Accepted Manuscript

Bayesian inference for higher-order ordinary differential equation models

Prithwish Bhaumik, Subhashis Ghosal

PII: S0047-259X(17)30160-4
DOI: http://dx.doi.org/10.1016/j.jmva.2017.03.003
Reference: YJMVA 4231

To appear in: *Journal of Multivariate Analysis*

Received date: 23 February 2016

Please cite this article as: P. Bhaumik, S. Ghosal, Bayesian inference for higher-order ordinary differential equation models, *Journal of Multivariate Analysis* (2017), http://dx.doi.org/10.1016/j.jmva.2017.03.003

Bayesian inference for higher-order ordinary differential equation models

Prithwish Bhaumik[a,*], Subhashis Ghosal[b]

[a] *Quantifind Inc., 8 Homewood Place, Menlo Park, CA 94025, USA*
[b] *Department of Statistics, North Carolina State University, SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695–8203, USA*

Abstract

Often the regression function appearing in fields like economics, engineering, and biomedical sciences obeys a system of higher-order ordinary differential equations (ODEs). The equations are usually not analytically solvable. We are interested in inferring on the unknown parameters appearing in such equations. Parameter estimation in first-order ODE models has been well investigated. Bhaumik and Ghosal [4] considered a two-step Bayesian approach by putting a finite random series prior on the regression function using a B-spline basis. The posterior distribution of the parameter vector is induced from that of the regression function. Although this approach is computationally fast, the Bayes estimator is not asymptotically efficient. Bhaumik and Ghosal [5] remedied this by directly considering the distance between the function in the nonparametric model and a Runge–Kutta (RK4) approximate solution of the ODE while inducing the posterior distribution on the parameter. They also studied the convergence properties of the Bayesian method based on the approximate likelihood obtained by the RK4 method. In this paper, we extend these ideas to the higher-order ODE model and establish Bernstein–von Mises theorems for the posterior distribution of the parameter vector for each method with $n^{-1/2}$ contraction rate.

[*] Corresponding author
Email addresses: prithwish.bhaumik@utexas.edu (Prithwish Bhaumik), sghosal@ncsu.edu (Subhashis Ghosal)

Keywords: Bayesian inference, Bernstein–von Mises theorem, higher order ordinary differential equation, Runge–Kutta method, spline smoothing.
2010 MSC: 62J02, 62G08, 62G20, 62F12, 62F15.

1. Introduction

The regression relationship between a response variable and a predictor variable, which usually stands for time, is sometimes implicitly given by a differential equation. Consider a regression model $Y = f_{\boldsymbol{\theta}}(t) + \boldsymbol{\varepsilon}$ with unknown parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ and $t \in [0,1]$. The functional form of $f_{\boldsymbol{\theta}}$ is not known but $f_{\boldsymbol{\theta}}$ is assumed to satisfy a qth order ordinary differential equation (ODE) given by

$$F\left(t, f_{\boldsymbol{\theta}}(t), \frac{df_{\boldsymbol{\theta}}(t)}{dt}, \ldots, \frac{d^q f_{\boldsymbol{\theta}}(t)}{dt^q}, \boldsymbol{\theta}\right) = 0, \tag{1}$$

where F is a known real-valued function which is sufficiently smooth in its arguments; we shall refer to it as the binding function. Regression models based on first-order ODE have been well studied from both non-Bayesian and Bayesian points of view; see, e.g., [4, 5, 9, 10].

Higher-order ODE models are encountered in different fields. For example, the system of ODE describing the concentrations of glucose and hormone in blood is given by

$$\begin{aligned} \frac{dk(t)}{dt} &= -m_1 k(t) - m_2 h(t) + J(t), \qquad (2)\\ \frac{dh(t)}{dt} &= -m_3 h(t) + m_4 k(t), \end{aligned}$$

where $k(t)$ and $h(t)$ denote the glucose and hormone concentrations at time t, respectively. Here the function J is known and m_1, m_2, m_3 and m_4 are unknown parameters. If we have only measurements on k, we can eliminate h by differentiating both sides of (2) to obtain the second-order ODE for k as

