

Inference for Differential Equation Models using Relaxation via Dynamical Systems

Kyoungjae Lee¹, Jaeyong Lee², and Sarat Dass³

¹Department of Applied and Computational Mathematics and Statistics, The University
of Notre Dame

²Department of Statistics, Seoul National University

³Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS

October 16, 2018

Abstract

Statistical regression models whose mean functions are represented by ordinary differential equations (ODEs) can be used to describe phenomena dynamical in nature, which are abundant in areas such as biology, climatology and genetics. The estimation of parameters of ODE based models is essential for understanding its dynamics, but the lack of an analytical solution of the ODE makes the parameter estimation challenging. The aim of this paper is to propose a general and fast framework of statistical inference for ODE based models by relaxation of the underlying ODE system. Relaxation is achieved by a properly chosen numerical procedure, such as the Runge-Kutta, and by introducing additive Gaussian noises with small variances. Consequently, filtering methods can be applied to obtain the posterior distribution of the parameters in the Bayesian framework. The main advantage of the proposed method is computation speed. In a simulation study, the proposed method was at least 14 times faster than the other methods. Theoretical results which guarantee the convergence of the posterior of the approximated dynamical system to the posterior of true model are presented. Explicit expressions are given that relate the order and the mesh size of the Runge-Kutta procedure to the rate of convergence of the approximated posterior as a function of sample size.

Key words: Ordinary differential equation, Dynamic model, Runge-Kutta Method, Extended Liu and West filter

1 Introduction

Many dynamical phenomena in the real world can be represented mathematically by ordinary differential equations (ODEs). Common examples include Newton's law of cooling, Lotka-Volterra equations for predator-prey populations (Alligood et al., 1997) and Lorenz equation for atmospheric convection (Lorenz, 1963). There are many other popular examples describing physical, chemical and biological phenomena using ODEs. Although observing the data sets from an ODE system is common, estimating the parameters of ODE models (ODEMs) can be challenging because of lack of an analytical solution to ODE. Here, we give a brief review of previous works on the ODEMs.

There are several frequentist methods in the literature for parameter estimation of ODEMs. Bard (1974) used numerical integration to approximate the solution of ODEs and minimized the objective function based on a gradient method. Varah (1982) suggested a two step estimation method using the cubic spline approximation. The two steps consist of estimation of the regression function and estimation of the parameters in the ODEM. Ramsay and Silverman (2005) modified the first step of Varah by adding the roughness penalty function which measures the difference between the ODE and the mean function. The parameter cascading method was proposed by Ramsay et al. (2007). They grouped the parameters into the regression coefficients, structural parameters, and regularization parameters. The parameters in each group are estimated in turn in a cascading fashion.

Bayesian inference of ODEMs is more challenging because naive application of Markov Chain Monte Carlo (MCMC) methods would require calculation of the numerical solution of ODE whenever parameters are sampled from the proposal distribution. Gelman et al. (1996) and Huang et al. (2006) proposed a Bayesian computation method for parameter inference of pharmacokinetic models and the longitudinal HIV dynamic system, respectively. Campbell (2007) combined the parallel tempering (Geyer, 1991) and collocation method (Ramsay et al., 2007) to get over the rough surface of the posterior, but this slows down the speed of computations significantly. Arnold et al. (2013) used particle filter framework for the inference of ODEMs with linear multistep methods for the numerical integration. Dass et al. (2017) suggested a Bayesian inference with Laplace approximation for a fast computation when the dimension of θ is moderate.

In this paper, we propose a Bayesian inference method for the ODEMs using a relaxation technique via dynamical systems and associated dynamic models. Relaxation is achieved by a properly chosen numerical procedure, such as the Runge-Kutta, and by introducing additive Gaussian noise variables with variance tending to zero. The variance of the additive noise variables works as a measure of fidelity to the original ODEM and by letting it tend to zero, we recover the original model. The relaxation introduces inefficiency of the inference, but we gain the speed of the computation in return.

For a fast computation, a filtering method is applied for inferring posterior distributions of parameters in a Bayesian framework. The relaxation technique provides a dynamical system and model to which a fast inference tool based on sequential Monte Carlo can be applied to. With these sequential methods, we do not need to calculate the whole path of the numerical solution for each realization of the new parameter. It reduces the computation time significantly compared to other standard Bayesian procedures and enables us to deal with the ODEM in reasonable computing time. In subsection 5.2, to emphasize its fast computation the proposed method is compared with the other methods: the parameter cascading, the delayed rejection adaptive Metropolis algorithm and the Bayesian inference with the Laplace approximation. In the simulation study, the proposed method is from 14 times to 78 times faster than other methods.

We also derive convergence results for the approximated posteriors under suitable regularity conditions. We present a guideline for the choice of the model parameters which give a reasonable relative error rate, and provide its theoretical basis. Theoretical results which guarantee the convergence of the posterior of the approximated dynamical system to the posterior of true model are presented. Explicit expressions are given that relate the order and the mesh size of the Runge-Kutta procedure and guarantee the rate of convergence of the approximated posterior to the true posterior.

The rest of the paper is organized as follows. In section 2, we describe a differential equation model and its corresponding relaxed dynamic model counterpart as well as prior choices. The method of posterior inference is described in section 3. Some theoretical support for the proposed method are given in section 4. In section 5, we give two simulated data examples to demonstrate the speed and performance of the proposed method. A real data set, the Lynx-Hare data set, is analyzed in section 6. The discussion is given in section 7. The proofs of theorems are given in

the appendix.

2 Ordinary Differential Equation Models and Nonlinear Dynamic Models

2.1 Ordinary Differential Equation Models (ODEMs)

The ODEM is the regression model with regression function $x(t)$ described by an ODE. The regression function $x(t)$ is the solution of the differential equation

$$\dot{x}(t) = f(x, u, t; \theta), \quad (1)$$

where f is a p -dimensional smooth function, $u(t)$ is a deterministic input function, $\theta \in \Theta \subset \mathbb{R}^q$ is the unknown parameter, and $\dot{x}(t)$ denotes the first derivative of $x(t)$ with respect to time t . Since the input function $u(t)$ does not affect the general ideas of inference in this paper, it is not considered subsequently. The data are observed at n points in the time interval $t \in [0, T] \subset \mathbb{R}$, given by $0 \leq t_1, t_2, \dots, t_n \leq T$. Thus,

$$y_i = x(t_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where y_i is a p -dimensional observation vector at time t_i , the error ϵ_i is drawn independently from the multivariate normal distribution $N_p(0, \sigma^2 I_p)$ with unknown $\sigma^2 > 0$, and $x(t_i) \equiv x_i$ is the underlying regression function measured at time t_i .

The regression model is given by

$$\begin{aligned} y_i &= x_i + \epsilon_i, \quad i = 1, \dots, n, \\ \dot{x}(t) &= f(x, u, t; \theta) \end{aligned} \quad (2)$$

where $x_i = x(t_i)$. The covariate x_i is determined by the initial value of x , $x_0 = x(0)$, and the parameter θ . In the rest of the paper, we call the model (2) as the regression model or the true model.

In most cases, ODE (1) does not have a closed form solution, so there is a need to approximate $x(t)$ numerically. We will use the Runge-Kutta method which is a standard numerical method for ODE. While there are many types of Runge-Kutta methods, we will only consider the 4th order method in this paper. However, our proposed method can be extended to the other

approximation methods for ODE as well as other Runge-Kutta methods with different orders easily. Letting $h_{i+1} = t_{i+1} - t_i$, the form of 4th order Runge-Kutta approximation for (2) is as follows:

$$x_{i+1} \equiv g(x_i, t_i; \theta) = x_i + \frac{1}{6}(k_{i1} + 2k_{i2} + 2k_{i3} + k_{i4}), \quad i = 0, \dots, n-1, \quad (3)$$

where

$$\begin{aligned} k_{i1} &= h_{i+1}f(x_i, t_i; \theta), \\ k_{i2} &= h_{i+1}f\left(x_i + \frac{1}{2}k_{i1}, t_i + \frac{1}{2}h_{i+1}; \theta\right), \\ k_{i3} &= h_{i+1}f\left(x_i + \frac{1}{2}k_{i2}, t_i + \frac{1}{2}h_{i+1}; \theta\right), \\ k_{i4} &= h_{i+1}f(x_i + k_{i3}, t_i + h_{i+1}; \theta). \end{aligned}$$

In the above equation, all x_i 's indicate the approximated values. For more details, see Spijker (1996).

With this approximation, we have the following model

$$\begin{aligned} y_i &= x_i + \epsilon_i, \quad i = 1, \dots, n, \\ x_{i+1} &= g(x_i, t_i; \theta), \quad i = 0, \dots, n-1. \end{aligned} \quad (4)$$

In the remainder of this paper, we call the model (4) as a differential equation model (DEM). Sometimes to obtain better approximation of x_{i+1} , we divide the interval $[t_{i-1}, t_i]$ into m small subintervals and apply the Runge-Kutta method for the subintervals. In this case, we will call the corresponding ODE model the m step ODE model and m the step size.

2.2 Nonlinear Dynamic Models

In practice, estimating the parameter from DEM can pose a significant computational challenge if the ODE does not have an analytical solution. Dass et al. (2017) marginalized out x_0 using Laplace approximation and conducted grid sampling to get posterior samples of θ . Their method is fast and accurate when the dimension of θ is small; however, the methodology suffers from heavy computations when the dimension of θ is large. The computation time increases exponentially as the dimension of θ increases due to the grid sampling. The griddy Gibbs sampler can be used on θ , but practical problems such as dependencies and slow convergence may arise.

In this paper, in order to make posterior inference on θ , we adopt a nonlinear dynamic model relaxation of the DEM in (4) given in terms of the model below with unknown initial condition x_0 :

$$\begin{aligned} y_i &= \tilde{x}_i + \epsilon_i, \quad i = 1, \dots, n, \\ \tilde{x}_{i+1} &= g(\tilde{x}_i, t_i; \theta) + \eta_i, \quad i = 0, \dots, n-1 \end{aligned} \tag{5}$$

where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2 I_p)$ and $\eta_i \stackrel{iid}{\sim} N(0, u^2 I_p)$ with $\sigma, u > 0$. The error term η_i reflects the fact that the approximation $g(x_i, t_i; \theta)$ of x_{i+1} is made with uncertainty. In the remainder of the paper, we call model (5) as the approximate dynamic model obtained as a relaxation of the DEM in (4) via the relaxation parameter u . The quantities \tilde{x}_i in (5) are not the same as x_i given in (4) since the former are quantities that are observed with error whereas the latter are not. However, note that the two models (4) and (5) become equivalent as the relaxation parameter $u \rightarrow 0$.

