the semiparametric approach, e.g., using the Dirichlet process or a combination of splines and wavelets. This semiparametric approach is more flexible in modelling the skewed or multi-mode distributions. For instance, Castro et al. (2018) proposed a Baysian semiparametric modelling framework for HIV longitudinal data with censoring and skewness. We will investigator this semiparametric approach in our future research. Another interesting work under investigation is to consider the robust estimations of semiparametric mixed-effect ODE models using heavy-tailed distributions with applications in gene regulatory activities. In this project, the ODE model has not only parametric parameters but also time-varying parameters.

Appendix

We use the Markov chain Monte Carlo (MCMC) methods which consist of the Metropolis-Hastings algorithm and the Gibbs sampling method to sample the parameters $\boldsymbol{\theta}_i$, $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}$, σ_ϵ^{-2}, U_i, W_i, κ, ν, λ_κ, and λ_ν. In this appendix, the symbol $\|\mathbf{a}\|_{\mathbf{A}}^2$ denotes $\mathbf{a}^T\mathbf{A}\mathbf{a}$ for the vector $\mathbf{a}$ and the matrix $\mathbf{A}$. When $\mathbf{A} = \mathbf{I}$, a symbol $\|\mathbf{a}\|^2$ is used instead. Define $\mathbf{X}_i = (X_i(t_{i1}), ..., X_i(t_{in_i}))^T, i = 1, ..., n$. The full conditional distributions for $\boldsymbol{\theta}_i$, $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}$, σ_ϵ^{-2}, U_i, W_i, κ, ν, λ_κ and λ_ν are displayed as follows (where $\sim$ denotes all variables except the one to be sampled):

(a) Full conditional distributions of $\boldsymbol{\theta}_i$ for $i = 1, ..., n$.

$$p(\boldsymbol{\theta}_i| \sim) \propto \exp\left\{-\frac{U_i}{2\sigma_\epsilon^2}\|\mathbf{Y}_i - \mathbf{X}_i\|^2\right\}\exp\left\{-\frac{W_i}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|_{\boldsymbol{\Sigma}^{-1}}^2\right\}.$$

(b) Full conditional distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}$.

$$\begin{aligned}
p(\boldsymbol{\xi}| \sim) &\propto \prod_{i=1}^{n}\exp\left\{-\frac{W_i}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|_{\boldsymbol{\Sigma}^{-1}}^2\right\}\exp\left\{-\frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|_{\boldsymbol{\Omega}_0}^2\right\},\\
p(\boldsymbol{\Sigma}| \sim) &\propto |\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}W_i\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|_{\boldsymbol{\Sigma}^{-1}}^2\right\}|\boldsymbol{\Sigma}|^{-(df+q+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{S}_0\boldsymbol{\Sigma}^{-1})\right\}.
\end{aligned}$$

Then the full conditional posterior distribution of $\boldsymbol{\xi}$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{\boldsymbol{\xi}} = \mathbf{B}(\sum_{i=1}^n W_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_i + \boldsymbol{\Omega}_0\boldsymbol{\xi}_0)$ and covariance matrix $\mathbf{B} = (\sum_{i=1}^n W_i\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Omega}_0)^{-1}$. The full conditional posterior distribution of $\boldsymbol{\Sigma}$ is an Inverse Wishart distribution with the scale matrix $\mathbf{S}_0 + \sum_{i=1}^n W_i\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|^2$ and degrees of freedom $n + q + 2$.

(c) Full conditional distributions of U_i and W_i.

$$\begin{aligned}
p(U_i| \sim) &\propto H_1(U_i; \kappa)U_i^{n_i/2}\exp\left\{-\frac{U_i}{2\sigma_\epsilon^2}\|\mathbf{Y}_i - \mathbf{X}_i\|^2\right\},\\
p(W_i| \sim) &\propto H_2(W_i; \nu)W_i^{q/2}\exp\left\{-\frac{W_i}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|_{\boldsymbol{\Sigma}^{-1}}^2\right\}.
\end{aligned}$$

Assuming that $U_i \sim Ga(\kappa/2, \kappa/2)$, then the full conditional posterior distribution of U_i is still a Gamma distribution with shape parameter $n_i/2 + \kappa/2$ and rate parameter $\kappa/2 + \frac{1}{2\sigma_\epsilon^2}\|\mathbf{Y}_i - \mathbf{X}_i\|^2$. Similarly, the full conditional posterior distribution of W_i is a Gamma distribution with shape parameter $\nu/2+q/2$ and rate parameter $\nu/2 + \frac{1}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\xi}\|^2_{\boldsymbol{\Sigma}^{-1}}$.

