Computational Statistics (2020) 35:1481–1499
https://doi.org/10.1007/s00180-020-00962-8

ORIGINAL PAPER

A Bayesian approach to estimate parameters of ordinary differential equation

Hanwen Huang[1] · **Andreas Handel**[1] · **Xiao Song**[1]

Received: 15 March 2018 / Accepted: 28 January 2020 / Published online: 10 February 2020

Abstract
We develop a Bayesian approach to estimate the parameters of ordinary differential equations (ODE) from the observed noisy data. Our method does not need to solve ODE directly. We replace the ODE constraint with a probability expression and combine it with the nonparametric data fitting procedure into a joint likelihood framework. One advantage of the proposed method is that for some ODE systems, one can obtain closed form conditional posterior distributions for all variables which substantially reduce the computational cost and facilitate the convergence process. An efficient Riemann manifold based hybrid Monte Carlo scheme is implemented to generate samples for variables whose conditional posterior distribution cannot be written in terms of closed form. Our approach can be applied to situations where the state variables are only partially observed. The usefulness of the proposed method is demonstrated through applications to both simulated and real data.

Keywords Noisy data · ODE constraint · Nonparametric fitting · Joint likelihood framework · Hybrid Monte Carlo

1 Introduction

Ordinary differential equation (ODE) is a simple but powerful framework for modeling the interactions of complex dynamic systems and has been widely used in many scientific fields including engineering, physics and biomedical sciences. In practice, we often need to estimate the parameters of ODE models from the observational data. This is an important but challenging statistical problem because the observed state

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-020-00962-8) contains supplementary material, which is available to authorized users.

✉ Hanwen Huang
huanghw@uga.edu

[1] Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602, USA

variables are not exactly the ODE solution but rather measured with noise. In addition, many ODE equations are nonlinear and thus do not have closed form solutions.

Two general strategies are commonly used to tackle parameter estimation problem for ODE models. The first strategy focuses on solving the ODE and uses the least square principle to fit the ODE solution to the observed data (Hemker 1972; Bates and Watts 1988; Seber and Wild 1989; Li et al. 2005). If the ODE has analytical solution, it is equivalent to a standard nonlinear least square problem. However, in practice, most of the ODEs do not have closed form solutions due to their nonlinear features. In this case, numerical methods such as the Runge-Kutta algorithm (see Hairer et al. 1993; Matteij and Molenaar 2002) have to be used to approximate the solution of the ODEs. This method has highly computational cost because the numerical solutions of ODEs have to be obtained for every updating of the parameters. In addition, the initial values of the state variables have to be estimated in this method in order to solve the ODE and this greatly increases the number of the unknown parameters in the model.

In the second strategy, ODE does not need to be solved explicitly, instead, smooth nonparametric methods are used. For example, Varah proposed a two-stage smoothing method to estimate the ODE parameters (Varah 1982). In the first stage, regression splines were used to estimate the state variables and their derivatives. In the second stage, parameters were estimated by minimizing the distance between the estimated derivatives and the derivatives determined by the ODE. Liang and Wu (2008) developed a similar idea but explored the kernel-based nonparametric methods to estimate the regression functions and their derivatives. As pointed by Dondelinger et al. (2013) and Wang and Barber (2014), a drawback of the two-step method is that the nonparametric fit in the first step is based on the data alone without any feedback mechanism from the ODE system. This limitation leads to rather poor parameter estimation from data subject to heavy noise. Another drawback of this method is that it cannot be applied to situations where some components are missing. To overcome this problem, Ramsay et al. (2007) proposed a parameter cascade approach to estimate the dynamic parameters in ODE models. In this approach, the smoothing of the state variables and the estimation of the ODE parameters are considered jointly. A linear combination of spline basis functions is employed to approximate the regression function and parameters are estimated by requiring that the regression functions satisfy the differential equation and fit the data simultaneously as well as possible. The second strategy avoids the high computational cost of repeatedly solving ODE. However, its optimization task is challenging. Efficient optimization techniques are required in practice to obtain the estimator and the convergence of the computational algorithms needs to be justified. The conventional gradient-based optimization methods such as the conjugate-gradient method and Qausi-Newton method may fail to converge or may converge to a local minimum if the initial values of the unknown parameters are not chosen to be close enough to the true values.

On the other hand, Bayesian based approaches may escape the local minimum in the optimization surface as mentioned by Gelman et al. (1996). For the first strategy, some Bayesian methods based on the likelihood centered on the numerical ODE solution of the state variables have been proposed in the literature, examples include Huang et al. (2006), Huang and Wu (2006) among many others. Since the state variables have no closed form solution, the posterior distribution for the parameters have no closed

form so the sampling has to be based on Metroplis-Hastings algorithm which is quite inefficient especially in high dimensional situations. In addition, at each iteration when new values for the parameters are proposed, the numerical ODE solution has to be obtained in order to compute the likelihood. This is also quite computationally expensive.

Bayesian analysis for the second strategy has not been studied until recently when Gaussian Process (GP) methods were considered as data models to infer the ODE parameters (Calderhead et al. 2009; Dondelinger et al. 2013; Chkrebtii et al. 2016; Wang and Barber 2014; Schober et al. 2014; Wang and Barber 2014). GPs provide a distribution over both the fitted functions and associated gradients. Using priors on the parameters of the GP model and the ODE model, this gives a flexible Bayesian parameter estimation procedure. In Calderhead et al. (2009), GP parameters are first fitted to the data, and subsequently the parameters of the ODE are estimated. The estimation accuracy of this approach is limited by the lack of feedback from ODE parameter inference to GP parameter inference. To address this, Dondelinger et al. (2013) and Wang and Barber (2014) introduced bidirectional interaction between ODE and GP parameters, demonstrating improved parameter estimation. However, these GP based approaches have similar computational complexity and all use Metroplis-Hastings algorithm in sampling the state variables and parameters which leaves room for improvement. Bhaumik and Ghosal (2015); Ranciati et al. (2016) developed two-stage Bayesian methods in which samples of the state variables were first generated from nonparametric regression and then the posterior distribution of the ODE parameters were determined by matching the fitted curves through ODE constraints. The main drawback of the approach is that the state variables are solely determined by the observed data without the influence of the ODE parameters.

We propose to improve on previous approaches by introducing a one-step generative model that directly combines the smoothness, system observations and ODE together. We replace the ODE constraints with a probability expression and combine it with nonparametric regression together into a joint likelihood function. The benefit of this Bayesian approach to parameter estimation in ODEs can be well-established. For some special ODE formulations, we can get closed form conditional distribution for every variable which substantially facilitates the convergence of MCMC process. For more complicated ODE models, we propose to use hybrid Monte Carlo scheme based on Hamiltonian dynamics to improve the acceptance rate. Hybrid Monte Carlo requires the computation of the derivative of the corresponding conditional distribution function for each variable, which can also be derived in closed forms in our one-step Bayesian model. Another advantage of our proposed model is that it can be applied to situations where only partial components of the state variables are observed. A similar idea has been applied to Partial Differential Equation Models in Xun et al. (2013) and ODE models in Mazur et al. (2009) and Campbell and Steele (2012). Mazur et al. (2009) only considered a single state variable and Campbell and Steele (2012), Xun et al. (2013) employed completely different computation methods.