$$\frac{d^2k(t)}{dt^2} + 2\alpha\frac{dk(t)}{dt} + \omega_0^2 k(t) = S(t),$$

where $\alpha = (m_1 + m_3)/2$, $\omega_0^2 = m_1 m_3 + m_2 m_4$ and $S(t) = m_3 J(t) + dJ(t)/dt$. Another popular example is the Van der Pol oscillator used in physical and biological sciences. The oscillator obeys the second-order ODE

$$\frac{d^2 f_\theta(t)}{dt^2} - \theta\{1 - f_\theta^2(t)\}\frac{df_\theta(t)}{dt} + f_\theta(t) = 0.$$

A related problem is a stochastic differential equation model where a signal is continuously observed in time with a noise process typically driven by a Brownian motion. Bergstrom [1, 2, 3] used the maximum likelihood estimation (MLE) technique to estimate the parameters involved in the higher-order stochastic differential equation given by

$$\frac{d^q y(t)}{dt^q} = A_1(\boldsymbol{\theta})\frac{d^{q-1}y(t)}{dt^{q-1}} + \ldots + A_{q-1}(\boldsymbol{\theta})\frac{dy(t)}{dt} + A_q y(t) + b(\boldsymbol{\theta}) + z(t) + W(t),$$

where $A_1, \ldots, A_q$ and b are functions on Θ, W is the noise process [8, p. 342] and z is a deterministic function. Bergstrom [1] showed that the maximum likelihood estimator of $\boldsymbol{\theta}$ is asymptotically normal and asymptotically efficient. An efficient algorithm was given in Bergstrom [2] to compute the Gaussian likelihood for estimating the parameters involved in a non-stationary higher-order stochastic ODE. Appropriate linear transformations are used in this algorithm to avoid the computation of the covariance matrix of the observations.

In this paper, we develop two Bayesian approaches for inference on $\boldsymbol{\theta}$ by embedding the ODE model in a nonparametric regression model, where a prior is put on the regression function through a B-spline basis expansion technique and the posterior is computed. The posterior on the parameter $\boldsymbol{\theta}$ of the ODE model is then induced directly by a distance minimization method, viewing $\boldsymbol{\theta}$ as a functional of the regression function. Depending on the choice of the distance function, two different two-step Bayesian methods are obtained. In the first approach we define

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\Theta} \int_0^1 \left| F\left(t, f(t,\boldsymbol{\beta}), \frac{df(t,\boldsymbol{\beta})}{dt}, \ldots, \frac{d^q f(t,\boldsymbol{\beta})}{dt^q}, \boldsymbol{\eta}\right)\right|^2 w(t)dt,$$

where the weight function and its first $q-1$ derivatives vanish at 0 and 1. Here $\boldsymbol{\beta}$ is the coefficient vector of the B-spline basis expansion. The posterior distribution of $\boldsymbol{\theta}$ is induced using the posterior distribution of $\boldsymbol{\beta}$. The method completely avoids a numerical resolution of the ODE and hence is computationally very convenient. Nevertheless this approach does not produce asymptotically efficient Bayes estimator. In our second approach, we use the Runge–Kutta method to obtain an approximate solution $f_{\boldsymbol{\theta},r_n}$ using r_n grid points, where n is the number of observations. This time the parameter is defined as

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\Theta} \int_0^1 \{f(t,\boldsymbol{\beta}) - f_{\boldsymbol{\eta},r_n}(t)\}^2 g(t)dt,$$

where g is an appropriate weight function. This method produces an asymptotically efficient Bayes estimator as shown later. As in the first approach we get the posterior of $\boldsymbol{\theta}$ from the posterior of $\boldsymbol{\beta}$. For the sake of simplicity we consider only one-dimensional regression functions. Extension to the multidimensional case where the binding function F is also vector-valued can be carried out similarly.

The rest of the paper is organized as follows. Section 2 contains descriptions of the estimation methods used. Convergence results are presented in Section 3. The algorithms associated with the different methods are described in Section 4. Section 5 reports the results of a simulation study. Proofs of the results are given in Section 6. The Appendix provides a useful description of the Runge–Kutta method for solving higher-order ODEs.