(d) Full conditional distributions of κ and ν.

$$\begin{aligned} p(\kappa|\sim) &\propto p(\kappa)\prod_{i=1}^{n} H_1(U_i;\kappa), \\ p(\nu|\sim) &\propto p(\nu)\prod_{i=1}^{n} H_2(W_i;\nu). \end{aligned}$$

Assuming that $U_i \sim Ga(\kappa/2, \kappa/2)$ and a truncated exponential prior $\exp(-\lambda_\kappa \cdot \kappa)I(\kappa > 2.0)$ is assigned on κ, then the full conditional posterior distribution of κ is proportional to $(\kappa/2)^{\kappa/2}/\Gamma(\kappa/2)\prod_{i=1}^{n} U_i^{\kappa/2-1}\exp(-\kappa U_i/2)\exp(-\lambda_\kappa\cdot\kappa)I(\kappa > 2.0)$. This is not a standard distribution; however, we can apply the Metropolis-Hastings algorithm to sample it. In the same way, under the assumption of $W_i \sim Ga(\nu/2, \nu/2)$ and the prior $p(\nu) \propto \exp(-\lambda_\nu \cdot \nu)I(\nu > 2.0)$, the full conditional posterior distribution of ν is given by

$$p(\nu|\sim) \propto (\nu/2)^{\nu/2}/\Gamma(\nu/2)\prod_{i=1}^{n} W_i^{\nu/2-1}\exp(-\nu W_i/2)\exp(-\lambda_\nu\cdot\nu)I(\nu > 2.0),$$

which is also sampled by the Metropolis-Hastings algorithm.

(e) Full conditional distributions of λ_κ and λ_ν.

$$\begin{aligned} p(\lambda_\kappa|\sim) &\propto p(\kappa|\lambda_\kappa)\cdot p(\lambda_\kappa), \\ p(\lambda_\nu|\sim) &\propto p(\nu|\lambda_\nu)\cdot p(\lambda_\nu). \end{aligned}$$

Assuming that a truncated exponential prior $\exp(-\lambda_\kappa \cdot \kappa)I(\kappa > 2.0)$ for κ and a Uniform prior distribution $U(c,d)$ for λ_κ, then the full conditional posterior distribution of λ_κ is a truncated Gamma distribution $Ga(2,\kappa)I(c,d)$. Similarly, under the assumption of $p(\nu|\lambda_\nu) \propto \exp(-\lambda_\nu\cdot\nu)I(\nu > 2.0)$ and a Uniform prior distribution $U(c,d)$ for λ_ν, the full conditional posterior distribution of ν is a truncated Gamma distribution $Ga(2,\nu)I(c,d)$.

(f) Sample σ_ϵ^{-2}.

$$p(\sigma_\epsilon^{-2}|\sim) \propto p(\sigma_\epsilon^{-2})(\sigma_\epsilon^{-2})^{N/2} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n} U_i\|\mathbf{Y}_i - \mathbf{X}_i\|^2\right\}.$$

Assuming that σ_ϵ^{-2} has a Gamma prior $Ga(a_0, b_0)$, then the full conditional posterior distribution of σ_ϵ^{-2} is a Gamma distribution with shape parameter $a_0 + N/2$ and rate parameter $b_0 + \frac{1}{2}\sum_{i=1}^{n} U_i\|\mathbf{Y}_i - \mathbf{X}_i\|^2$ where $N = \sum_{i=1}^{n} n_i$.