The remaining of the paper is organized as follows. Section 2 is devoted to the detailed formulation of our Bayesian method. The Markov chain Monte Carlo procedure for sampling the posterior distribution is provided in Sect. 2.3. Especially, the implementation of the newly developed Riemann manifold based hybrid Monte Carlo

scheme is introduced in details. The performance of the proposed method is tested in Sect. 3 for simulated data and in Sect. 4 for real data. The paper is concluded with a discussion in Sect. 5.

2 Model formulation

2.1 ODE system description

In a general continuous time dynamical system, the evolution of K states $\mathbf{x}(t) \equiv \{x_1(t), \ldots, x_K(t)\}$ is described by a set of K ODEs

$$\dot{\mathbf{x}}(t) \equiv \frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t), \tag{1}$$

where $\boldsymbol{\theta} \in R^p$ is the parameter vector of ODE and $\mathbf{f}(\cdot) = \{f_1(\cdot), \ldots, f_K(\cdot)\}$ is a vector of known appropriately smoothing functions. Without loss of generality, assume that the first K_0 states are measured at time points $t_1, \ldots, t_n$. A common situation in practice is that only part of the system is measured, i.e. $K_0 < K$. For the k-th observed state, the n observations $y_k(t_1), \ldots, y_k(t_n)$ are obtained according to independent additive noise model $y_k(t) = x_k(t) + \epsilon_k(t)$ where the noise $\epsilon_k(t)$ is Gaussian with a state-specific error variance σ_k^2, i.e., $\epsilon_k(t) \sim N(0, \sigma_k^2)$. Other noise models are possible though not trivial. Further define the state matrix $\mathbf{X} \equiv [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_n)]$ and the observation matrix $\mathbf{Y} \equiv [\mathbf{y}(t_1), \ldots, \mathbf{y}(t_n)]$, where $\mathbf{y}(t) \equiv (y_1(t), \ldots, y_{K_0}(t))^T$. This gives an observation model

$$p_{OBS}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{K_0}\prod_{i=1}^{n} p_{OBS}(y_k(t_i)|x_k(t_i), \sigma_k^2), \tag{2}$$

where $p_{OBS}(y_k(t)|x_k(t), \sigma_k^2) = N(y_k(t)|x_k(t), \sigma_k^2)$. Given potentially noisy observations $\mathbf{Y}$, we aim to estimating the parameters $\boldsymbol{\theta}$. Parameter estimation for ODE system is challenging because typically Eq. (1) does not have closed form solutions for a general nonlinear function $\mathbf{f}(\cdot)$, thus it is difficult to apply traditional nonlinear least square method.

2.2 Bayesian ODE parameter estimation approach

In our proposed method, we first generate an estimator of $\mathbf{x}_k$ from the observation $\mathbf{y}_k$ using nonparametric regression in which $\mathbf{x}_k$ is expressed in terms of a basis function expansion

$$x_k(t) = \sum_{j=1}^{q} c_{kj}\phi_j(t) = \mathbf{c}_k^T\boldsymbol{\phi}(t), \tag{3}$$

where $\boldsymbol{\phi}(t)$ is a vector of q basis functions, the common choices of which include polynomial bases, B-spline bases, and Laguerre functions. The derivative of $x_k(t)$ can be written as

$$\dot{x}_k(t) = \mathbf{c}_k^T \dot{\boldsymbol{\phi}}(t). \tag{4}$$

Denote matrix $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]$. The likelihood function of the model is

$$p_{OBS}(\mathbf{Y}|\mathbf{C}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{K_0} \prod_{i=1}^{n} N(y_k(t_i)|\mathbf{c}_k^T \boldsymbol{\phi}(t_i), \sigma_k^2). \tag{5}$$

We tackle the inferential procedure with a Bayesian approach by assigning prior probabilities to the parameters $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \ldots, \sigma_K^2\}$ and $\mathbf{C}$. For each σ_k^2, we employ an improper prior $p(\sigma_k^2) = 1/\sigma_k^2$. For each $\mathbf{c}_k$, we employ a Gaussian prior distribution of the form

$$p(\mathbf{c}_k|\lambda_k) = N(0, \boldsymbol{\Sigma}/\lambda_k^2). \tag{6}$$

The hyperparameter matrix $\boldsymbol{\Sigma} = \int_{t_1}^{t_n} \boldsymbol{\phi}''(t)[\boldsymbol{\phi}''(t)]^T dt$ which is defined in the way that we want the basis to be penalized. The parameter λ_k controls the trade-off between smoothness of the functions $x_k(t)$ and fit of the noisy data. The prior for λ_k is

$$p(\lambda_k) = \text{Gamma}(\alpha_k, \beta_k), \tag{7}$$

for some shape and rate hyperparameters (α_k, β_k).

Assuming additive Gaussian noise with a state specific error variance γ_k^2, one can include ODE model (1) using

$$p_{ODE}(\mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\gamma}^2) \approx \prod_{k=1}^{K} \prod_{i=1}^{n} \frac{1}{\gamma_k} \exp\left[-\frac{\{\mathbf{c}_k^T \dot{\boldsymbol{\phi}}(t_i) - f_k(\mathbf{C}^T \boldsymbol{\phi}(t_i), \boldsymbol{\theta}, t_i)\}^2}{2\gamma_k^2} \right], \tag{8}$$

where "$\approx$" means "equal up to a constant" and $\boldsymbol{\gamma}^2 = \{\gamma_1^2, \ldots, \gamma_K^2\}$. Here we replace the exact ODE constraint (1) with the probability expression (8) which models a normal distribution for the discrepancy between $\dot{\mathbf{x}}(t)$ and $\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t)$. As $\gamma_k^2 \to 0$, we get an exact ODE constraint for the k-th component. For each γ_k^2, we employ an inverse Gamma prior

$$p(1/\gamma_k^2) = \text{Gamma}(\alpha_\gamma, \beta_\gamma),$$

for some shape and rate hyperparameters $(\alpha_\gamma, \beta_\gamma)$. For ODE parameters $\boldsymbol{\theta}$, we assume the following prior distribution

$$p(\boldsymbol{\theta}|\lambda_\theta) = N(0, \mathbf{I}_p/\lambda_\theta^2), \tag{9}$$

where $\mathbf{I}_p$ denotes the p-dimensional identity matrix. This is equivalent to Bayesian ridge regression with λ_θ as a penalizing term. The prior distribution for λ_θ is

$$p(1/\lambda_\theta^2) = \text{Gamma}(\alpha_\theta, \beta_\theta), \tag{10}$$

with hyperparameters $(\alpha_\theta, \beta_\theta)$.