In the above model (5), there are four unknown quantities, namely, $x_0, \theta, \lambda = 1/\sigma^2$ and u . The Bayesian approach proceeds by considering priors for these quantities. We do not consider a prior for the relaxation parameter u since it is artificially introduced to control the quality of the approximation. We fix u to be a small positive quantity in the subsequent numerical computations. The priors on x_0 and λ are taken as

$$\begin{aligned} x_0 | \lambda &\sim N_p(\mu_{x_0}, c\lambda^{-1} I_p) \quad \text{and} \\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \end{aligned} \tag{6}$$

where $c > 0$ and $\text{Gamma}(a, b)$ represents the Gamma distribution with mean a/b and variance a/b^2 . The prior for θ , $\pi(\theta)$, is taken independently of the rest of the unknown quantities above.

2.3 Sequential Monte Carlo

Sequential Monte Carlo (SMC) is a simulation-based method for estimating the states and the parameters of the nonlinear dynamic model. The basic idea of SMC is using the importance samples to approximate posterior at each state and updating the samples sequentially through a proper kernel. There exists an extensive literature on SMC which includes sequential importance sampling (Handschin and Mayne, 1969), bootstrap filter (Gordon et al., 1993), auxiliary particle filter (Pitt and Shephard, 1999), Rao-Blackwellised particle filter (Doucet et al., 2000), sequential Monte Carlo sampler (Del Moral et al., 2006), Liu and West filter (Liu and West, 2001), particle

learning (Carvalho et al., 2010), multilevel sequential Monte Carlo sampler (Beskos et al., 2016), to name just a few. For an extensive review of SMC, see Doucet et al. (2001), Kantas et al. (2009), Lopes and Tsay (2011) or Särkkä (2013).

The SMC has advantages over other alternative posterior computation methods such as Kalman filter, extended Kalman filter and Markov chain Monte Carlo (MCMC). The Kalman filter and the extended Kalman filter are applicable to the linear dynamic model, while the SMC can be applied to the nonlinear dynamic model as well. The SMC has advantages over MCMC. First, SMC methods are much faster than MCMC methods. Whenever the new parameter is propagated in each stage of SMC, we only calculate the next step of the numerical solution. Fast computation is the biggest advantage of our method. Second, they are able to be implemented in an on-line learning scenario. When a new data point is observed, SMC just need to update one step of the algorithm, while MCMC must implement the whole algorithm again to get the new posterior samples. Due to these advantages, we choose SMC for the posterior computation of the nonlinear dynamic model, which approximates the ODE model.

3 Posterior Computations for the Approximate Dynamic Model via Sequential Monte Carlo

To obtain inference for θ based on the approximated dynamic model of (5), we will use the extended Liu and West (ELW) filter to estimate parameters and states (Rios and Lopes, 2013). We call the proposed method of computation relaxed DEM with ELW filter (RDEM-ELW) or simply RDEM. The ELW filter uses the idea of auxiliary particle filter to sample the states, and it divides the parameters into two sets, θ and γ , representing parameters with and without sufficient statistic, respectively. The parameters denoted by θ (i.e., without the sufficient statistic) is the same set of parameters denoted by θ in (5). For the θ -set, the ELW filter introduces artificial random errors onto the static parameter θ , thus converting and combining it with the other evolving parameters which are the states x_i (see Liu and West, 2001). Furthermore, in the ELW filter, the marginal posterior of θ at each time point is approximated by a finite mixture of normal distributions. The mean and variance of the evolution distribution are determined so that the mixture of normals does not increase the posterior variance. For the posterior update of the γ -set of parameters, the idea of Storvik (2002) and Fearnhead (2002) is used. For the

idea of ELW to be successfully applied, the posterior of γ , $p(\gamma | y_{1:i}, x_{0:i}, \theta)$, $i = 1, \dots, n$, needs to be tractable, that is from which samples can be drawn directly. In particular, we assume $p(\gamma | y_{1:i}, x_{0:i}, \theta)$ depends on a sufficient statistic $s_i = s_i(y_{1:i}, x_{0:i}, \theta)$.

Incorporating the evolution of θ into (5) according to the ELW methodology creates a further relaxation of the former model. The ELW model for the approximate dynamical model in (5) is given by

$$y_i \sim N(x_i, \sigma^2 I_p), \quad (7)$$

$$x_i \sim N(g(x_{i-1}, t_i; \theta_i), u^2 I_p), \quad \text{and} \quad (8)$$

$$\theta_i \sim N(a\theta_{i-1} + (1-a)\bar{\theta}_{i-1}, \tilde{h}^2 V_i), \quad (9)$$

for $i = 1, 2, \dots, n$ with $\theta_0 \sim \pi_\theta$ and x_0 distributed according to its prior specification in (6). In (8), g is as defined in (3), and u is a small fixed positive real number representing the relaxation parameter. In (9), $\bar{\theta}_{i-1}$ represents the posterior mean of θ given $y_{1:i-1}$ at time $i-1$, $a = (1 - \tilde{h}^2)^{1/2}$ where $\tilde{h}^2 = 1 - ((3\delta - 1)/(2\delta))^2$, δ is a discounting factor usually taken to be a high value such as 0.95 or 0.99, and V_i is the covariance matrix corresponding to the evolution equation of θ_i . Equation (9) is the further relaxation and evolution model for θ prescribed by the ELW methodology (see Liu and West, 2001). The selection of the parameters a and \tilde{h} guarantees that the posterior variance of θ_i remains stable (i.e., does not increase) with the progression of the time index i .

Several posterior distributions will be needed for the subsequent discussion and we derive their forms here. Consider $\gamma = \lambda = \sigma^{-2}$, the inverse of the variance of observation error. ELW methodology requires the distribution $p(\gamma | y_{1:i}, x_{0:i}, \theta)$ be tractable and easily sampled from. In our case, the posterior distribution for γ , conditional on observations $y_{1:i}$, states $x_{0:i}$ and θ , is given by

$$\pi(\gamma | y_{1:i}, x_{0:i}, \theta) = \text{Gamma} \left(a_\lambda + \frac{(i+1)p}{2}, b_\lambda + \frac{1}{2} \left(\frac{\|x_0 - \mu_{x_0}\|^2}{c} + \sum_{k=1}^i \|y_k - x_k\|^2 \right) \right) \quad (10)$$

which is a tractable distribution. Note also from the above equation that the distribution of γ depends on $y_{1:i}$ and $x_{0:i}$ through the sufficient statistic $s_i = s_i(y_{1:i}, x_{0:i}, \theta) = (a_\lambda + (i+1)p/2, b_\lambda + (\|x_0 - \mu_{x_0}\|^2/c + \sum_{k=1}^i \|y_k - x_k\|^2)/2)$, where a_λ, b_λ, c and μ_{x_0} are all fixed and known hyperparameters (see (6)). Next, the two distributions, that is (i) the conditional distribution of x_i given x_{i-1}, y_i, θ_i and γ , and (ii) the marginal distribution of y_i given x_{i-1}, θ_i and γ , can

be obtained by considering the joint density of x_i and y_i , conditional on x_{i-1} , θ_i and γ , from (7) and (8). From these two equations, it follows that (x_i, y_i) is jointly normal, and thus, the conditional density of x_i given y_i is

$$p(x_i | x_{i-1}, y_i, \theta_i, \gamma) = N \left(\frac{y_i/\sigma^2 + g(x_{i-1}, t_i, \theta_i)/u^2}{1/\sigma^2 + 1/u^2}, \frac{1}{1/\sigma^2 + 1/u^2} I_p \right), \quad (11)$$

whereas the marginal distribution of y_i given x_{i-1} , θ_i and γ , obtained by integrating out x_i , is given by

$$p(y_i | x_{i-1}, \theta_i, \gamma) = N \left(g(x_{i-1}, t_i, \theta_i), (\sigma^2 + u^2) I_p \right). \quad (12)$$

We now give the ELW algorithm for obtaining inference for θ based on the approximate dynamic model (5) and the posteriors defined above. Let the notation $[A, B, \dots | C, D, \dots]$ denote the conditional density of random entities (either scalars or vectors) A, B, \dots conditional on either random or fixed constant entities C, D, \dots . The ELW model of (7)-(9) can be written based on this notation as

$$y_{i+1} \sim [y_{i+1} | x_{i+1}, \gamma], \quad (13)$$

$$x_{i+1} \sim [x_{i+1} | x_i, \theta_{i+1}], \quad \text{and} \quad (14)$$

$$\theta_{i+1} \sim [\theta_{i+1} | \theta_i, y_{1:i}]. \quad (15)$$

Equation (13)-(15) gives the joint distribution of $(y_{i+1}, x_{i+1}, \theta_{i+1})$ conditional on the observations, states and θ -values at previous time points, that is,

$$[y_{i+1}, x_{i+1}, \theta_{i+1} | x_i, \theta_i, y_{1:i}, \gamma] = [y_{i+1} | x_{i+1}, \gamma] \cdot [x_{i+1} | x_i, \theta_{i+1}] \cdot [\theta_{i+1} | \theta_i, y_{1:i}]$$

based on (13)-(15). The auxiliary particle filter (APF) technique rewrites this joint density as

$$[y_{i+1}, x_{i+1}, \theta_{i+1} | x_i, \theta_i, y_{1:i}, \gamma] = [x_{i+1} | x_i, \theta_{i+1}, y_{i+1}, \gamma] \cdot [y_{i+1} | x_i, \theta_{i+1}, \gamma] \cdot [\theta_{i+1} | \theta_i, y_{1:i}]. \quad (16)$$

The first term on the right hand side of (16) is given by (11), thus available in closed form for sampling in our examples. The second term on the right hand side of (16) is given by (12), which is again available in closed form for *evaluation* in our examples. The third term in (16) is the Liu and West filter for θ given by (15), which can be easily sampled from. We give our sampling methodology to sample from the posteriors using sequential Monte Carlo. Suppose $\{x_i^{(j)}, \theta_i^{(j)}, \gamma_i^{(j)}, s_i^{(j)}\}$ for $j = 1, 2, \dots, N$ are N samples from the posterior $[x_i, \theta_i, \gamma_i, s_i | y_{1:i}]$.