2. Description of proposed methodology

We observe n samples $t_1,\ldots,t_n$ of the predictor variable along with the corresponding values $Y_1,\ldots,Y_n$ of the response variable. The relation between the two variables is modeled by non-linear regression

$$Y_i = f_{\boldsymbol{\theta}}(t_i) + \varepsilon_i, \quad i \in \{1,\ldots,n\} \tag{3}$$

where the unknown parameter $\boldsymbol{\theta}$ belongs to a compact subset Θ of $\mathbb{R}^p$. The regression function $f_{\boldsymbol{\theta}}$ is not given explicitly but is assumed to be q times differentiable on an open set containing $[0,1]$ and satisfies the higher-order ODE given by

$$F(t, f_{\boldsymbol{\theta}}(t), f_{\boldsymbol{\theta}}^{(1)}(t),\ldots,f_{\boldsymbol{\theta}}^{(q)}(t),\boldsymbol{\theta}) = 0, \tag{4}$$

where for every fixed $\boldsymbol{\theta}$, we assume that $F(\cdot,\cdot,\boldsymbol{\theta}) \in \mathcal{C}^{m-q+1}[(0,1) \times \mathbb{R}^{q+1}]$ for some integer $m \geq q$. Then, by successive differentiation we have $f_{\boldsymbol{\theta}} \in \mathcal{C}^m[(0,1)]$. We also assume that the function $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\cdot)$ is two times continuously differentiable. We do not assume that the true regression function f_0 satisfies the ODE (4) for any value of $\boldsymbol{\theta}$, i.e., the model may be misspecified, but only assume the mild regularity condition $f_0 \in \mathcal{C}^m([0,1])$.

The regression errors $\varepsilon_1,\ldots,\varepsilon_n$ are assumed to be independently and identically distributed with mean zero and finite fourth order moment. Let the common variance be denoted by σ_0^2. We use $\mathcal{N}(0,\sigma^2)$ as the working model for the error, which may be different from the true distribution of the errors. We treat σ^2 as a nuisance parameter and assign an inverse gamma prior on σ^2 with shape and scale parameters a and b respectively. Additionally it is assumed that $t_1,\ldots,t_n$ form a random sample from distribution G with density g.

Due to the intractable nature of the likelihood function, the standard method of assigning a prior distribution on the parameter of interest $\boldsymbol{\theta}$ and then updating to the posterior distribution is extremely computationally expensive. To avoid the problem, we embed the parametric model implicitly given by the solution of the ODE in a nonparametric regression model, where we assign a prior distribution on the regression function using the basis expansion method of constructing priors on infinite-dimensional spaces. Then the parameter θ can be viewed as a functional of the regression function.

In the basis expansion method, a prior is induced on the regression function through that on the coefficient vector. The assumed normal regression model conjugacy can be obtained by using a multivariate normal prior. The resulting posterior distribution directly induces a posterior distribution on the parameter of interest. We refer to this technique as the Bayesian two-step method.

The functional expressing the parameter in terms of the regression function, can be defined by minimization method as in a minimum contrast estimation approach. Depending on the choice of our distance, we can define different functionals and hence different two-step posterior distributions. Below we describe two such methods, one building on a distance based on the differential equation's structure and the other based on a numerical solution of the equation. The first approach avoids numerical resolution of the differential equation and hence is computationally faster but is incapable of giving asymptotically efficient Bayes estimators. The second approach is computationally more involved but will be shown to produce asymptotically efficient Bayes estimators.