Combining observation model (5), basis expansions (3) and (4), and ODE requirement (8) together, we propose the following joint distribution of the whole system over state parameters $\mathbf{C}$, ODE parameters $\boldsymbol{\theta}$, observations $\mathbf{Y}$, and remaining parameters

$$\begin{aligned} p(\mathbf{Y}, \mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}^2, \boldsymbol{\sigma}^2, \lambda_\theta) = {} & p_{OBS}(\mathbf{Y}|\mathbf{C}, \boldsymbol{\sigma}^2) p_{ODE}(\mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\gamma}^2) p(\mathbf{C}|\boldsymbol{\lambda}) \\ & p(\boldsymbol{\theta}|\lambda_\theta^2) p(\boldsymbol{\lambda}^2) p(\boldsymbol{\gamma}^2) p(\boldsymbol{\sigma}^2) p(\lambda_\theta^2), \end{aligned} \tag{11}$$

where $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_K\}$ and

$$p(\mathbf{C}|\boldsymbol{\lambda}^2) = \prod_{k=1}^{K} p(\mathbf{c}_k|\lambda_k^2), \quad p(\boldsymbol{\lambda}^2) p(\boldsymbol{\gamma}^2) p(\boldsymbol{\sigma}^2) = \prod_{k=1}^{K} p(\lambda_k^2) p(\gamma_k^2) p(\sigma_k^2).$$

In this way we combine a smoothness assumption on the state variable $\mathbf{X}$ together with derivative information obtained from the ODE in a single model. For those components of the $\mathbf{X}$ which are not observed, they are not included in $p_{OBS}(\mathbf{Y}|\mathbf{C}, \boldsymbol{\sigma}^2)$ but can still be estimated because they appear in $p_{ODE}(\mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\gamma}^2)$.

The main difference between our model and the traditional Bayesian ODE model (Huang et al. 2006; Huang and Wu 2006) is that we replace the strict ODE requirement (1) by the probabilistic expression (8). Our method is also different from the two-step Bayesian approach proposed in Bhaumik and Ghosal (2015) which first sample the state variables using Bayesian nonparametric method, and then infer the ODE parameters deterministically by minimizing the distance between the derivatives of the state variables and those predicted from the ODEs. The drawback of the two-step approach is that the ODE parameters never feed back into the the first step and thus have no bearing on the inference of the state variables. In contrast, our model is described by a one-step expression (11) where the interpolation of the state variables fits both the noisy data and the derivatives from the ODEs simultaneously, allowing the system of ODEs to feed back into the interpolation. The smoothness of the state variables, data fitting, and ODE constraint are balanced in the Bayesian model (11) through the change of three parameters $\boldsymbol{\lambda}$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\gamma}^2$. Instead of being treated as tuning parameters like in Ramsay et al. (2007), they all can be estimated adaptively through MCMC procedure in our model.

2.3 Parameter estimation

Statistical inference can be obtained based on the joint posterior distribution $p(\mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}^2, \boldsymbol{\sigma}^2, \lambda_\theta|\mathbf{Y})$ which will be estimated using Markov Chain Monte Carlo (MCMC) sampling scheme.

The full conditional posterior distributions of $\boldsymbol{\sigma}^2$, $\boldsymbol{\gamma}^2$, $\boldsymbol{\lambda}$, and λ_θ are inverse Gammas. Generally, the distribution (11) is a complicated function of the state variables $\mathbf{X}$ owing to the nonlinear dependence via $\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t)$. The full conditional distributions of $\mathbf{X}$ and $\boldsymbol{\theta}$ depend on the explicit form of function $\mathbf{f}(\cdot)$. For situations where $\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t)$ only

depends on each argument up to the first order of power, we can derive a closed form full conditional posterior distribution for each component of $\mathbf{X}$ and $\boldsymbol{\theta}$. For example, if $\mathbf{f}(\cdot)$ is a linear function of any component of $\mathbf{X}$, the full conditional distribution of the corresponding component will be multivariate normal. Similarly, if $\mathbf{f}(\cdot)$ is a linear function of any component of $\boldsymbol{\theta}$, the full conditional distribution of that component will be normal as well. However, in some applications where $\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t)$ tends to have more complicated forms, there is no standard form distribution. The standard Metropolis-Hastings method using a simple random-walk proposal distribution is not efficient if the state variables are highly correlated among different time points due to the slow exploration of the state space. To bypasses this problem, we exploit the efficient hybrid Monte Carlo (HMC) method, which is also called Hamiltonian Monte Carlo method to update the state variables and model parameters that do not have closed form full conditional posterior distribution.

To draw samples for a random variable $\mathbf{q} \in R^d$ with density $p(\mathbf{q})$, HMC introduces an independent auxiliary variable $\mathbf{p} \in R^d$ with density $p(\mathbf{p}) = N(\mathbf{p}|0, \mathbf{M})$, where $\mathbf{M}$ is a symmetric, positive-define mass matrix. The joint density follows in factorized form as $p(\mathbf{q}, \mathbf{p}) = p(\mathbf{q})p(\mathbf{p})$ and the negative joint log-probability is

$$H(\mathbf{q}, \mathbf{p}) = -\log(p(\mathbf{q})) + \frac{1}{2}\log((2\pi)^d|\mathbf{M}|) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}.$$

From the joint distribution of $\mathbf{q}$ and $\mathbf{p}$, negating $\mathbf{p}$ variable, we will get the marginal distribution of $\mathbf{q}$. The physical analog of this negative joint log-probability is a Hamilton which describes the movement of a particle with position $\mathbf{q}$ and momentum $\mathbf{p}$ in a potential $-\log(p(\mathbf{q}))$. The time evolution of the system with respective to a fictitious time τ is determined by Hamilton's equations

$$\begin{aligned} \frac{d\mathbf{q}}{d\tau} &= \nabla_{\mathbf{p}} H(\mathbf{q}, \mathbf{p}) = \mathbf{M}^{-1}\mathbf{p}, \\ \frac{d\mathbf{p}}{d\tau} &= -\nabla_{\mathbf{q}} H(\mathbf{q}, \mathbf{p}) = \nabla_{\mathbf{q}} \log(p(\mathbf{q})). \end{aligned} \tag{12}$$

The HMC procedure is to first assume that the current state is at time 0, denoted by $(\mathbf{q}(0), \mathbf{p}(0))$, and propose a new state at time τ as $(\mathbf{q}(\tau), \mathbf{p}(\tau))$ by solving the Hamilton equation (12). Then this proposed state is accepted as the next state of the Markov chain with probability

$$\min[1, \exp\{-H(\mathbf{q}(\tau), \mathbf{p}(\tau)) + H(\mathbf{q}(0), \mathbf{p}(0))\}].$$

If the proposed state is not accepted, the next state is the same as the current state.