The subscript i on γ_i does not imply any evolution equation for γ . It just denotes the random variable γ for marginal realizations of γ from the posterior $[\gamma | s_i]$. Similarly, s_i denotes realizations of the sufficient statistic at time point i based on its functional equation, namely, $\mathcal{S}(y_{1:i}, x_{0:i}, \theta_i)$ when $x_{0:i}$ and θ_i are samples from the posterior $[x_{0:i}, \theta_i | y_{1:i}]$.

The steps of our sampling algorithm is as follows:

- First, sample $\theta_{i+1}^{(j)} \sim [\theta_{i+1} | \theta_i^{(j)}, y_{1:i}]$ according to (9) for $j = 1, 2, \dots, N$.
- Compute weights $w_i^{(j)} \propto [y_{i+1} | x_i^{(j)}, \theta_{i+1}^{(j)}, \gamma_i^{(j)}]$ for $j = 1, 2, \dots, N$.
- Obtain N resamples $\{\tilde{x}_i^{(j)}, \tilde{\theta}_{i+1}^{(j)}, \tilde{\gamma}_i^{(j)}, \tilde{s}_i^{(j)}\}_{j=1}^N$ by sampling from the collection $\{x_i^{(j)}, \theta_{i+1}^{(j)}, \gamma_i^{(j)}, s_i^{(j)}\}_{j=1}^N$ according to the weights $\{w_i^{(j)}\}_{j=1}^N$.
- Sample $\tilde{x}_{i+1}^{(j)} \sim [x_{i+1} | \tilde{x}_i^{(j)}, \tilde{\theta}_{i+1}^{(j)}, y_{i+1}, \tilde{\gamma}_i^{(j)}]$ for $j = 1, 2, \dots, N$.
- Compute $\tilde{s}_{i+1}^{(j)} = \mathcal{S}(\tilde{s}_i^{(j)}, y_{i+1}, \tilde{x}_{i+1}^{(j)}, \tilde{\theta}_{i+1}^{(j)})$ for $j = 1, 2, \dots, N$.
- Sample $\tilde{\gamma}_{i+1}^{(j)} \sim [\gamma | \tilde{s}_{i+1}^{(j)}]$ for $j = 1, 2, \dots, N$.

Then, it follows that the N samples $\{\tilde{x}_{i+1}^{(j)}, \tilde{\theta}_{i+1}^{(j)}, \tilde{\gamma}_{i+1}^{(j)}, \tilde{s}_{i+1}^{(j)}\}$ for $j = 1, 2, \dots, N$ are realizations from the posterior $[x_{i+1}, \theta_{i+1}, \gamma_{i+1}, s_{i+1} | y_{1:i+1}]$. As the tuning parameter $\tilde{h} \rightarrow 0$, the posterior of θ at every time point i from the approximate dynamic model becomes closer to the true posterior from the DEM.

As mentioned earlier, in the above algorithm, the subscripts i on γ_i and s_i do not imply any kind of evolution over time. They just represent the update of the parameter and statistic, respectively, as new data become available. The tuning parameter a determines the extent of shrinkage of the normal mixture through its mean. It also controls the smoothness through the variance term $\tilde{h}^2 V_i$. It is usually prescribed to be chosen around the value 0.95. The tuning parameter a was fixed at 0.95 throughout the rest of examples. This corresponds to taking $\tilde{h}^2 = 1 - a^2 = 0.0975$ and $\delta = 1/(3 - 2a) = 0.909$. For the covariance matrix V_i , we chose $V_i = (N - 1)^{-1} \sum_{j=1}^N (\theta_{i-1}^{(j)} - \bar{\theta}_{i-1})(\theta_{i-1}^{(j)} - \bar{\theta}_{i-1})^T$.

The initial proposal density $q(x_0, \theta, \gamma)$ affects the performance of the algorithm. The proposal density which is concentrated around the true parameter has a better performance than the other proposal densities even with relatively small number of particles. In practice, we suggest that one

run the ELW filter with initial particles $\theta^{(j)}$ and $\gamma^{(j)}$ from $\pi(\theta, \gamma)$ and rerun with the particles $\hat{\theta}^{(j)}$ and $\hat{\gamma}^{(j)}$ from the first inference. It is equivalent to consider the proposal density

$$q(x_0, \theta, \gamma) \equiv \pi(x_0) \times \pi(\theta, \gamma \mid \mathbf{y}_n).$$

We call the resulting particles the refined particles. It was used throughout the rest of examples.

4 Convergence of the Posterior

4.1 Convergence of the Posterior as the relaxation parameter decreases

In this subsection, we show that as the relaxation parameter u converges to 0, the posterior density of (x_0, θ, λ) from the approximate dynamic model converges to the posterior from the DEM, i.e.

$$\pi(x_0, \theta, \lambda \mid \mathbf{y}_n, u^2) = \frac{\int L(\Lambda) \pi(dx_1, \dots, dx_n \mid x_0, \theta, u^2) \pi(x_0, \theta, \lambda)}{\int \int L(\Lambda) \pi(dx_1, \dots, dx_n \mid x_0, \theta, u^2) \pi(dx_0, d\theta, d\lambda)} \quad (17)$$

converges to

$$\pi(x_0, \theta, \lambda \mid \mathbf{y}_n) = \frac{L^*(x_0, \theta, \lambda) \pi(x_0, \theta, \lambda)}{\int L^*(x_0, \theta, \lambda) \pi(dx_0, d\theta, d\lambda)} \quad (18)$$

as $u^2 \rightarrow 0$, where $\Lambda = (x_0, \dots, x_n, \theta, \lambda)$,

$$\begin{aligned} L(\Lambda) &= (\lambda)^{np/2} \exp\left(-\frac{\lambda}{2} \cdot \sum_{i=1}^n \|y_i - x_i\|^2\right) \text{ and} \\ L^*(x_0, \theta, \lambda) &= (\lambda)^{np/2} \exp\left(-\frac{\lambda}{2} \cdot \sum_{i=1}^n \|y_i - g^i(x_0, t_{i-1}; \theta)\|^2\right) \end{aligned}$$

with $g^i(x_0, t_{i-1}; \theta) = g(g^{i-1}(x_0, t_{i-2}; \theta), t_{i-1}; \theta)$. Note that $\pi(x_0, \theta, \lambda \mid \mathbf{y}_n)$ is the posterior of DEM.

Theorem 4.1 *Consider model (5) and prior (6). Suppose $f(x, t; \theta)$ is continuous in x . Then, the posterior density of the dynamic model (5) converges to that of the differential equation model (4), i.e.*

$$\pi(x_0, \theta, \lambda \mid \mathbf{y}_n, u^2) \rightarrow \pi(x_0, \theta, \lambda \mid \mathbf{y}_n)$$

for all x_0, θ, λ as $u^2 \rightarrow 0$.

4.2 Convergence of the Posterior as the step size increases

We have shown that the posterior of the dynamic model (5) converges to that of the differential equation model (4) as $u^2 \rightarrow 0$. In this subsection, we will prove that the posterior of the differential equation model converges to that of the true model.

If the step size is m , each time interval $[t_{i-1}, t_i]$ is divided into m segments of length $(t_i - t_{i-1})/m$, and the Runge-Kutta method is applied to each subinterval to obtain x'_i s. To clarify the difference, let x^m be the approximated solution of the differential equation by the fourth-order Runge-Kutta method with m segments. Similarly, let π_m and π_{true} be the posterior distributions corresponding to x^m and the true x , respectively. Note $x^m(t_1) = x(t_1)$ for all m .

Theorem 4.2 *Consider model (4) and prior (6). Suppose $f(x, t; \theta)$ satisfies Lipschitz condition in x , i.e. there exists the constant $K > 0$ such that*

$$\|f(x, t; \theta) - f(x', t; \theta)\| < K\|x - x'\| \quad (19)$$

for any $x, x' \in \mathbb{R}^p, t \in [T_0, T_1]$ and $\theta \in \Theta$. Then, the posterior density of the differential equation model (4) converges that of the true model, i.e.

$$\pi_m(x_0, \theta, \lambda | \mathbf{y}_n) \rightarrow \pi_{true}(x_0, \theta, \lambda | \mathbf{y}_n)$$

for all x_0, θ, λ as $m \rightarrow \infty$.

This result guarantees that the differential equation model works well with a reasonable segments parameter m under the Lipschitz condition.

4.3 Choice of the relaxation parameter and the step size

In practice, the choice of u^2 and m can affect the performance of the approximation. The approximate posterior distribution may vary by different choice of these values. Theoretically, the smaller the relaxation parameter u^2 is, the closer the approximate posterior is to the true posterior. But in practice we may need moderately large value of u^2 to get stable posterior approximation. We suggest following strategy for choosing the variance of state u^2 . Consider various u^2 values from large to small values in turn. For each u^2 value, check the stability of posteriors by running two or three ELW filters simultaneously. Here, the stability means that all

posterior densities based on ELW runs are closed enough to each other. Finally, use the smallest u^2 value for the inference which gives the stable result.

For convenience, let $h \equiv t_{i+1} - t_i$ for all $i = 1, 2, \dots, n - 1$. For the choice of m , we assume $h/m = O(n^{-\alpha})$. Theoretically, the larger value of m gives more accurate inference, but it would require heavier computation. In the following theorem, we relate the step size h/m to the approximation error rate of the posterior, and based on the theorem we suggest values of m for computation according to the acceptable error rate. The theorem requires the following assumptions.

- A1. $\{x(t) : t \in [0, T]\}$ is a compact subset of \mathbb{R}^p ;
- A2. $\{y(t) : t \in [0, T]\}$ is a bounded subset of \mathbb{R}^p ; and
- A3. the K th order derivative of $f(x, t; \theta)$ with respect to t exists and is continuous in x and t , where K is the order of the numerical method g .