Generally, in the above Gibbs sampler algorithm, the full conditional distribution in (a) has no closed form. We apply the Metropolis-Hastings method to sample $\boldsymbol{\theta}_i$. The details are as follows: in the ℓth iteration, a candidate, $\boldsymbol{\theta}_i^{cand}$, is generated from a proposal distribution, $q(\cdot|\boldsymbol{\theta}_i^{(\ell-1)})$, like a multivariate normal distribution, $N(\boldsymbol{\theta}_i^{(\ell-1)}, \sigma_0^2\mathbf{I}_q)$, where $\sigma_0^2 > 0$ is a pre-specified scalar to control the acceptance rate. Then, the acceptance probability is calculated by $\alpha(\boldsymbol{\theta}_i^{cand}|\boldsymbol{\theta}_i^{(\ell-1)}) = \min\{1, \frac{p(\boldsymbol{\theta}_i^{cand}|\sim)q(\boldsymbol{\theta}_i^{(\ell-1)}|\boldsymbol{\theta}_i^{cand})}{p(\boldsymbol{\theta}_i^{(\ell-1)}|\sim)q(\boldsymbol{\theta}^{cand}|\boldsymbol{\theta}_i^{(\ell-1)})}\}$. However, this acceptance probability depends on the ODE solution $X_i(t)$ which generally has no explicit expression and has to be obtained numerically. Conditioning on $\boldsymbol{\theta}_i$, $X_i(t)$ is estimated by minimizing Equation (10) numerically.

Acknowledgements

The authors are very grateful to the Editor, the Associate Editor and a reviewer for their very constructive comments. These comments are extremely helpful for us to improve our work. This research was supported by the Liaoning Provincial Education Department (No. LN2017ZD001) to B. Liu and the discovery grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to J. Cao and L. Wang.

Supplementary Files

The simulation programs are included in the supplementary document, which is available with this paper at the Computational Statistics & Data Analysis website on Wiley Online Library.

References

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistics Society, Series B 36, 99–102.

Azzalini, A., Capitanio, A., 2014. The Skew-Normal and Related Families. Chapman and Hall, London.

Bhaumik, P., Ghosal, S., 2015. Bayesian two-step estimation in differential equation models. Electronic Journal of Statistics 9, 3124–3154.

Brunel, N.J., Clairon, Q., d'Alché Buc, F., 2014. Parametric estimation of ordinary differential equations with orthogonality conditions. Journal of the American Statistical Association 109, 173–185.

Burden, R.L., Douglas, F.J., 2000. Numerical Analysis. Brooks/Cole Publishing Company, Pacific Grove, California.

Campbell, D., Steele, R.J., 2012. Smooth functional tempering for nonlinear differential equation models. Statistics and Computing 22, 429–443.

Cancho, V., Dey, D., Lachos, V., , Andrade, M., 2011. Bayesian nonlinear regression models with scale mixtures of skew normal distributions: Estimation and case influence diagnostics. Computational Statistics and Data Analysis 55, 588–602.

Cao, J., Fussmann, G., Ramsay, J.O., 2008. Estimating a predator-prey dynamical model with the parameter cascades method. Biometrics 64, 959–967.

Cao, J., Huang, J.Z., Wu, H., 2012. Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. Journal of Computational and Graphical Statistics 21, 42–56.

Cao, J., Wang, L., Xu, J., 2011. Robust estimation for ordinary differential equation models. Biometrics 67, 1305–1313.

Carlin, B.P., Louis, T.A., 2008. Bayesian Methods for Data Analysis. Third ed., Chapman/Hall, London.

Castro, L.M., Wang, W.L., Lachos, V.H., Ináio de Carvalho, W., Bayes, C.L., 2018. Bayesian semiparametric modeling for hiv longitudinal data with censoring and skewness. Statistical Methods in Medical Research doi:10.1177/0962280218760360.

Chen, J., Wu, H., 2008. Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. Journal of the American Statistical Association 103, 369–383.

Chen, M.H., Shao, Q.M., Ibrahim, J.G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag Inc., New York.

Choy, S.T.B., Smith, A.F.M., 1997. Hierarchical models with scale mixtures of normal distributions. Test 6, 205–221.

De la Cruza, R., 2014. Bayesian analysis for nonlinear mixed-effects models under heavy-tailed distributions. Pharmaceutical Statistics 13, 81–93.

Dass, S.C., Lee, J., Lee, K., Park, J., 2017. Laplace based approximate posterior inference for differential equation models. Statistics and Computing 27, 679–698.