For practical applications of interest, the differential equations (12) cannot be solved analytically and numerical methods are required. For implementation, Hamilton's equations are approximated by discretizing time, using some small step-size, ϵ and starting with the state at time 0, iteratively computing the state at time $\epsilon, 2\epsilon$, until time τ. The choice of the step size ϵ and number of integration steps can be tuned on the basis of the overall acceptance rate of the HMC sampler. Traditionally, when $\mathbf{M}$ is diagonal,

a leapfrog method can be implemented to update the state at each step (Neal 1994). However, as mentioned in Girolami and Calderhead (2011), if $\mathbf{M}$ is chosen adaptively by allowing its dependence on the position variable $\mathbf{q}$, significant improvement can be achieved for the overall mixing of the chain. Instead of the diagonal matrix, they choose $\mathbf{M}$ to be the Fisher information matrix at position $\mathbf{q}$ and propose a generalized leapfrog algorithm which updates the state at each step using

$$\begin{aligned}
\mathbf{p}(\epsilon/2) &= \mathbf{p}(0) - \frac{\epsilon}{2} \bigtriangledown_{\mathbf{q}} H(\mathbf{q}(0), \mathbf{p}(\epsilon/2)), \\
\mathbf{q}(\epsilon) &= \mathbf{q}(0) + \frac{\epsilon}{2}[\bigtriangledown_{\mathbf{p}} H(q(0), \mathbf{p}(\epsilon/2)) + \bigtriangledown_{\mathbf{p}} H(q(\epsilon), \mathbf{p}(\epsilon/2))], \\
\mathbf{p}(\epsilon) &= \mathbf{p}(\epsilon/2) - \frac{\epsilon}{2} \bigtriangledown_{\mathbf{q}} H(\mathbf{q}(\epsilon), \mathbf{p}(\epsilon/2)).
\end{aligned}$$

This modified HMC scheme is also called Riemann manifold hybrid Monte Carlo (RMHMC) which can give better performance than standard HMC scheme in situations where the components of $\mathbf{x}$ are highly correlated.

The required condition for the HMC sampling scheme is that we need to have closed form first and second order derivatives for the full conditional density function. Clearly this is satisfied in our method (11). For example, the conditional distribution of the spline coefficients $\mathbf{c}_k$ is

$$p(\mathbf{c}_k) \approx p(\mathbf{c}_k|\lambda_k) p_{ODE}(\mathbf{C}|\boldsymbol{\theta}, \boldsymbol{\gamma}^2) \prod_{i=1}^{n} N(y_k(t_i)|\mathbf{c}_k^T \boldsymbol{\phi}(t_i), \sigma_k^2).$$

To sample $\mathbf{c}_k$, we need to compute the derivative $\bigtriangledown_{\mathbf{c}_k} \log(p(\mathbf{c}_k))$, which has a closed form expression. The Fisher information matrix of $p(\mathbf{c}_k)$ can also be derived in a closed form. Therefore, the implementation of the RMHMC algorithms is fairly straightforward for our method.

3 Simulation

In our simulation studies, we illustrate the performance of our method on three dynamical systems: Lotka–Volterra model, complex reaction model, and Fitzhugh–Nagumo model. We applied our method to simulated data and test the performance by comparing the estimated parameters and state variables with the true ones. In all our numerical studies, we use 11 cubic B-spline bases to approximate the state variables. Our numerical studies have shown that the results are not sensitive to the number of basis functions used in the nonparametric fit. Priors for the Bayesian methods are taken to be almost noninformative, i.e. the parameters $\alpha_k, \beta_k, \alpha_\theta, \beta_\theta, \alpha_\gamma, \beta_\gamma$ in the Gamma priors are chosen to be 0.01. For the analysis of the data, the MCMC sampler was run for a total of 11,000 cycles. The first 1000 cycles were discarded as burn-in, and the remainder of the chain was thinned by keeping one out of every ten samples, resulting in a total of 1000 samples for post-MCMC analysis. Our simulations have also shown that the results are not sensitive to the choice of the starting points for the

MCMC sampler and our MCMC algorithms converge fast. For comparison, we also include the results based on the generalized profiling approach (GP) of Ramsay et al. (2007) and the classic nonlinear least squares method (NLS) by using the R packages *CollocInfer* and *nloptr* respectively. For the GP method, we choose the tuning parameter λ through a grid search such that the smallest residual sum square are obtained. For the NLS method, in addition to the ODE parameters, the initial values of the state variables also need to be estimated using the optimization package. For each of the following configurations, we run 100 replications. To evaluate the performance of different methods, we define the average relative estimation error (ARE) of a parameter θ as $ARE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{\theta}_i - \theta|}{|\theta|}$, where $\hat{\theta}_i$ is the estimate of θ for the i-th replication and N is the number of simulation runs (here $N = 100$).

3.1 Lotka–Volterra model

The Lotka–Volterra model is an ecological system that is used to describe the periodical interaction between a prey species S and a predator species W (Lotka 1909)

$$\begin{aligned} \dot{S}(t) &= S(t)(\alpha - \beta W(t)), \\ \dot{W}(t) &= -W(t)(\gamma - \delta S(t)), \end{aligned} \tag{13}$$

where $\boldsymbol{\theta} = [\alpha, \beta, \gamma, \delta]^T$ and $\mathbf{X} = [S, W]$. The observed data are generated using numerical integration over the interval [0,4] with $\boldsymbol{\theta} = [2, 1, 4, 1]$ and initial state values $S(0) = 5$ and $W(0) = 3$. Then we add Gaussian noise $N(0, \sigma^2)$ to the numerical ODE solution to form the observation $\mathbf{Y}$. We set $\sigma = 0.2$ and $\sigma = 1$ to represent situations of small noise and large noise respectively. Data are collected at each time point with interval 0.1. For this model, the function $\mathbf{f}(\cdot)$ is linear with respect to each of its arguments. Thus we can get closed form expression for full conditional distributions of all variables. The full conditional distributions for $\boldsymbol{\theta}$, $S(t)$ and $W(t)$ are all multivariate normals. This substantially facilitates the convergence speed of MCMC. We consider two scenarios. In the first scenario, both S and W are observed. In the second scenario, only S is observed and the observation for W is missing. Then by eliminating variable $W(t)$ in (13), we can show that the parameter β is not identifiable. Meaning that for a given data, there are more than one set of parameters resulting in a similar data fit. However, if the initial value $W(0)$ is known, then all four parameters are structurally identifiable.

We summarize the point estimation results in Table 1 and the associated AREs in Table 2. The upper blocks are for the first scenario and the lower blocks are for the second scenario. Table 1 reports the averages of the estimated values, standard errors, and the average time for analyzing each data set. The 95% highest posterior density intervals are also included for our Bayesian method. For the second scenario, we didn't include the GP results because the software does not provide the option on how to get the output with the initial value $W(0)$ fixed. Table 1 shows that the point estimates of the ODE parameters are reasonable close to the true values for all methods when the noise is small. In terms of the speed, the fastest method is GP. The computational time increases with the noise for NLS but changes very little for Bayesian and GP.