Theorem 4.3 *Consider model (4) and prior (6). Suppose $f(x, t; \theta)$ satisfies Lipschitz condition (19) in x , and suppose A1 – A3 hold. Let K be the order of the numerical method g and $h/m = O(n^{-\alpha})$. If $\alpha \geq (1+R)/K$, the error rate of the posterior approximation is $O(n^{-R})$ for sufficiently large n , i.e.,*

$$\pi_m(x_0, \theta, \lambda | \mathbf{y}_n) = \pi(x_0, \theta, \lambda | \mathbf{y}_n) \times (1 + O(n^{-R}))$$

for all x_0, θ, λ , then $\alpha \geq (1 + R)/K$ is sufficient.

Note that the order of Runge-Kutta method is 4, and the rate of h is n^{-1} because we consider a bounded time interval $[0, T] \subset \mathbb{R}$ with $T < \infty$. By the above theorem, if we want to get the error rate $O(n^{-3})$ or larger, we know that it can be achieved by $m = 1$ for large n . However, in practice, one should notice that the additional error from the SMC sampling may arise. In such case, we may need to use m bigger than 1.

5 Simulated Data Examples

5.1 Newton's law of Cooling

5.1.1 Description of model and data generation step

Newton's law of cooling, made by English physicist Isaac Newton, is a model describing the temperature change of an object. According to the model, the temperature of an object changes proportional to the temperature difference between the object and its surroundings. This notion is given by the following ODE form

$$\dot{x}(t) = \theta_1(x(t) - \theta_2), \quad (20)$$

where $x(t)$ is the temperature of the object at time t , θ_1 is a negative constant and θ_2 is the temperature of the surroundings. All of the temperature are in Celcius. For more details, see Incropera (2006).

We chose this model as a testbed for our method. Since the solution of (20) is known as

$$x(t) = \theta_2 - (\theta_2 - x_0)e^{\theta_1 t} \quad (21)$$

where $x_0 = x(0)$, we can calculate the true posterior directly. The data $y_i = y(t_i)$ was generated with the true mean function (21) and we set the model parameters as $x_0 = 20$, $\theta = (-0.5, 80)^T$, $\sigma^2 = 25$ and time points $t_i = ih$ for $i = 1, \dots, n$ where the sample size $n = 100$ and the step size $h = 0.15$. The simulated data and the true mean function are shown in Figure 1.

The priors were set by

$$\begin{aligned} x_0 \mid \lambda &\sim N(\mu_{x_0}, c/\lambda) \\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda) \\ \theta = (\theta_1, \theta_2) &\sim \text{Uniform}((-100, 0) \times (50, 150)) \end{aligned}$$

where $\mu_{x_0} = y_1$, $a_\lambda = 1$, $b_\lambda = 1$ and $c = 1$. The values of y_i are in the interval $[65, 90]$ after 50th observation, and the temperature of the surroundings, θ_2 , must be the around the interval. The prior of θ_2 is set by $\text{Uniform}(50, 150)$ whose support includes $[65, 90]$. With a similar reasoning, we set $\theta_1 \sim \text{Uniform}(-100, 0)$.

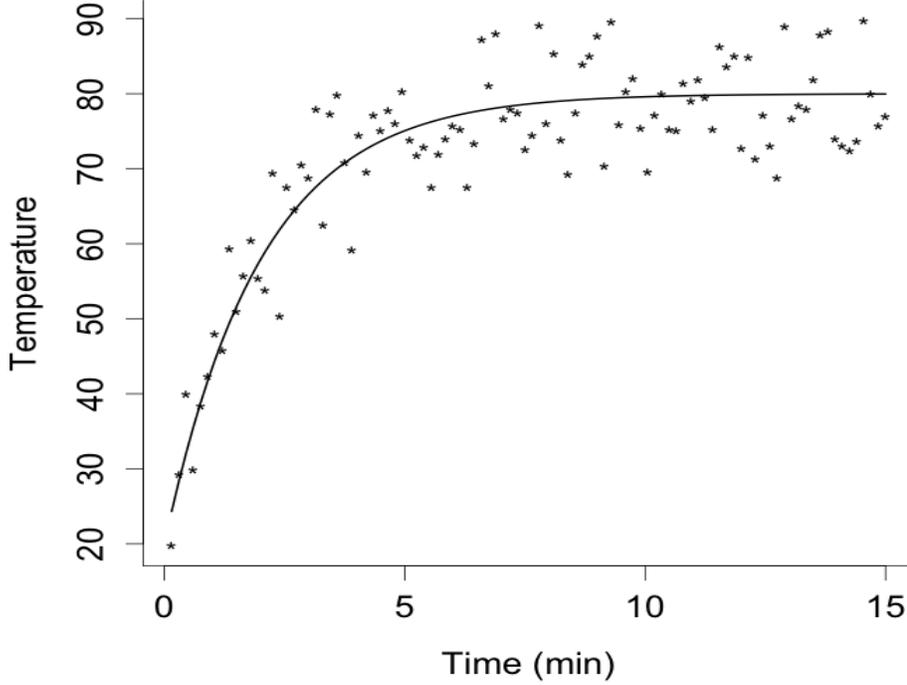


Figure 1: The solid line is the true temperature as a function of time from the Newton's law of cooling model with $x_0 = 20$, $\theta = (-0.5, 80)^T$. The star-shaped points are the generated data of temperatures with $\sigma^2 = 25$.

The true posterior of θ and λ can be obtained as follows:

$$\begin{aligned} \lambda \mid \theta, y_{1:n} &\sim \text{Gamma}\left(\frac{np}{2} + a_\lambda, \frac{1}{2}\tilde{u}(\theta) + b_\lambda\right) \\ \theta \mid y_{1:n} &\sim \frac{1}{\left(\frac{1}{2}\tilde{u}(\theta) + b_\lambda\right)^{\frac{np}{2} + a_\lambda}} I(-100 < \theta_1 < 0) I(50 < \theta_2 < 150), \end{aligned} \quad (22)$$

where

$$\begin{aligned} \tilde{u}(\theta) &= \mu_{x_0}^2/c + \sum_{i=1}^n z_i^2 - (1/c + \sum_{i=1}^n e^{2\theta_1 i h})^{-1} (\mu_{x_0}/c + \sum_{i=1}^n z_i e^{\theta_1 i h})^2, \\ z_i &= z_i(\theta) = y_i - \theta_2 + \theta_2 e^{\theta_1 i h}. \end{aligned}$$

5.1.2 Assessment of the convergence of the posteriors

We assessed the convergence of posteriors which is described at Theorem 4.1. To show that the posterior of dynamic model converges to that of DEM, we got the simulation results for

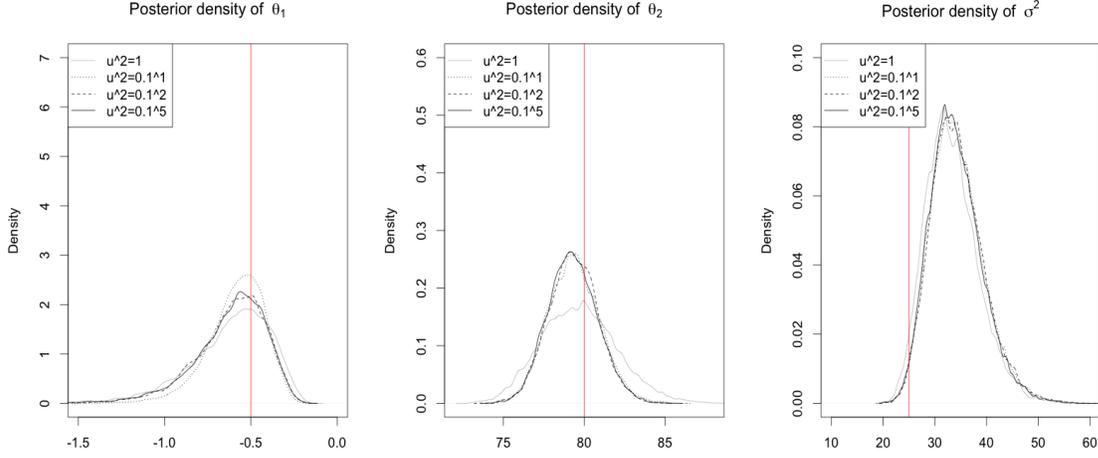


Figure 2: The histograms of the marginal posterior distributions of the dynamic models with $u^2 = 1, 0.1, 0.1^2, 0.1^5$ and $m = 1$ from the Newton's law of cooling. The red lines are the true values of parameters, $(\theta_1, \theta_2, \sigma^2) = (-0.5, 80, 25)$.

RDEM with $u^2 = 1, 0.1^1, 0.1^2$ and 0.1^5 . The DEM was treated as a dynamic model with small value of u^2 . We ran the ELW filter based on 20,000 particles and fixed the number of segments m at 1. For all of the settings, the ELW filter takes less than 3 seconds for 20,000 particles. The histogram of the marginal posterior distributions are drawn at Figure 2. It seems that the posterior of dynamic model approaches that of the DEM as u^2 decreases to zero. Thus, it supports the theoretical result, Theorem 4.1.

To show that the posterior of DEM converges to that of true model, we got the simulation results for the DEM with the number of segments $m = 1, 2, 4$ and the true model. We approximated DEM by the dynamic model with $u^2 = 0.1^5$. For the true model, we used a grid sampling algorithm for the true posterior (22). For each setting, the ELW filter takes less than 3 seconds for 20,000 particles. The grid set was chosen by $[-2, 0] \times [70, 90]$, and each axis was divided into 50 equal length intervals resulting 51 points. 20,000 posterior samples were drawn. The histograms of the marginal posterior distributions are drawn at Figure 3. The posterior densities of DEM are quite similar to each other, but they have the larger variation than the true posterior densities.