Fang, Y., Wu, H., Zhu, L.X., 2011. A two-stage estimation method for random-coefficient differential equation models with application to longitudinal hiv dynamic data. Statistica Sinica 21, 1145–1170.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515–534.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for bayesian models. Statistics and Computing 24, 997–1016.

Guedj, J., Thiébaut, R., Commenges, D., 2007. Maximum likelihood estimation in dynamical models of hiv. Biometrics 63, 1198–1206.

Hall, P., Ma, Y., 2014. Quick and easy kernel based one-step estimation of parameters in differential equations. Journal of the Royal Statistical Society, Series B 76, 735–748.

Huang, Y., Liu, D., Wu, H., 2006. Hierachical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics 62, 413–423.

Huang, Y., Wu, H., 2006. A bayesian approach for estimating antiviral efficacy in hiv dynamic models. Journal of Applied Statistics 33, 155–174.

Lachos, V.H., Bandyopadhyay, D., Dey, D.K., 2011. Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions. Biometrics 67, 1594–1604.

Lahiri, S.N., 2003. A necessary and sufficient condition for asymptotic independence of discrete fourier transforms under short- and long-range dependece. The Annals of Statistics 31, 613–641.

Lange, K., Sinsheimer, J., 1993. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics 2, 175–198.

Lange, K.L., Little, R.J.A., Taylor, J.M.G., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84, 881–896.

Li, L., Brown, M.B., Lee, K.H., Gupta, S., 2002. Estimation and inference for a spline-enhanced population pharmacokinetic model. Biometrics 58, 601–611.

Li, Y., Zhu, J., Wang, N., 2015. Regularized semiparametric estimation for ordinary differential equations. Technometrics 57, 341–350.

Liang, H., Wu, H., 2008. Parameter estimation for differential equation models using a framework of measurement error in regression. Journal of the American Statistical Association 103, 1570–1583.

Liu, C., 1996. Bayesian robust multivariate linear regression with incomplete data. Journal of the American Statistical Association 91, 1219–1227.

Massuia, M.B., Garay, A.M., Lachos, V.H., Cabral, C.R., 2017. Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions. Statistics and its interface 10, 425–439.

Meza, C., Osorio, F., De la Cruz, R., 2012. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Statistics and Computing 22, 121–139.

Peng, F., Dey, D.K., 1995. Bayesian analysis of outlier problems using divergence measures. The Canadian Journal of Statistics 23, 199–213.

Perelson, A.S., Nelson, P.W., 1999. Mathematical analysis of hiv-1 dynamics in vivo. SIAM review 41, 3–44.

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., Ho, D.D., 1996. Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271, 1582–1586.

Putter, H., Heisterkamp, S.H., Lange, J.M., De Wolf, F., 2002. A bayesian approach to parameter estimation in hiv dynamical models. Statistics in Medicine 21, 2199–2214.

Ramsay, J.O., Hooker, G., Campbell, D., Cao, J., 2007. Parameter estimation for differential equations: a generalized smoothing approach (with discussion). Journal of the Royal Statistical Society, Series B 69, 741–796.

Rosa, G.J.M., Gianola, D., Padovani, C.R., 2004. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using mcmc. Journal of Applied Statistics 31, 855–873.

Rosa, G.J.M., Padovani, C.R., Gianola, D., 2003. Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. Biometrical Journal 45, 573–590.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science 32, 1–28.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistics Society, Series B 64, 583–639.

Wang, L., Cao, J., Ramsay, J.O., Burger, D., Laporte, C., Rockstrohk, J., 2014. Estimating mixed-effects differential equation models. Statistics and Computing 24, 111–121.

Wasmuth, J., la Porte, C.J., Schneider, K., Burger, D.M., Rockstroh, J.K., 2004. Comparison of two reduced-dose regimens of indinavir (600 mg vs. 400 mg twice daily) and ritonavir (100 mg twice daily) in healthy volunteers (coredir). International Medical Press 2, 1359–6535.

Watanabe, S., 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.

Wu, H., Ding, A., 1999. Population hiv-1 dynamics in vivo: applicable models and inferential tools for virological data from aids clinical trials. Biometrics 55, 410–418.