Table 1 Summary table of the estimated ODE coefficients for Lotka–Volterra model based on 100 simulated data sets for $\alpha = 2, \beta = 1, \gamma = 4, \delta = 1$

σ	Method	α	β	γ	δ	Time
0.2	Bayesian	2.00 (0.10)	1.00 (0.05)	3.76 (0.26)	0.93 (0.07)	18.1
		[1.60, 2.41]	[0.93, 1.06]	[3.20, 4.33]	[0.82, 1.04]	
	NLS	2.02 (0.11)	1.01 (0.05)	3.98 (0.21)	1.00 (0.05)	12.0
	GP	1.88 (0.12)	0.94 (0.05)	3.39 (0.28)	0.84 (0.07)	7.8
1.0	Bayesian	1.52 (0.47)	0.78 (0.22)	3.14 (1.28)	0.77 (0.35)	18.0
		[0.94, 2.13]	[0.55, 1.02]	[1.75, 4.53]	[0.43, 1.10]	
	NLS	1.78 (0.83)	0.88 (0.40)	4.08 (1.97)	1.01 (0.49)	35.5
	GP	1.92 (0.58)	0.93 (0.27)	3.67 (1.62)	0.89 (0.42)	7.0
0.2	Bayesian	1.70 (0.28)	0.87 (0.12)	4.35 (0.50)	1.07 (0.13)	18.7
		[0.66, 2.77]	[0.73, 1.02]	[3.18, 5.47]	[0.82, 1.30]	
	NLS	2.85 (0.46)	1.28 (0.15)	2.87 (0.66)	0.72 (0.16)	20.0
1.0	Bayesian	2.58 (1.34)	1.17 (0.48)	3.50 (1.30)	0.83 (0.34)	18.5
		[1.13, 4.07]	[0.91, 1.44]	[2.35, 4.66]	[0.61, 1.10]	
	NLS	2.62 (1.37)	1.20 (0.46)	6.09 (6.97)	1.51 (1.72)	26.7

The numbers in parentheses denote standard deviations over 100 samples. The numbers in the square brackets denote the average 95% posterior credible intervals for the proposed Bayesian method. Time denotes the average time for analyzing each data set in unit of seconds

From Table 2, we can see that in the first scenario, NLS gives the smallest ARE if the noise is small and GP gives the smallest ARE if the noise is large. The performances of Bayesian are in-between in both situations. In the second scenario, Bayesian is consistently better than NLS.

To evaluate the goodness of fit, we obtained the predicted values of S and W by numerically integrating ODEs (13) using estimated parameters. We present the predicted curves for the case of $\sigma^2 = 0.5$ and the corresponding true curves (by solving the ODEs using the true parameter values) in Fig. 1 in which the associated 95% confidence intervals of these state variables are also delineated. The upper panel is for results based on estimation using observed data of both S component and W component. The lower panel is for results estimated from S component alone with known $W(0)$. We can see that the predicted curves of S and W have good agreement with the corresponding true curves. Figure 1 shows that our method can give quite reasonable estimation even in situations where some components of state variables are not observable. This is one of the big advantages of our Bayesian method.

3.2 Complex reaction model

This model describes the complex reaction with segregation in a semi-batch reactor which includes five state variables y_1, y_2, y_3, y_4, y_5 and four parameters $\beta, D_a, \theta_{mix}, \eta$. The dynamics of the system can be described by the following differential equations (Schittkowski 2002)

Table 2 Relative error for the simulated data from the Lotka–Volterra model

σ	Method	α	β	γ	δ
0.2	Bayesian	0.04	0.04	0.07	0.09
	NLS	0.04	0.04	0.04	0.04
	GP	0.07	0.06	0.15	0.16
1.0	Bayesian	0.29	0.27	0.32	0.35
	NLS	0.31	0.31	0.35	0.36
	GP	0.24	0.24	0.34	0.36
0.2	Bayesian	0.18	0.16	0.13	0.12
	NLS	0.45	0.29	0.31	0.31
1.0	Bayesian	0.41	0.34	0.28	0.30
	NLS	0.65	0.43	1.05	1.04

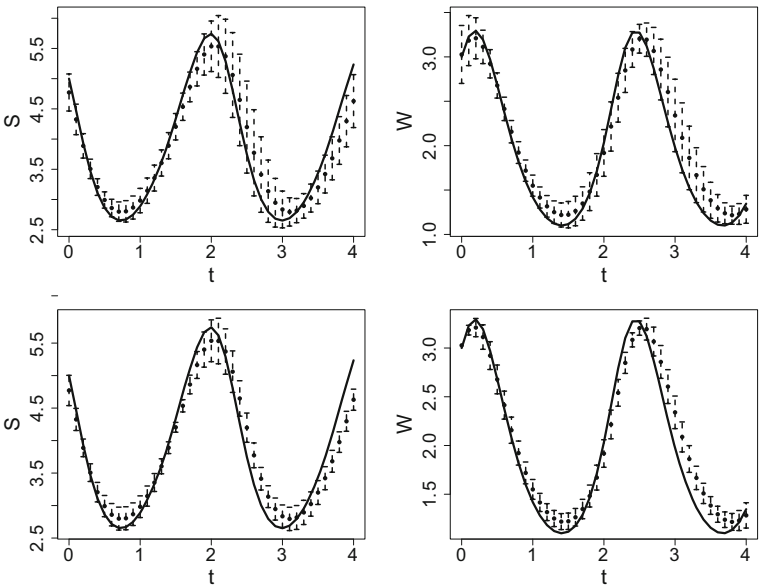

Fig. 1 Bayesian inference for Lotka–Volterra model. The results for state variables $S(t)$ and $W(t)$ are shown in the left and right panel respectively. The solid lines indicate the true trajectories. The solid circles are the estimated trajectories with 95% confidence intervals represented by the error bars. The results based on the observed data of both S and W components are shown in the upper panel. The results based on observed data of S component alone and $W(0)$ are shown in the lower panel

$$\dot{y}_1(t) = \frac{1 - y_1(t)}{\Theta(t)} - \frac{y_1(t)}{\theta_{mix}},$$
$$\dot{y}_2(t) = \frac{1 - y_2(t)}{A(t)} - \beta D_a(y_2(t)y_3(t) - \eta y_2(t)),$$
$$\dot{y}_3(t) = \frac{\eta - y_3(t)}{A(t)} - \beta D_a(y_2(t)y_3(t) - \eta y_1(t)) - D_a y_2(t)y_4(t)$$

Table 3 Summary table of the estimated ODE coefficients for complex reaction model based on 100 simulated data sets for $\beta = 10$, $D_a = 5$, $\theta_{mix} = 3$, $\eta = 0.3$

σ	Method	β	D_a	θ_{mix}	η	Time
0.1	Bayesian	7.41 (0.76)	4.51 (0.37)	2.95 (0.20)	0.31 (0.02)	31.0
		[7.13, 7.70]	[4.37, 4.65]	[2.84, 3.06]	[0.31, 0.32]	
	NLS	5.26 (5.00)	4.57 (0.54)	2.97 (0.08)	0.45 (0.17)	224.9
	GP	3.51 (1.75)	4.16 (0.61)	3.00 (0.12)	0.29 (0.12)	469.8
0.5	Bayesian	2.37 (1.58)	3.60 (1.61)	2.61 (0.83)	0.30 (0.17)	29.3
		[1.89, 2.87]	[3.14, 4.05]	[2.14, 3.09]	[0.23, 0.38]	
	NLS	19.9 (61.85)	4.44 (1.18)	3.09 (0.44)	0.45 (0.19)	152.7
	GP	0.14 (1.32)	1.63 (0.97)	2.71 (0.58)	−0.10 (0.33)	233.7