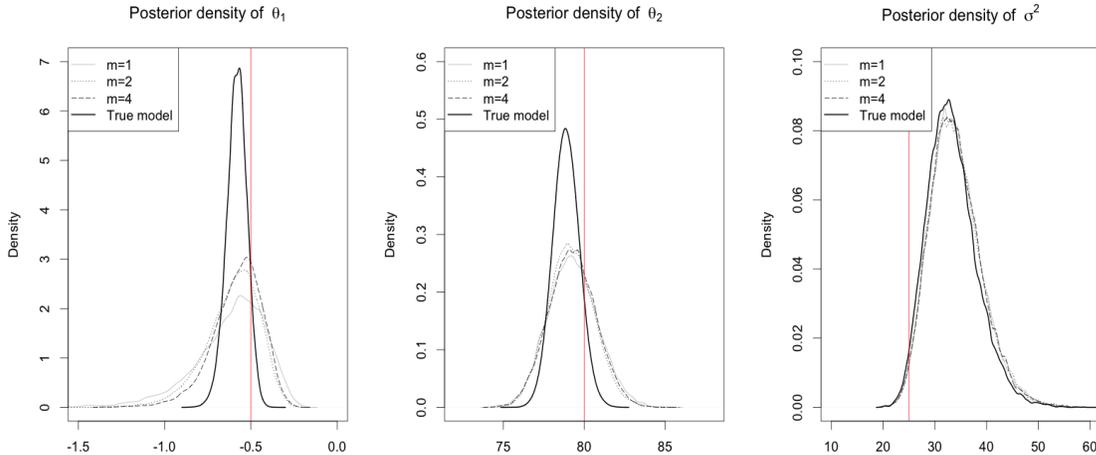


Figure 3: The histograms of the marginal posterior distributions of the dynamic models with $u^2 = 0.1^5$ and $m = 1, 2, 4$, and those of the true model from the Newton's law of cooling. The red lines are the true values of parameters, $(\theta_1, \theta_2, \sigma^2) = (-0.5, 80, 25)$.

5.2 FitzHugh-Nagumo model

5.2.1 Description of model and data generation step

FitzHugh-Nagumo model (FitzHugh, 1961; Nagumo et al. 1962) describes the action of spike potential in the giant axon of squid neurons by an ODE with two state variables and three parameters:

$$\begin{aligned}\dot{x}_1(t) &= \theta_3 \left(x_1(t) - \frac{1}{3}x_1^3(t) + x_2(t) \right), \\ \dot{x}_2(t) &= -\frac{1}{\theta_3} (x_1(t) - \theta_1 + \theta_2 x_2(t)),\end{aligned}$$

where $-0.8 < \theta_1, \theta_2 < 0.8$ and $0 < \theta_3 < 8$. The two state variables, $x_1(t)$ and $x_2(t)$, are the voltage across an membrane and outward currents at time t , respectively.

Using the FitzHugh-Nagumo model, we compare the proposed method with the parameter cascading method (Ramsay et al., 2007), the delayed rejection adaptive Metropolis (DRAM) algorithm (Soetaert and Petzoldt, 2010) and the Laplace approximated posterior (LAP) method (Dass et al., 2017). The data $y_i = y(t_i)$ was generated from DEM (4) with the model parameters $x_0 = (-1, 1)^T$, $\theta = (0.2, 0.2, 3)^T$, $\sigma^2 = 25$ and time points $t_i = ih$ for $i = 1, \dots, n$, where the sample size $n = 100$ and the step size $h = 0.2$, $m = 400$. The simulated data and the true mean

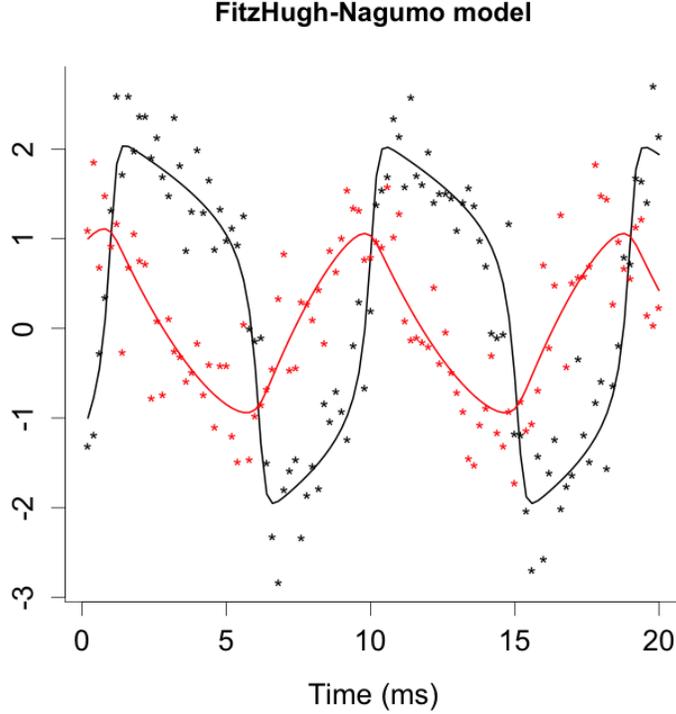


Figure 4: The solid lines are x_1 (black line) and x_2 (red line) as a function of time from the FitzHugh-Nagumo model with $x(t_0) = (-1, 1)^T$, $\theta = (0.2, 0.2, 3)^T$. The star-shaped points are the generated data of the populations with $\sigma^2 = 0.25$.

function are shown in Figure 4. The priors were set by

$$\begin{aligned} x_0 \mid \lambda &\sim N(\mu_{x_0}, c\lambda^{-1}I_2) \\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda) \\ \theta &\sim \text{Uniform}(A) \end{aligned}$$

where $\mu_{x_0} = y_1$, $a_\lambda = 1$, $b_\lambda = 1$, $c = 1$ and $A = \{(\theta_1, \theta_2, \theta_3) : -0.8 < \theta_1, \theta_2 < 0.8, 0 < \theta_3 < 8\}$.

5.2.2 Comparison with other methods

To compare the proposed method (RDEM-ELW) with other methods, the parameter cascading (PC) method, DRAM algorithm and LAP method were applied to the same data set. We used the R packages `CollocInfer` and `FME` for the parameter cascading and DRAM, respectively.

The PC method is one of the popular frequentist methods for estimating the parameters in

ODE. It uses the collocation method which represents the state vector $x(t)$ as a series of basis expansion. The penalized likelihood criterion has three components: the matrix of coefficients of basis expansions C , the unknown parameter θ and the smoothing parameter λ . PC optimizes the penalized likelihood by two steps. In the inner optimization, the criterion is optimized with respect to the coefficient C while θ and λ are fixed. After that, in the outer optimization, the penalized likelihood is optimized with respect to θ while λ is kept fixed. The smoothing parameter λ is chosen based on the appropriate criteria such as the numerical stability of parameter estimates or the forward prediction error (Hooker et al., 2000). For more details about PC method, see Ramsay et al. (2007). For the PC method, we used the third-order B-spline basis and $2n - 1$ equally spaced knots on $[t_0, t_n]$. The smoothing parameter was set by $\lambda = 10^5$. The initial parameter were drawn from $N(\theta_0, (0.01)^2 I_q)$ where θ_0 is the true parameter value.

The DRAM algorithm, a variant of the standard Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), is chosen as a benchmark in the Bayesian side. With the R package FME (Soetaert and Petzoldt, 2010), one can infer the DEM with DRAM algorithm for the parameters and numerical integration for the state variables. We applied the DRAM algorithm with the initial parameter as the maximum likelihood estimate using `modFit()` function and the maximal number of tries 1. The parameter covariance was updated in every 100 iteration. We got 20,000 posterior samples for the inference.

LAP method is another benchmark in the Bayesian side. It is fast when the dimension of parameter is small and empirically has comparable or better performance than PC method and DRAM algorithm (Dass et al., 2017). Since the dimension of parameter is small, the grid sampling method for θ was chosen. For each parameter θ_i , the grid range was chosen by $[\hat{\theta}_i^R \pm 4sd(\hat{\theta}_i^R)]$ where $\hat{\theta}_i^R$ is the parameter estimate for θ_i from the PC method. Each axis was divided into 31 intervals of equal length, and the step size for numerical integration was set at $m = 2$. The priors for parameters were set as in subsection 5.2.1, and 20,000 posterior samples were obtained.

For the RDEM-ELW, the step size for numerical integration and the variance for the state were chosen by $m = 2$ and $u^2 = 0.1^5$, respectively. The priors for parameters were set as described in subsection 5.2.1, the number of particles was chosen by $N = 20,000$. We generated 100 simulated data set using the 4th order Runge-Kutta. The model parameters were set as described in subsection 5.2.1.

For RDEM, PC and DRAM methods, R and C/C++ were used for implementation. R and

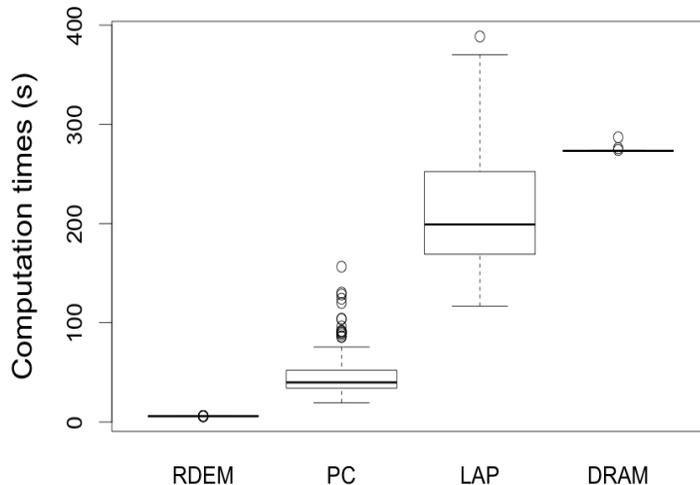


Figure 5: The boxplots of the computation times for $\hat{\theta}$ based on 100 simulated date sets. The results for the relaxed DEM with ELW filter (RDEM), the parameter cascading (PC)method, Laplace approximated procedure (LAP) and delayed rejection adaptive Metropolis (DRAM) algorithm are shown.

Fortran90 were used for LAP method. On average based on 100 simulations, it took only 3.523 seconds for estimation, while the PC method, DRAM algorithm and LAP method took 49.152, 276.700 and 215.591 seconds, respectively. The boxplot of computation times for each method is given at Figure 5. The proposed RDEM method significantly reduced the computation time. It was even faster than the frequentist method, the PC method. Thus, the RDEM method has an enormous advantage in computation speed over other methods. Table 1 represents the absolute biases, standard deviations for $\hat{\theta}$ and root mean squared errors (rmse) for $\hat{\theta}$ in the FitzHugh-Nagumo model. It seems RDEM method provides reasonable estimates in terms of bias, but larger standard deviation than others.