Let us denote $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ and $\boldsymbol{t} = (t_1, \dots, t_n)^\top$. The true joint distribution of (t_i, ε_i) is denoted by P_0. Assume that for all $i \in \{1, \dots, n\}$, $Y_i = f(t_i) + \varepsilon_i$, where f is a sufficiently smooth function. We construct a finite random series prior for f using a B-spline basis expansion

$$f(t) = f(t, \boldsymbol{\beta}) = \sum_{j=1}^{k_n+m-1} \beta_j N_j(t)$$

with coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{k_n+m-1})^\top$, where $\{N_j(\cdot)\}_{j=1}^{k_n+m-1}$ are the B-spline basis functions of order m with uniformly spread $k_n - 1$ interior knots; see Chapter IX of De Boor [6]. It is then possible to express the nonparametric model as a linear model controlled by the parameter $\boldsymbol{\beta}$

$$\boldsymbol{Y} = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}_n = ((N_j(t_i)))_{1 \le i \le n, 1 \le j \le k_n+m-1}$. Given σ^2, we let $\boldsymbol{\beta}$ have the $(k_n + m - 1)$-dimensional multivariate normal prior

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 n^2 k_n^{-1} \boldsymbol{I}_{k_n+m-1}).$$

Using normal-inverse gamma conjugacy, a simple calculation yields that the conditional posterior distribution of $\boldsymbol{\beta}$ given σ^2 is

$$\mathcal{N}\left[\left(\boldsymbol{X}_n^T \boldsymbol{X}_n + \frac{k_n}{n^2}\boldsymbol{I}_{k_n+m-1}\right)^{-1} \boldsymbol{X}_n^T \boldsymbol{Y}, \sigma^2 \left(\boldsymbol{X}_n^T \boldsymbol{X}_n + \frac{k_n}{n^2}\boldsymbol{I}_{k_n+m-1}\right)^{-1}\right]. \quad (5)$$

2.1. Two-step Bayesian Method (TSB)

Since the parameter $\boldsymbol{\theta}$ is defined only within the ODE model but the regression function in the nonparametric model can be general, it is essential to extend the definition of the parameter $\boldsymbol{\theta}$ for any smooth regression function.

Naturally, the extended definition must agree with the original definition on the ODE model.

We use a projection method, i.e., locating a parameter value such that the solution of the ODE for that parameter value is, in some sense, closest to the targeted regression function, among all possible regression functions described by the ODE model. A sensible choice of a distance is given by the ODE itself. If f belongs to the ODE model, then $F\left(\cdot, f(\cdot,\boldsymbol{\beta}), f^{(1)}(\cdot,\boldsymbol{\beta}), \ldots, f^{(q)}(\cdot,\boldsymbol{\beta}), \boldsymbol{\theta}\right)$ is identically 0 for some value of $\boldsymbol{\theta}$. This suggests, for an arbitrary smooth function f, the value of $\boldsymbol{\theta}$ that makes some appropriate norm of the function F closest to 0 may be defined as the parameter value corresponding to f. That is, define a functional $\boldsymbol{\theta} = \boldsymbol{\theta}(f)$ by

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\Theta} \left\| F\left(\cdot, f(\cdot,\boldsymbol{\beta}), f^{(1)}(\cdot,\boldsymbol{\beta}), \ldots, f^{(q)}(\cdot,\boldsymbol{\beta}), \boldsymbol{\eta}\right)\right\|_w, \tag{6}$$

where $f^{(r)}(t,\boldsymbol{\beta}) = \frac{d^r}{dt^r} f(t,\boldsymbol{\beta})$ for every $r \in \{1,\ldots,q\}$, and $\|\phi\|_w^2 = \int |\phi(t)|^2 w(t)dt$ stands for the squared L_2-norm with respect to a nonnegative weight function w. Clearly, $\boldsymbol{\theta}(f_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$, so the new definition of the parameter $\boldsymbol{\theta}$ truly extends the original definition. The true value of the parameter is defined by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\eta}\in\Theta} \left\| F\left(\cdot, f_0(\cdot), f_0^{(1)}(\cdot), \ldots, f_0^{(q)}(\cdot), \boldsymbol{\eta}\right)\right\|_w. \tag{7}$$