The notations are the same as in Table 1

Table 4 Relative error for the simulated data from the complex reaction model

σ	Method	β	D_a	θ_{mix}	η
0.2	Bayesian	0.09	0.08	0.09	0.05
	NLS	0.53	0.11	0.02	0.54
	GP	0.29	0.14	0.07	0.14
0.5	Bayesian	0.60	0.41	0.27	0.67
	NLS	2.00	0.21	0.11	0.63
	GP	1.44	0.81	0.17	5.92

$$\dot{y}_4(t) = \frac{-y_4(t)}{A(t)} + \beta D_a(y_2(t)y_3(t) - \eta y_1(t)) - D_a y_3(t)y_4(t)$$
$$\dot{y}_5(t) = \frac{-y_5(t)}{A(t)} + D_a y_3(t)y_4(t),$$

where $\Theta(t) = t + 0.01$ and

$$A(t) = \begin{cases} \Theta(t) & if\ \Theta(t) \le 1.0 \\ 1 & otherwise \end{cases}.$$

We generate data using parameters $[\beta, D_a, \theta_{mix}, \eta] = [10, 5, 3, 0.3]$ and initial state values $[y_1(0), y_2(0), y_3(0), y_4(0), y_5(0)] = [1, 1, 1, 1, 1]$. The Gaussian noise is generated from $N(0, \sigma^2)$. Data are collected at every 0.03 time unit on the interval $[0, 3]$. We consider two situations: $\sigma = 0.1$ and $\sigma = 0.5$. The averages of the estimated ODE parameters, standard errors, 95% posterior credible intervals, and average analysis times are summarized in Table 3 based on 100 simulated data sets. Table 4 shows the corresponding AREs. From Table 3, we can see that our Bayesian method is much faster than the optimization based NLS and GP methods in solving this estimation problem. Table 4 shows that our method gives the smallest AREs for β, D_a, η in situation of $\sigma = 0.1$ and also the smallest AREs for β in situation of $\sigma = 0.5$. In other cases, our estimations are not substantially worse. Figure 2 compares the curves using the estimated parameters with the curves using the true parameters for five state

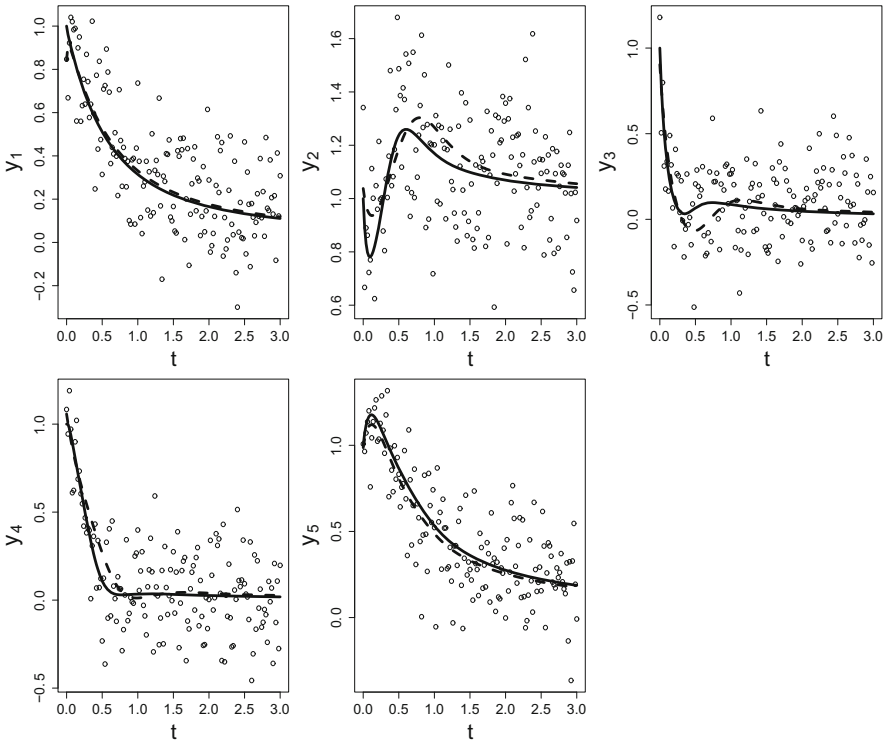

Fig. 2 Bayesian inference for complex reaction model. The results for five state variables $y_1(t)$, $y_2(t)$, $y_3(t)$, $y_4(t)$, and $y_5(t)$ are shown in five panels respectively. The circles represent the observed data. The solid lines indicate the true trajectories and the dashed lines represent the estimated trajectories based on the Bayesian method

variables based on one simulated example in the case of $\sigma = 0.2$. The results show that our estimation can successfully recover the shapes of state variables.

3.3 Fitzhugh–Nagumo model

The Fitzhugh–Nagumo equations describe the behavior of spike potentials in the giant axon of squid neurons (Ramsay et al. 2007):

$$\dot{V}(t) = c\left(V(t) - \frac{V(t)^3}{3}\right) + R(t),$$
$$\dot{R}(t) = -(V(t) - a + bR(t)),$$

where V describes the voltage across an axon membrane and R is a recovery variable summarizing outward currents. The parameter values used to generate the data are $a = 0.2, b = 0.1, c = 3$ and $V(0) = 1, R(0) = 3$. The simulated values are then perturbed with Gaussian noise from $N(0, \sigma^2)$. The data consists of 201 evenly-

Table 5 Summary table of the estimated ODE coefficients for Fitzhugh–Nagumo model based on 100 simulated data sets

σ	Method	a	b	c	Time
0.1	Bayesian	0.19 (0.01)	0.10 (0.01)	2.75 (0.08)	50.0
		[0.05, 0.32]	[0.08, 0.12]	[2.62, 2.88]	
	NLS	0.07 (0.09)	0.12 (0.03)	3.00 (0.07)	126.4
	GP	0.29 (0.07)	0.20 (0.02)	2.45 (0.11)	8.7
0.5	Bayesian	0.19 (0.06)	0.10 (0.03)	1.40 (0.27)	48.9
		[0.05, 0.33]	[0.02, 0.18]	[1.06, 1.75]	
	NLS	0.13 (0.19)	0.12 (0.12)	2.91 (0.41)	99.0
	GP	0.11 (0.22)	0.16 (0.08)	0.66 (0.30)	12.1

The notations are the same as in Table 1

Table 6 Relative error for the simulated data from the Fitzhugh–Nagumo model

σ	Method	a	b	c
0.1	Bayesian	0.08	0.05	0.08
	NLS	0.67	0.32	0.02
	GP	0.46	1.03	0.18
0.5	Bayesian	0.22	0.24	0.53
	NLS	0.89	0.83	0.08
	GP	0.95	0.77	0.78

spaced observations in the interval [0, 20]. For this ODE system, the full conditional distributions for state variable $R(t)$ and parameters a, b and c are multivariate normals based on our model (11). But there is no closed form expression for state variable $V(t)$. So special attention needs to be taken. We use RMHMC method proposed in Girolami and Calderhead (2011) to generate posterior samples for $V(t)$. The mass matrix $\mathbf{M}$ is chosen to be the Fisher information matrix at the current value of the position variable.