6 Lynx-hare data: Lotka-Volterra equation

There are large number of models to express predator-prey relationships because predation is often direct, conspicuous and easy to study. Lotka-Volterra model is one of the simplest model of predator-pray interactions. Lotka (1925) and Volterra (1926) independently developed the

Table 1: The table of mean of the absolute biases, standard deviations and root mean squared errors (rmse) for $\hat{\theta}$ in the FitzHugh-Nagumo model. The results for the relaxed DEM with ELW filter (RDEM), parameter cascading (PC) method, Laplace approximated procedure (LAP) and delayed rejection adaptive Metropolis (DRAM) algorithm are shown.

		RDEM	PC	LAP	DRAM
Absolute bias	θ_1	0.051	0.024	0.024	0.024
	θ_2	0.135	0.106	0.099	0.100
	θ_3	0.108	0.039	0.044	0.047
Standard deviation	θ_1	0.063	0.027	0.027	0.028
	θ_2	0.130	0.123	0.117	0.119
	θ_3	0.194	0.060	0.056	0.059
rmse	θ_1	0.084	0.038	0.038	0.040
	θ_2	0.198	0.171	0.161	0.164
	θ_3	0.233	0.076	0.075	0.079

model of the form:

$$\begin{aligned}\dot{x}_1(t) &= x_1(t)(\theta_1 - \theta_2 x_2(t)), \\ \dot{x}_2(t) &= -x_2(t)(\theta_3 - \theta_4 x_1(t)),\end{aligned}\tag{23}$$

where x_1 denotes the number of preys, and x_2 denotes the number of their predators. The model parameters $\theta_1, \theta_2, \theta_3$ and θ_4 are the intrinsic rate of prey population increase, the predation rate, the predator mortality rate and the offspring rate of the predator, respectively.

Lynx-hare data is a popular data set representing the number of captured lynx and snowshoe hares in North Canada which was collected by Hudson Bay company. It contains the number of furs of lynx and hares, so it implies the actual populations of them. We obtained the annual data between 1900 and 1920 recorded in thousands from Li (2012) which is given at Figure 6. The Lotka-Volterra equation, the equation (23), is fitted to the data set and used to predict the future values of trapped lynxes and hares.

The same model and prior in subsection 5.2 were used. As we mentioned in subsection 4.3, we ran the ELW filter 10 times based on $N = 500,000$ particles with $u^2 = 20, 10, 5, 1$ and 0.1^5 , in turn. In this case, u^2 values smaller than 5 lead somewhat unstable approximation even with 3,000,000 particles. Finally, the state variance was chosen by $u^2 = 5$ based on the criterion in subsection 4.3, because it gives stable posterior densities for each ELW run. The other model parameters were chosen as the subsection 5.2. On average, it took approximately 17 seconds for each run.

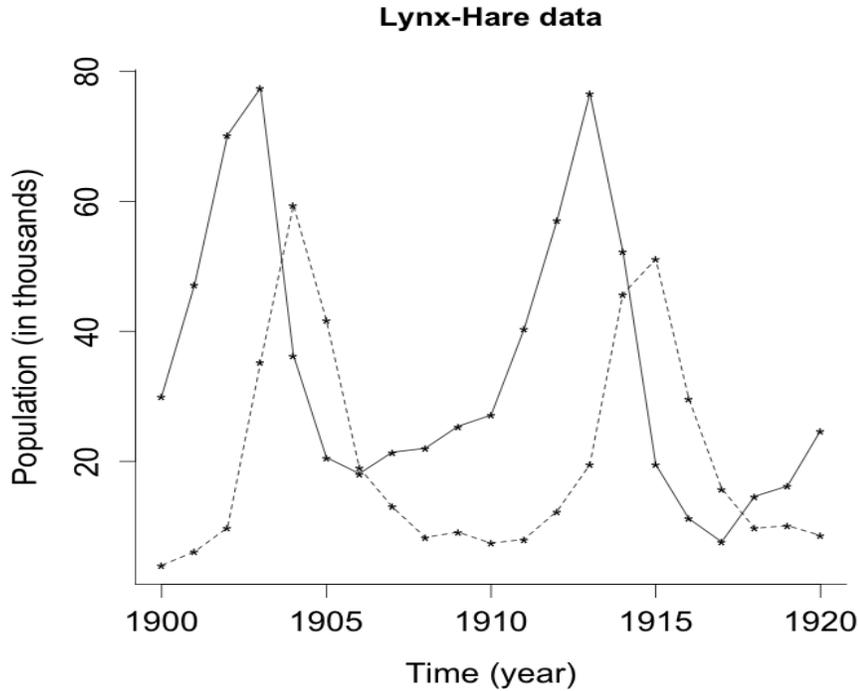


Figure 6: The numbers of trapped lynx and snowshoe hares between 1900 and 1920 is drawn. The solid line is the number of hares, and the dotted line is the number of lynx.

Table 2: Posterior summary statistics for the parameter of the Lotka-Volterra equation for the lynx-hare data with $m = 2$ and $u^2 = 10$.

	Mean	Median	90% credible interval
θ_1	0.526	0.525	(0.491, 0.562)
θ_2	0.026	0.026	(0.024, 0.027)
θ_3	0.986	0.985	(0.906, 1.067)
θ_4	0.028	0.028	(0.026, 0.030)
σ^2	4.087	3.818	(2.018, 7.065)

The marginal posterior densities of parameters are given at Figure 7. Posterior summary statistics for the first run are represented at Table 2. Figure 8 contains the scatter plots of the observations and 90% posterior credible lines for prediction values at 10 future time points when $m = 2$ and $u^2 = 5$. The predicted values of trapped lynxes and hares follow oscillation patterns. The size of prediction interval gets wider as the prediction time gets further ahead and also the predicted value become larger.

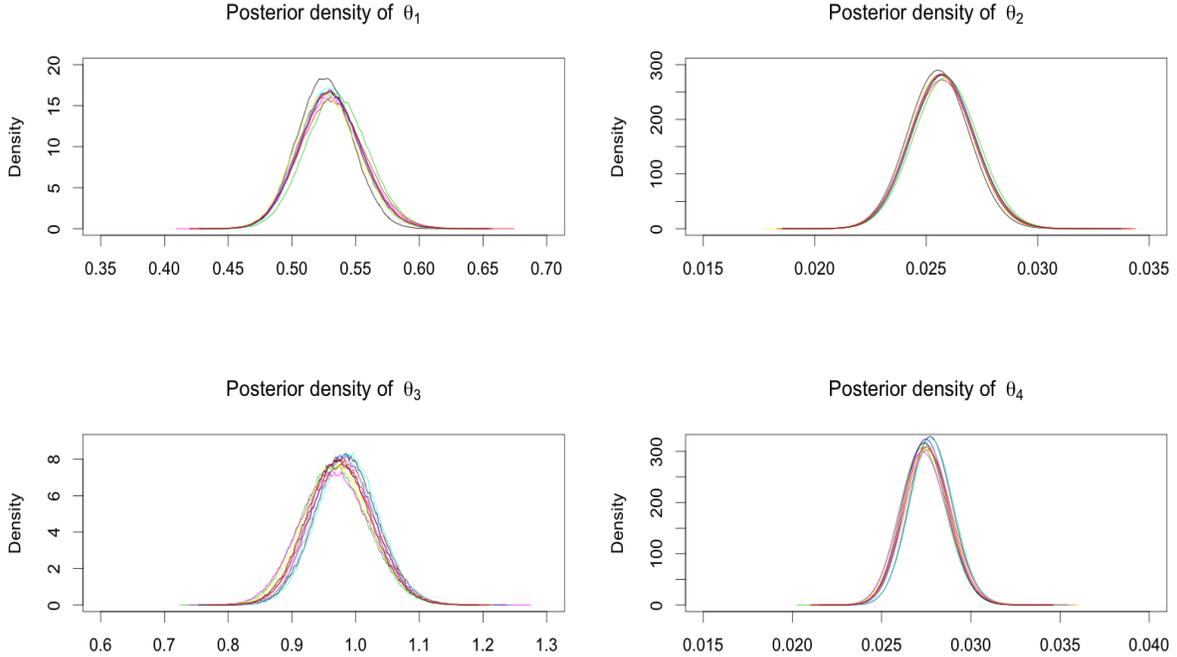


Figure 7: The posterior densities of the Lotka-Volterra equation for the lynx-hare data based on 10 ELW filter runs with $m = 2$ and $u^2 = 5$.

7 Discussion

A lot of biological or physical systems are given by a set of differential equations. To understand these processes, estimation of their parameters is essential. However, especially in Bayesian literature, there is no standard framework to analyze differential equation model. In many cases, the posterior of parameter does not belong a well-known family, so grid sampling or MCMC methods are used to get posterior samples. They usually suffer from heavy computation. We propose a general framework to analyze DEM using relaxation via dynamical systems. The dynamic model enables a fast inference for DEM and provides convenient sampling methods. Among the sampling algorithms for dynamic models, we adopted the ELW filter suggested by Rios and Lopes (2013). We argue that our method can be an alternative to the existing inference methods when one needs a fast and reasonable result. This argument is supported by the example in subsection 5.2. Section 4 guarantees the convergence of the approximated posterior to the true posterior. However, the theoretical results in this paper does not consider the additional error from the SMC sampling. The proposed method may be improved if a better SMC algorithm is