We induce posterior distribution on $\boldsymbol{\theta}$ through the posterior of $\boldsymbol{\beta}$ given by (5). Note that in the well-specified case where $f_{\boldsymbol{\theta}_0}$ is the true regression function with corresponding true parameter $\boldsymbol{\theta}_0$, then the minimum is automatically located at the true value $\boldsymbol{\theta}_0$. Multiple minima of (6) are allowed (in which case we choose the value of $\boldsymbol{\theta}$ arbitrarily among all minima), but for good large-sample behavior of the posterior distribution, the true value must be uniquely defined. In fact we shall assume that the location of the minimum (7) is strongly separated in the sense that the minimum value is not approached by other parameter values not close to the true value, i.e., for all $\epsilon > 0$,

$$\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} \left\| F\left(\cdot, f_0(\cdot), f_0^{(1)}(\cdot), \ldots, f_0^{(q)}(\cdot), \boldsymbol{\eta}\right)\right\|_w > \left\| F\left(\cdot, f_0(\cdot), f_0^{(1)}(\cdot), \ldots, f_0^{(q)}(\cdot), \boldsymbol{\theta}_0\right)\right\|_w. \tag{8}$$

Note that if the ODE model is identifiable, the uniqueness of $\boldsymbol{\theta}_0$ is automatically satisfied, but the strong separation condition will be an additional condition.

95 2.2. Runge–Kutta Two-Step Bayesian Method (RKTSB)

Here we use the same nonparametric model and prior specifications as in the two-step Bayesian method, but $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ are defined by

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\eta}\in\Theta} \int_0^1 |f(t,\boldsymbol{\beta}) - f_{\boldsymbol{\eta},r_n}(t)|^2 g(t)dt, \qquad (9) \\ \boldsymbol{\theta}_0 &= \arg\min_{\boldsymbol{\eta}\in\Theta} \int_0^1 |f_0(t) - f_{\boldsymbol{\eta}}(t)|^2 g(t)dt, \end{aligned}$$

g being the density function of the regressor variable. We also assume the strong separability condition that, for all $\epsilon > 0$,

$$\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} \int_0^1 |f_0(t) - f_{\boldsymbol{\eta}}(t)|^2 g(t)dt > \int_0^1 |f_0(t) - f_{\boldsymbol{\theta}_0}(t)|^2 g(t)dt,$$

i.e., $\int_0^1 |f_0(t) - f_{\boldsymbol{\eta}}(t)|^2 g(t)dt$ has a well separated unique minimum at $\boldsymbol{\theta}_0$.

3. Asymptotic properties

100 3.1. Two-step Bayesian Method (TSB)

Let $\boldsymbol{h}(\cdot) = (f(\cdot,\boldsymbol{\beta}), f^{(1)}(\cdot,\boldsymbol{\beta}), \ldots, f^{(q)}(\cdot,\boldsymbol{\beta}))^\top$ and $\boldsymbol{h}_0$ stand for $\boldsymbol{h}$ with f being replaced by f_0. We assume that w is sufficiently smooth and w as well as its first $q-1$ derivatives vanish at 0 and 1. Let

$$\boldsymbol{G}(t,\boldsymbol{h}(t),\boldsymbol{\theta}) = \{D_{\boldsymbol{\theta}} F(t,\boldsymbol{h}(t),\boldsymbol{\theta})\}^\top F(t,\boldsymbol{h}(t),\boldsymbol{\theta}),$$

where $D_\bullet$ stands for the differentiation operation with respect to the indicated argument at its stated value.

The key step in studying asymptotic properties of the two-step Bayesian method is a linearization step representing the nonlinear functional $\boldsymbol{\theta}$ of f as an approximate linear functional. The following lemma controls the error in the linearziation step. Below $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$, $\mathcal{C}^m(E)$ refers to the space of m times continuously differentiable functions on an open set containing E and the symbols $\boldsymbol{A}_{i,}$ and $\boldsymbol{A}_{,j}$ stand for the ith row and jth column of a matrix $\boldsymbol{A}$, respectively. The symbol $o_P(1)$ stands for a sequence of random variables converging in P-probability to 0.