We consider two situations: $\sigma = 0.1$ and $\sigma = 0.5$. The averages of the estimated ODE parameters, standard errors, 95% posterior credible intervals, and average analysis times are summarized in Table 5 based on 100 simulated data sets. Table 6 shows the corresponding AREs. From Tables 5 and 6, we can see that GP is much faster than Bayesian and NLS. With respect to the accuracy, Bayesian gives the smallest errors for the estimation of parameters a and b while NLS gives the smallest errors for the estimation of parameter c in both cases.

4 Real data application

4.1 England and Wales measles data

The first application of the proposed method is to the data of England and Wales weekly case reports of measles from 1948 to 1965 taken from http://ns.mcmaster.ca/~bolker/measdata.html. We use the following model:

$$\begin{aligned}\dot{S}(t) &= -\beta S(t)I(t),\\ \dot{I}(t) &= -\beta S(t)I(t) - \delta I(t),\end{aligned} \tag{14}$$

where $S(t)$ and $I(t)$ stand for the number of susceptible and infected individuals at time t respectively. The parameter β represents the infecting rate of the susceptible individuals while the parameter δ represents the recovering rate of the infected individuals. Since the right hand sides of euqations in (14) are linear function of all ODE parameters as well as the state variables, all posterior samples can be generated from closed form distribution in our MCMC procedure, which substantially reduced the computational cost.

In this study, S was not measured and only I was collected over time. Similar to Dattner (2015), we focus on the observations for 53 weeks for the years 1948–1949. Therefore, n is taken to be 53. For model (14) with observed $I(t)$, parameter δ is not identifiable and its estimation depends on the intial value $S(0)$. To verify this, we take integration over the first equation of (14) and obtain

$$S(t) = S(0)\exp\left(-\int_0^t \beta I(s)ds\right). \tag{15}$$

Then substitute it into the second equation, we obtain

$$\dot{I}(t) = \left\{\beta S(0)\exp\left(-\int_0^t \beta I(s)ds\right) - \delta)\right\} I(t). \tag{16}$$

Clearly, for fixed $I(t)$, different combinations of β, $S(0)$ and δ can satisfy the same equation. Therefore, the solution for (16) is not unique. In practice, δ is usually considered as fixed and known, then the solution of β and $S(0)$ are unique. In Dattner (2015), it was assumed that an individual experiences one recovery in 5 days which is equivalent to setting $\delta = 7/5$ because the data are reported weekly. Table 2 shows how the estimated parameters change with the given recovery time using our Bayesian method. The parameters are estimated by taking average over 1000 posterior samples. It is indicated from Table 1 that $\hat{\beta}$, $\hat{S}(0)$, and $\hat{I}(0)$ increases, decreases and does not change with δ respectively. Dattner (2015) used the nonlinear least square method (NLS) with carefully chosen initial values to obtain $\hat{\beta} = 3.87 \times 10^{-7}$, $\hat{S}(0) = 3.95 \times 10^{6}$, and $\hat{I}(0) = 2012$, which are very close to our results. Our method is computationally much cheaper because instead of numerically solving the ODE, we only need to draw Gibbs sampling from closed form distribution in each step (Table 7).

The solutions of (14) with respect to the final estimated parameters for fixed $\delta = 7/5$ are displayed in Fig. 3. The resulting data fit looks very reasonable.

4.2 HIV dynamics data from an AIDS clinical trial

The second application is to the HIV dynamics data from an AIDS clinical trial in which HIV-1 infected patients were recruited to be treated by antiviral therapies and immune-based treatment. This study measured HIV viral load $V(t)$ and total CD4 T

Table 7 Estimated ODE coefficients and initial state values for different given recovery times in the England and Wales measles study

Recovery time in days	δ	$\hat{\beta}$	$\hat{S}(0)$	$\hat{I}(0)$
3	7/3	2.41×10^{-7}	1.02×10^{7}	2595
4	7/4	3.11×10^{-7}	6.04×10^{6}	2595
5	7/5	3.86×10^{-7}	3.96×10^{6}	2595
6	7/6	4.62×10^{-7}	2.80×10^{6}	2595
7	1	5.39×10^{-7}	2.10×10^{6}	2595

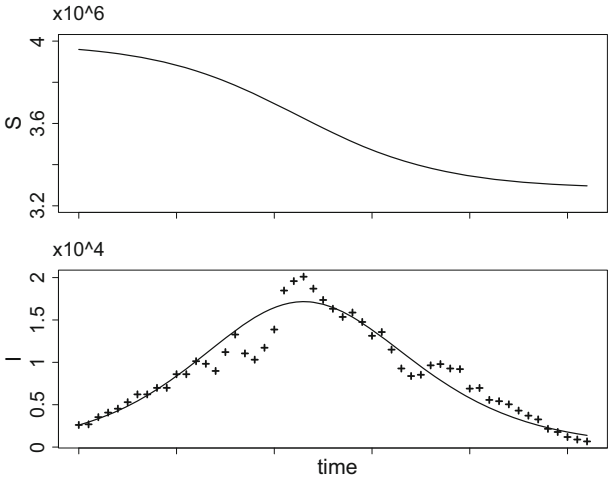

Fig. 3 The resulting data fit of our method. The solid lines represent the solution of (14) with respect to the estimated parameters $\hat{\beta} = 3.86 \times 10^{-7}$, $\hat{S}(0) = 3.96 \times 10^{6}$, and $\hat{I}(0) = 2595$ while observations are displayed with plus signs

cell counts $T(t)$. As discussed in Liang and Wu (2008), the HIV dynamics can be described by

$$V'(t) = \alpha_0 + \alpha_1 T(t) + \alpha_2 T'(t) - cV(t), \tag{17}$$

where the functions $V(t)$ and $T(t)$ are state variables and $(\alpha_0, \alpha_1, \alpha_2, c)$ are unknown dynamic parameters. If we obtain the estimates of $(\alpha_0, \alpha_1, \alpha_2)$, we can derive the estimates of important viral dynamics parameters using the relationships: $\lambda = -\alpha_0/\alpha_2$, $\rho = \alpha_1/\alpha_2$. Here λ represents the rate at which new T cells are continuously generated, ρ is the death rate of T cells, and c is the clearance rate of free virions. We fitted model (17) to the viral load data using the proposed Bayesian method and present the parameter estimation results as follows: $\lambda = 45.5$, s.e. 3.1, and 95% posterior credible interval [39.9, 52.3]; $\rho = 0.085$, s.e. 0.006, and 95% posterior credible interval [0.074, 0.098]; $c = 0.2937$, s.e. 0.0003, and 95% posterior credible interval [0.2929, 0.2943]. Our estimates are quite similar to the results reported in Liang and

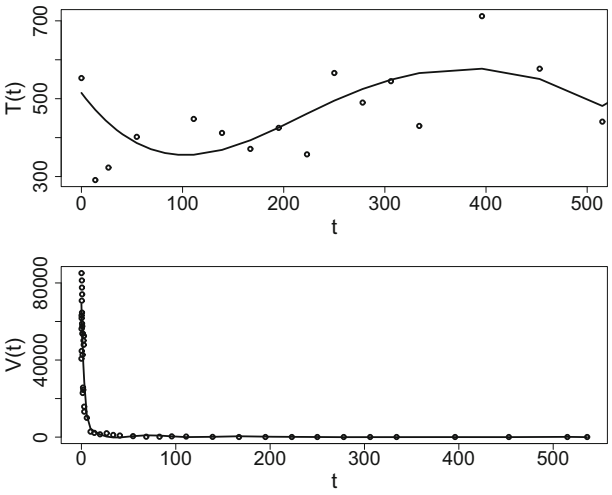

Fig. 4 Fitted curves of $T(t)$ and $V(t)$ for a patient from an HIV dynamics study. Circles indicate the observations and the solid lines are the fitted curves by the Bayesian method

Wu (2008) by using the PsLS method and SIMEX methods. The fitted curves of viral load and total CD4 T cell counts are shown in Fig. 4. The results show consistence of our estimation with the observed data.