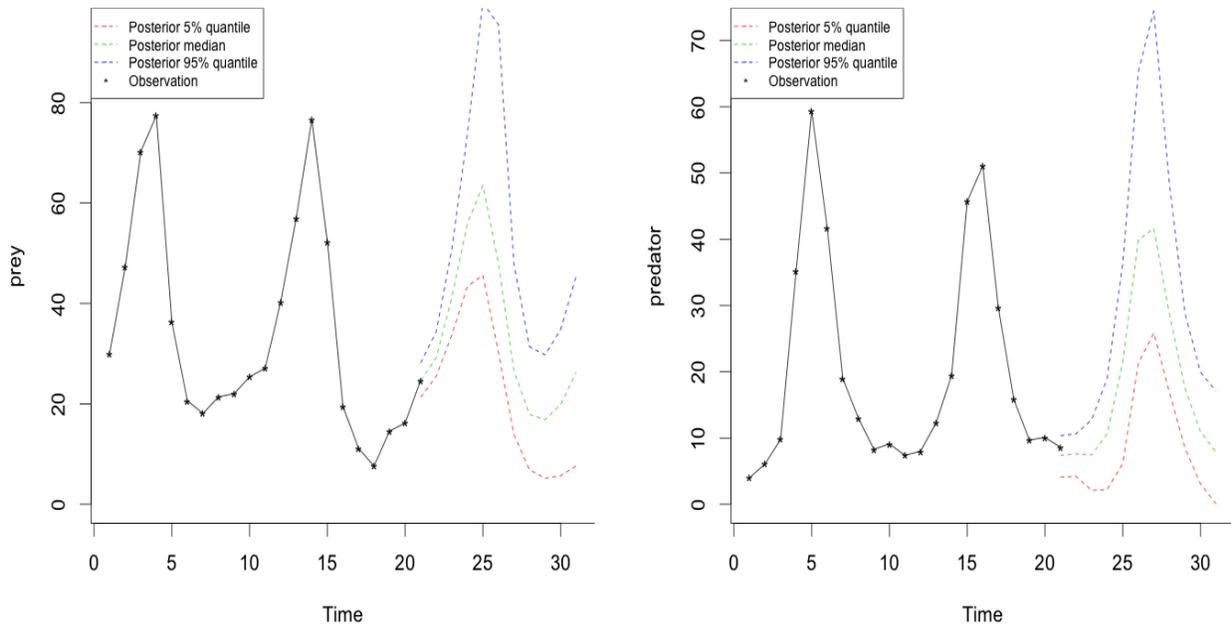


Figure 8: Scatter plot of the lynx-hare data and plots of 90% credible set lines for predictions of 10 time points ahead are drawn when $m = 2$ and $u^2 = 5$. The upper, lower and middle dotted lines are the 95% and 5% quantiles and median of the posterior, respectively. The star-shaped points are the lynx-hare data.

developed.

Appendix

The following lemma shows that each x_i given x_{i-1}, θ, u^2 converges to $g(x_{i-1}, t_{i-1}; \theta)$ in probability as $u^2 \rightarrow 0$.

Lemma 7.1 *Consider model (5). Then, for $i = 1, \dots, n$, x_i given x_{i-1}, θ and u^2 converges to $g(x_{i-1}, t_{i-1}; \theta)$ in probability as $u^2 \rightarrow 0$.*

Proof of Lemma 7.1 Note that $r^T x_i | x_{i-1}, \theta, u^2 \sim N(r^T g(x_{i-1}, t_{i-1}; \theta), u^2 \|r\|^2)$ for all $r \in \mathbb{R}^p, i = 1, \dots, n$. If we denote $\phi_{[Z]}$ as a moment generating function (mgf) of random variable Z ,

then for any $r \in \mathbb{R}^p$,

$$\begin{aligned}\phi_{[r^T x_i | x_{i-1}, \theta, u^2]}(z) &= \exp(r^T g(x_{i-1}, t_{i-1}; \theta)z + \frac{1}{2}u^2 \|r\|^2 z^2) \\ &\rightarrow \exp(r^T g(x_{i-1}, t_{i-1}; \theta)z)\end{aligned}\tag{24}$$

as $u^2 \rightarrow 0$, for $i = 1, \dots, n$. Note that (24) is mgf of $[r^T g(x_{i-1}, t_{i-1}; \theta) | x_{i-1}, \theta]$. Since the convergence of mgf implies the convergence of distribution, it implies

$$[r^T x_i | x_{i-1}, \theta, u^2] \rightarrow [r^T g(x_{i-1}, t_{i-1}; \theta) | x_{i-1}, \theta]$$

for any $r \in \mathbb{R}^p$. Hence, by the Cramer-Wold theorem (Billingsley, 1995), it implies that $[x_i | x_{i-1}, \theta]$ converges to $g(x_{i-1}, t_{i-1}; \theta)$ in distribution, as $u^2 \rightarrow 0$. Note that given x_{i-1} and θ , $g(x_{i-1}, t_{i-1}; \theta)$ is a constant. Thus, by Portmanteau theorem (Dudley, 2002), it implies the convergence in probability. \square

With the continuity condition of $f(x, t; \theta)$ in x , Lemma 7.1 can be extended to the joint convergence in probability using the mathematical induction. Lemma 7.2 describes the result.

Lemma 7.2 *Consider model (5). Suppose $f(x, t; \theta)$ is continuous in x . Then, $[x_1, \dots, x_n | x_0, \theta, u^2]$ converges to $(g(x_0, t_0; \theta), \dots, g^n(x_0, t_{n-1}; \theta))$ in probability as $u^2 \rightarrow 0$.*

Proof of Lemma 7.2 Let $X = (x_1, \dots, x_n)$ and $\bar{X} = (g(x_0, t_0; \theta), \dots, g^n(x_0, t_{n-1}; \theta))$ where

$$x_i^m = g^i(x_0, t_{i-1}; \theta), \quad i = 1, \dots, n\tag{25}$$

by the relation (3) where $g^i(x_0, t_i; \theta) = g(g^{i-1}(x_0, t_{i-1}; \theta), t_i; \theta)$ is defined recursively. We want to show

$$\lim_{u^2 \rightarrow 0} P\left(\|X - \bar{X}\| \geq \epsilon | x_0, \theta, u^2\right) = 0$$

for given $\epsilon > 0$. It suffices to prove

$$\lim_{u^2 \rightarrow 0} P\left(\|x_i - g^i(x_0, t_{i-1}; \theta)\| \geq \frac{\epsilon}{n} | x_0, \theta, u^2\right) = 0\tag{26}$$

for given $\epsilon > 0$ and $i = 1, \dots, n$. We use the mathematical induction.

When $i = 1$, we can check

$$\lim_{u^2 \rightarrow 0} P\left(\|x_1 - g(x_0, t_0; \theta)\| \geq \frac{\epsilon}{n} | x_0, \theta, u^2\right) = 0$$

by Lemma 7.1. Suppose (26) holds for $i = k$. Note

$$\begin{aligned} & P(\|x_{k+1} - g^{k+1}(x_0, t_k; \theta)\| \geq \frac{\epsilon}{n} | x_0, \theta, u^2) \\ & \leq P(\|x_{k+1} - g(x_k, t_k; \theta)\| \geq \frac{\epsilon}{2n} | x_0, \theta, u^2) \end{aligned} \quad (27)$$

$$+ P(\|g(x_k, t_k; \theta) - g(g^k(x_0, t_{k-1}; \theta), t_k; \theta)\| \geq \frac{\epsilon}{2n} | x_0, \theta, u^2). \quad (28)$$

By assumption, $g(x, t|\theta)$ is continuous in x . Thus, (28) converges to 0 as $u^2 \rightarrow 0$ because (26) holds for $i = k$. Also note that (27) is

$$E_{x_2|x_0, \theta, u^2} \dots E_{x_k|x_{k-1}, \theta, u^2} \left[P(\|x_{k+1} - g(x_k, t_k; \theta)\| \geq \frac{\epsilon}{2n} | x_k, \theta, u^2) \right].$$

Since $P(\|x_{k+1} - g(x_k, t_k; \theta)\| \geq \epsilon/(2n) | x_k, \theta, u^2) \leq 1$ and Lemma 7.1, (27) converges to 0 as $u^2 \rightarrow 0$ by the bounded convergence theorem. \square

Proof of Theorem 4.1 Note that we need to prove

$$\int L(\Lambda) \pi(dx_1, \dots, dx_n | x_0, \theta, u^2) \pi(x_0, \theta, \lambda) \rightarrow L^*(x_0, \theta, \lambda) \pi(x_0, \theta, \lambda), \quad (29)$$

$$\int \int L(\Lambda) \pi(dx_1, \dots, dx_n | x_0, \theta, u^2) \pi(dx_0, d\theta, d\lambda) \rightarrow \int L^*(x_0, \theta, \lambda) \pi(dx_0, d\theta, d\lambda) \quad (30)$$

as $u^2 \rightarrow 0$ where $\Lambda = (x_1, \dots, x_n, \theta, \lambda)$.

To show (29), we only need to prove

$$\int L(\Lambda) \pi(dx_1, \dots, dx_n | x_0, \theta, u^2) \rightarrow L^*(x_0, \theta, \lambda)$$

as $u^2 \rightarrow 0$. Since $L(\Lambda) = \lambda^{np/2} \exp(-\frac{\lambda}{2} \sum_{i=1}^n \|y_i - x_i\|^2)$, it suffices to prove

$$\int e^{-\frac{\lambda}{2} \sum_{i=1}^n \|y_i - x_i\|^2} \pi(dx_1, \dots, dx_n | x_0, \theta, u^2) \rightarrow e^{-\frac{\lambda}{2} \sum_{i=1}^n \|y_i - g^{i-1}(x_0, t_{i-1}; \theta)\|^2}. \quad (31)$$

By Lemma 7.2, we have

$$[x_1, \dots, x_n | x_0, \theta, u^2] \rightarrow [g(x_0, t_1; \theta), \dots, g^{n-1}(x_0, t_{n-1}; \theta) | x_0, \theta]$$

as $u^2 \rightarrow 0$. Note that the right hand side of (31) is the expectation of $\exp(-\lambda/2 \cdot \sum_{i=1}^n \|y_i - x_i\|^2)$ with respect to $[g(x_0, t_1; \theta), \dots, g^{n-1}(x_0, t_{n-1}; \theta) | x_0, \theta]$. Also note that $\exp(-\lambda/2 \cdot \sum_{i=1}^n \|y_i - x_i\|^2)$ is bounded by 1 and is continuous in x_1, \dots, x_n . Thus, the Portmanteau theorem implies (29).