Lemma 1. *Let the matrix $\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0) = \int_0^1 D_{\boldsymbol{\theta}_0} \{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\} w(t)dt$ be non-singular and assume that (8) holds. If $m > 2q+2$ and $n^{1/2m} \ll k_n \ll n^{1/(4q+4)}$, then there exists $E_n \subseteq \boldsymbol{\Theta} \times \mathcal{C}^m[(0,1)]$ with $\Pi(E_n^c|\boldsymbol{t}, \boldsymbol{Y}) = o_{P_0}(1)$, such that*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{h})\in E_n} \left\| \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \{\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0)\}^{-1} \sqrt{n} \{\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0)\} \right\| \to 0,$$

where

$$\boldsymbol{\Gamma}(z) = -\sum_{r=0}^{q} \int_0^1 (-1)^r \frac{d^r}{dt^r} \left[D_{\boldsymbol{h}_0} \{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\} w(t)\right]_{,r} z(t)dt$$

is a linear functional of z for any function $z : [0,1] \mapsto \mathbb{R}$.

Denoting

$$\boldsymbol{A}(t) = -\{\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0)\}^{-1} \sum_{r=0}^{q} (-1)^r \frac{d^r}{dt^r} \left[D_{\boldsymbol{h}_0} \{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\} w(t)\right]_{,r},$$

we have

$$\{\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0)\}^{-1} \boldsymbol{\Gamma}(f) = \int_0^1 \boldsymbol{A}(t)\boldsymbol{\beta}^\top \boldsymbol{N}(t)d\boldsymbol{t} = \boldsymbol{H}_n^\top \boldsymbol{\beta}, \tag{10}$$

where $\boldsymbol{H}_n^\top = \int_0^1 \boldsymbol{A}(t)\boldsymbol{N}^\top(t)dt$, a matrix of order $p \times (k_n + m - 1)$. Then in order to approximate the posterior distribution of $\boldsymbol{\theta}$, it suffices to study the asymptotic posterior distribution of the linear functional of $\boldsymbol{\beta}$ given by (10). The next theorem describes the approximate posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. We denote the posterior probability measure of the two-step method by Π_n^*.

Theorem 1. *Let us denote*

$$\begin{aligned} \boldsymbol{\mu}_n^* &= \sqrt{n}\, \boldsymbol{H}_n^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{Y} - \sqrt{n} \{\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0)\}^{-1} \boldsymbol{\Gamma}(f_0), \\ \boldsymbol{\Sigma}_n^* &= n\boldsymbol{H}_n^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{H}_n \end{aligned}$$

and $\boldsymbol{D} = ((\int_0^1 A_k(t)A_{k'}(t)dt))_{k,k'=1,\ldots,p}$. If $\boldsymbol{D}$ is non-singular, then under the conditions of Lemma 1,

$$\left\| \Pi_n^* \left\{ \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot | \boldsymbol{t}, \boldsymbol{Y} \right\} - \mathcal{N}\left(\boldsymbol{\mu}_n^*, \sigma_0^2 \boldsymbol{\Sigma}_n^*\right) \right\|_{TV} = o_{P_0}(1).$$

Remark 1. Following the steps of the proof of Lemma 4 of Bhaumik and Ghosal [4], it can be proved that both $\boldsymbol{\mu}_n^*$ and $\boldsymbol{\Sigma}_n^*$ are stochastically bounded. Hence, with high true probability the posterior distribution of $\boldsymbol{\theta}$ contracts at $\boldsymbol{\theta}_0$ at the rate $n^{-1/2}$.

Remark 2. Similar results will follow for deterministic covariates provided that

$$\sup_{t\in[0,1]} |Q_n(t) - Q(t)| = o(k_n^{-1}),$$

125 where Q_n is the empirical distribution function of the covariate sample and Q is a distribution function with positive density on [0,1]. Note that this condition holds with probability tending to 1 when the covariates are random.