5 Discussion

In this paper, we develop a one-step Bayesian approach to estimate the ODE parameters based on observed noisy data. In contrast to the previous two-step Bayesian methods, our method jointly analyzes the nonparametric data fitting and ODE equations based on a single likelihood function. The main advantage is that we allow the system of ODEs to act back in an adaptive manner on the nonparametric interpolation of the state variables over the observed data. In this way, we address the main shortcoming of the method proposed in Bhaumik and Ghosal (2015). Our next step is to study the asymptotic properties of the proposed and establish Bernstein-von-Mises theorem for the posterior distribution of the ODE parameter estimation.

We also compared our method to two popular existing methods: the standard NLS method (Bates and Watts 1988; Seber and Wild 1989) and penalized spline method (Ramsay and Silverman 2005; Ramsay et al. 2007; Ramsay and Hooker 2017). Although our method is not the best in all situations in terms of accuracy and speed, we do provide an alternative estimation approach to avoid some critical problems of these existing methods. Particularly, our method avoids the sensitivity of initial values of the state variables on the parameter estimation as well as the high computational cost due to solving the ODEs numerically or due to the complicated optimization techniques. However, as pointed in Liang and Wu (2008), one limitation of the smoothing spline-based methods is that they require frequent measurement data of state variables.

One possible solution is to combine the proposed method with some existing methods such as NLS to overcome this limitation.

Another big advantage of our proposed framework is that it can naturally incorporate the hierarchical mixed-effect to allow the investigation of variability among different levels. Toward this end, we only need to decompose the ODE parameters $\boldsymbol{\theta}$ into two parts in our model (11), one for fixed effect and one for random effect. Then an extra layer of MCMC sampling procedure needs to be introduced to accomplish the task. We will investigate these issues in a new paper.

Acknowledgements The authors thank the editor, associate editor, and three referees for many helpful comments and suggestions which led to a much improved presentation. This research is supported in part by Division of Mathematical Sciences (National Science Foundation) Grant DMS-1916411 (Huang, Song).

References

Bates D, Watts D (1988) Nonlinear regression analysis and its applications. Wiley, New York

Bhaumik P, Ghosal S (2015) Bayesian two-step estimation in differential equation models. Electron J Stat 9(2):3124–3154

Calderhead B, Girolami M, Lawrence ND (2009) Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. Adv Neural Inf Process Syst 21:217–224

Campbell D, Steele RJ (2012) Smooth functional tempering for nonlinear differential equation models. Stat Comput 22(2):429–443

Chkrebtii OA, Campbell DA, Calderhead B, Girolami MA (2016) Bayesian solution uncertainty quantification for differential equations. Bayesian Anal 11(4):1239–1267

Dattner I (2015) A model-based initial guess for estimating parameters in systems of ordinary differential equations. Biometrics 71(4):1176–1184

Dondelinger F, Husmeier D, Rogers S, Filippone M (2013) ODE parameter inference using adaptive gradient matching with Gaussian processes. In: Carvalho CM, Ravikumar P (eds) Proceedings of the sixteenth international conference on artificial intelligence and statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29–May 1, 2013, pp 216–228. JMLR.org

Gelman A, Bois F, Jiang J (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. J Am Stat Assoc 91(436):1400–1412

Girolami M, Calderhead B (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J R Stat Soc Ser B Stat Methodol 73(2):1-37

Hairer E, Nørsett SP, Wanner G (1993) Solving ordinary differential equations I (2nd revised). Nonstiff problems. Springer, Berlin

Hemker P (1972) Numerical methods for differential equations in system simulation and in parameter estimation. In: Hemker HC, Hess B (eds) Analysis and simulation of biochemical systems. Elsevier, North Holland, pp 59–80

Huang Y, Wu H (2006) A bayesian approach for estimating antiviral efficacy in HIV dynamic models. J Appl Stat 33(2):155–174

Huang Y, Liu D, Wu H (2006) Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics 62(2):413–423

Li Z, Osborne MR, Pravan T (2005) Parameter estimation of ordinary differential equations. IMA J Numer Anal 25:264–285

Liang H, Wu H (2008) Parameter estimation for differential equation models using a framework of measurement error in regression models. J Am Stat Assoc 103(484):1570–1583

Lotka AJ (1909) Contribution to the theory of periodic reactions. J Phys Chem 14(3):271–274

Matteij R, Molenaar J (2002) Ordinary differential equations in theory and practice. SIAM, Philadelphia

Mazur J, Ritter D, Reinelt G, Kaderali L (2009) Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. BMC Bioinform 10(1):448

Neal RM (1994) An improved acceptance procedure for the hybrid monte carlo algorithm. J Comput Phys 111(1):194–203

Ramsay J, Hooker G (2017) Dynamic data analysis: modeling data with differential equations. Springer, New York
Ramsay J, Silverman B (2005) Functional data analysis. Springer, New York
Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. J R Stat Soc Ser B Stat Methodol 69(5):741–796
Ranciati S, Viroli C, Wit E (2016) Bayesian smooth-and-match strategy for ordinary differential equations models that are linear in the parameters. Preprint: arXiv:1604.02318
Schittkowski K (2002) Numerical data fitting in dynamical systems: a practical introduction with applications and software. Kluwer, Norwell
Schober M, Duvenaud D, Hennig P (2014) Probabilistic ODE solvers with Runge–Kutta means. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) Advances in neural information processing systems 27. Curran Associates Inc, Red Hook, pp 739–747
Seber GAF, Wild CJ (1989) Nonlinear regression. Wiley, New York
Varah JM (1982) A spline least squares method for numerical parameter estimation in differential equations. SIAM J Sci Stat Comput 3(1):28–46
Wang Y, Barber D (2014) Gaussian processes for Bayesian estimation in ordinary differential equations. In: Xing EP, Jebara T (eds) Proceedings of the 31st international conference on international conference on machine learning—volume 32, ICML'14, Beijing, China, pp II-1485–II-1493. JMLR.org
Xun X, Cao J, Mallick B, Maity A, Carroll RJ (2013) Parameter estimation of partial differential equation models. J Am Stat Assoc 108(503):1009–1020