Since we have proved (29), it suffices for (30) to show that $\int L(\Lambda) \pi(dx_2, \dots, dx_n | x_0, \theta, u^2)$ is dominated by an integrable random variable. It is easy to check because

$$\int L(\Lambda) \pi(dx_2, \dots, dx_n | x_0, \theta, u^2) \leq (\lambda)^{\frac{np}{2}}$$

and $(\lambda)^{np/2}$ is integrable with respect to $\pi(x_0, \theta, \lambda)$. The dominated convergence theorem gives the desired result. ■

Proof of Theorem 4.2 Denote the likelihood of approximated x with the number of segments m as $L_m(x_0, \theta, \lambda)$, and let $L_{\text{true}}(x_0, \theta, \lambda)$ be the likelihood of true x . We should prove that

$$\pi_m(x_0, \theta, \lambda | \mathbf{y}_n) = \frac{L_m(x_0, \theta, \lambda) \pi(x_0, \theta, \lambda)}{\int L_m(x_0, \theta, \lambda) \pi(dx_0, d\theta, d\lambda)}$$

converges to

$$\pi_{\text{true}}(x_0, \theta, \lambda | \mathbf{y}_n) = \frac{L_{\text{true}}(x_0, \theta, \lambda) \pi(x_0, \theta, \lambda)}{\int L_{\text{true}}(x_0, \theta, \lambda) \pi(dx_0, d\theta, d\lambda)}$$

for any x_0, θ and λ . It is well known that if $f(x, t; \theta)$ satisfies Lipschitz condition in x , then Runge-Kutta method converges to the true solution, i.e.

$$x_i^m(x_0, \theta) \rightarrow x_i(x_0, \theta) \quad \text{for all } x_0 \text{ and } \theta \text{ as } m \rightarrow \infty. \quad (32)$$

See Cartwright and Piro (1992) for the proof. The convergence (32) implies that $L_m(x_0, \theta, \lambda)$ converges to $L_{\text{true}}(x_0, \theta, \lambda)$ for all x_0, θ and λ because an exponential function is continuous. It implies the convergence of numerator part.

For the denominator part, recall that

$$L_m(x_0, \theta, \lambda) \leq (\lambda)^{\frac{np}{2}}$$

and $(\lambda)^{np/2}$ is integrable with respect to $\pi(x_0, \theta, \lambda)$. Again, the dominated convergence theorem gives the desired result. ■

Proof of Theorem 4.3 At first, we want to show that under A1 – A3, $|ng_n(x_0) - ng_n^m(x_0)| = O(n(h/m)^K)$ for sufficiently large n . Since we assume the Lipschitz continuity of f , the ODE has a unique solution with initial condition $x(t_1) = x_0$. Assumptions A1 and A3 implies

$$\sup_{x,t} \left\| \frac{d^K}{dt^K} f(x, t; \theta) \right\| =: B < \infty$$

for some constants $B > 0$. The local errors of the K th order numerical method are given by

$$\|x(t_i) - x(t_{i-1}) - h\phi(x_{i-1}, t_{i-1}; \theta)\| \leq B'h^{K+1}, \quad i = 1, \dots, n$$

for some $B' > 0$, which depends only on $\sup_t \|d^K f(x, t; \theta)/(dt^K)\| \leq B$ (Palais and Palais, 2009). Thus, the local errors are uniformly bounded. It implies the global errors uniformly bounded by

$$\|x_i - x_i^h\| \leq Ch^K$$

for some constant $C > 0$. Thus,

$$\begin{aligned} |ng_n(x_0) - ng_n^m(x_0)| &= \left| \sum_{i=1}^n \|y_i - x_i\|^2 - \sum_{i=1}^n \|y_i - x_i^m\|^2 \right| \\ &= \sum_{i=1}^n (\|y_i - x_i\| + \|y_i - x_i^m\|) \left| \|y_i - x_i\| - \|y_i - x_i^m\| \right| \\ &\leq \sum_{i=1}^n (2\|y_i - x_i\| + \|x_i - x_i^m\|) \|x_i - x_i^m\| \\ &\leq \sum_{i=1}^n (2C_y + 2C_x + \|x_i - x_i^m\|) \|x_i - x_i^m\| \\ &\leq \sum_{i=1}^n \left(2C_y + 2C_x + C \left(\frac{h}{m} \right)^K \right) C \left(\frac{h}{m} \right)^K \\ &\asymp n \left(\frac{h}{m} \right)^K, \end{aligned}$$

where $\sup_{t \in [T_0, T_1]} \|y(t)\| < C_y < \infty$, $\sup_{t \in [T_0, T_1]} \|x(t)\| < C_x < \infty$ for sufficiently large n .

By the above inequality, for fixed $x_0 \in \mathbb{R}^p$, $\lambda > 0$,

$$\begin{aligned} e^{-\frac{\lambda}{2} ng_n^m(x_0)} &= e^{-\frac{\lambda}{2} [ng_n(x_0) + ng_n^m(x_0) - ng_n(x_0)]} \\ &= e^{-\frac{\lambda}{2} ng_n(x_0)} \times e^{-\frac{\lambda}{2} [ng_n^m(x_0) - ng_n(x_0)]} \\ &= e^{-\frac{\lambda}{2} ng_n(x_0)} \times e^{-\frac{\lambda}{2} O(n(\frac{h}{m})^K)} \\ &= e^{-\frac{\lambda}{2} ng_n(x_0)} \times \left(1 + O\left(n \left(\frac{h}{m} \right)^K\right) \right) \end{aligned}$$

because $e^x = 1 + O(x)$ for sufficiently small x . It implies

$$\begin{aligned} \pi_m(x_0, \theta, \lambda \mid \mathbf{y}_n) &\propto L_m(\theta, \lambda, x_0) \pi(\theta, \lambda, x_0) \\ &= L^*(\theta, \lambda, x_0) \pi(\theta, \lambda, x_0) \times \left(1 + O\left(n \left(\frac{h}{m} \right)^K\right) \right) \\ &\propto \pi(x_0, \theta, \lambda \mid \mathbf{y}_n) \times \left(1 + O\left(n \left(\frac{h}{m} \right)^K\right) \right) \end{aligned}$$

for sufficiently large n . If $\alpha > (1 + R)/K$, then we have $n(h/m)^K \leq n^{-R}$. \blacksquare

References

- [1] K.T. Alligood, T.D. Sauer, and J.A. Yorke. *Chaos: An Introduction to Dynamical Systems*. Chaos: An Introduction to Dynamical Systems. Springer, 1997.
- [2] Andrea Arnold, Daniela Calvetti, and Erkki Somersalo. Linear multistep methods, particle filtering and sequential monte carlo. *Inverse Problems*, 29(8):085007, 2013.
- [3] Yonathan Bard. *Nonlinear parameter estimation*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974.
- [4] Alexandros Beskos, Ajay Jasra, Kody Law, Raul Tempone, and Yan Zhou. Multilevel sequential monte carlo samplers. *Stochastic Processes and their Applications*, 2016.
- [5] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.
- [6] D.A. Campbell. *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models*. Canadian theses. McGill University (Canada), 2007.
- [7] Julyan H. E. Cartwright and Oreste Piro. The dynamics of runge-kutta methods. *Int. J. Bifurcation and Chaos*, 2:427–49, 1992.
- [8] Carlos M. Carvalho, Michael Johannes, Hedibert F. Lopes, and Nicholas Polson. Particle learning and smoothing. *Statistical Science*, pages 88–106, 2010.
- [9] Sarat C. Dass, Jaeyong Lee, Kyoungjae Lee, and Jonghun Park. Laplace based approximate posterior inference for differential equation models. *Statistics and Computing*, 27(3):679–698, 2017.
- [10] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.
- [11] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

- [12] Arnaud Doucet, Nando De Freitas, Kevin Murphy, and Stuart Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 176–183. Morgan Kaufmann Publishers Inc., 2000.
- [13] R.M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.
- [14] Paul Fearnhead. Markov chain monte carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862, 2002.
- [15] R. FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1:445–466, 1961.
- [16] Andrew Gelman, Frederic Bois, and Jiming Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91:1400–1412, 1996.
- [17] Charles J. Geyer and Minnesota Univ Minneapolis School Of Statistics. *Markov Chain Monte Carlo Maximum Likelihood*. Defense Technical Information Center, 1992.
- [18] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113, 1993.
- [19] JE Handschin and David Q Mayne. Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering? *International journal of control*, 9(5):547–559, 1969.
- [20] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [21] Giles Hooker, Stephen P Ellner, Laura De Vargas Roditi, and David JD Earn. Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in ontario. *Journal of The Royal Society Interface*, page rsif20100412, 2010.

- [22] Yangxin Huang, Dacheng Liu, and Hulin Wu. Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, 62(2):413–423, 2006.
- [23] Frank P. Incropera. *Fundamentals of Heat and Mass Transfer*. John Wiley & Sons, 2006.
- [24] Nicholas Kantas, Arnaud Doucet, Sumeetpal Sindhu Singh, and Jan Marian Maciejowski. An overview of sequential monte carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes*, 42(10):774–785, 2009.
- [25] Yun Li. *Regularized Statistical Methods for Data of Grouped or Dynamic Nature*. PhD thesis, The University of Michigan, 2012.
- [26] Jane Liu and Mike West. Combined parameter and state estimation in simulation-based filtering. In De Freitas and N. J. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.
- [27] Hedibert F Lopes and Ruey S Tsay. Particle filters and bayesian inference in financial econometrics. *Journal of Forecasting*, 30(1):168–209, 2011.
- [28] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.
- [29] A.J. Lotka. *Elements of Physical Biology*. Williams & Wilkins Company, 1925.
- [30] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [31] J. Nagumo, S. Arimoto, and S. Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50:2061–2070, 1962.
- [32] Richard S. Palais and Robert A. Palais. *Differential Equations, Mechanics, and Computation*, volume 51 of *Student mathematical library; IAS/Park City mathematical subseries*. 2009.
- [33] Michael K. Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):pp. 590–599, 1999.

- [34] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(5):741–796, 2007.
- [35] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition, June 2005.
- [36] Maria Paula Rios and Hedibert Freitas Lopes. The extended liu and west filter: Parameter learning in markov switching stochastic volatility models. In Yong Zeng and Shu Wu, editors, *State-Space Models*, volume 1 of *Statistics and Econometrics for Finance*, pages 23–61. Springer New York, 2013.
- [37] Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.
- [38] Karline Soetaert and Thomas Petzoldt. Inverse modelling, sensitivity and monte carlo analysis in R using package FME. *Journal of Statistical Software*, 33(3):1–28, 2010.
- [39] KER Soetaert and Thomas Petzoldt. Inverse modelling, sensitivity and monte carlo analysis in r using package fme. *Journal of Statistical Software*, 33, 2010.
- [40] M. N. Spijker. Error propagation in Runge–Kutta methods. 22(1–3):309–325, December 1996. Special issue celebrating the centenary of Runge–Kutta methods.
- [41] Geir Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on signal Processing*, 50(2):281–289, 2002.
- [42] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.*, 3(1):28–46, 1982.
- [43] V. Volterra. *Variazioni e fluttuazioni del numero d’individui in specie animali conviventi*. Atti della R. Accademia Nazionale dei Lincei. C. Ferrari, 1927.