3.2. Runge–Kutta Two-Step Bayesian Method (RKTSB)

In RKTSB we assume that the matrix

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = -\int_0^1 \ddot{f}_{\boldsymbol{\theta}_0}(t)\{f_0(t) - f_{\boldsymbol{\theta}_0}(t)\}g(t)dt + \int_0^1 \{\dot{f}_{\boldsymbol{\theta}_0}(t)\}^\top \{\dot{f}_{\boldsymbol{\theta}_0}(t)\}g(t)dt$$

130 is nonsingular, where $\dot{f}_{\boldsymbol{\theta}_0}$ refers to the vector derivative of $f_{\boldsymbol{\theta}}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\ddot{f}_{\boldsymbol{\theta}_0}$ is the matrix of mixed partial derivatives of order two at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Note that in the well-specified case, the first term vanishes and hence $\boldsymbol{J}(\boldsymbol{\theta}_0)$ equals the second term which is positive definite. Let us denote $\boldsymbol{C}(t) = \{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1}\{\dot{f}_{\boldsymbol{\theta}_0}(t)\}^\top$ and $\boldsymbol{G}_n^\top = \int_0^1 \boldsymbol{C}(t)\boldsymbol{N}^\top(t)g(t)dt$. Also, we denote the posterior probability measure 135 of RKTSB by Π_n^{**}. Now we have the following result.

Theorem 2. *Let*

$$\begin{aligned}
\boldsymbol{\mu}_n^{**} &= \sqrt{n}\,\boldsymbol{G}_n^\top \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_n^\top \boldsymbol{Y} - \sqrt{n}\{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1} \int_0^1 \{\dot{f}_{\boldsymbol{\theta}_0}(t)\}^\top f_0(t)g(t), \\
\boldsymbol{\Sigma}_n^{**} &= n\,\boldsymbol{G}_n^\top \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n\right)^{-1} \boldsymbol{G}_n, \\
\boldsymbol{B} &= ((\langle C_k(\cdot), C_{k'}(\cdot)\rangle_g))_{k,k'=1,\ldots,p},
\end{aligned}$$

where $\langle \cdot,\cdot\rangle_g$ refers to inner product with respect to the density g. If $\boldsymbol{B}$ is nonsingular, then for $r_n \gg n^{1/8}$, $m \geq 3$ and $n^{1/(2m)} \ll k_n \ll n^{1/4}$,

$$\left\|\Pi_n^{**}\{\sqrt{n}\,(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot|\boldsymbol{t}, \boldsymbol{Y}\} - \mathcal{N}\left(\boldsymbol{\mu}_n^{**}, \sigma_0^2 \boldsymbol{\Sigma}_n^{**}\right)\right\|_{TV} = o_{P_0}(1).$$

We also get the following important corollary.

Corollary 1. *When the regression model* (3) *is correctly specified and the true distribution of error is Gaussian, the Bayes estimator based on* Π_n^{**} *is asymptotically efficient.*

As in the two-step Bayesian approach, if the regressor is deterministic, we also get similar results under appropriate conditions.

Remark 3. We do not require that the true regression function f_0 is a solution of the ODE. Like any statistical model, an ODE model is only a relatively simple mathematical description of a phenomenon under study and is deemed only approximate in real life. Nonetheless the parameter values stand for certain characteristics of the system. Although the model may be misspecified, we may be interested in the parameters rather than the regression function described by the ODE model. Our posterior concentration results show that the definition of $\boldsymbol{\theta}_0$ as the parameter value which brings the ODE model closest to the true regression function is appropriate.

Remark 4. Note that in both approaches the true regression function is assumed to have a certain degree of smoothness to ensure the contraction rate $n^{-1/2}$, which obviously cannot be improved by exploiting additional smoothness, if it exists, in the regression function as a function of the predictor variable t. Therefore, the issue of adapting with smoothness does not arise in the context of ODE models. As a result, we may simply choose the number of knots of the spline functions guided by the conditions given in Lemma 1 or Theorem 2, as appropriate, corresponding to the minimum allowed smoothness level m and then the knots themselves deterministically, for instance as uniformly spread out. This avoids putting priors on these quantities and hence substantially simplifies the computation.

Remark 5. In the frequentist framework, nonlinear least squares (NLS) is a popular way of estimating $\boldsymbol{\theta}$. In the absence of an explicit functional form of the regression function, the likelihood function can be computed numerically using the RK4 method ODE solver with a sufficient number of grid points, so that the error in the ODE solver can be appropriately controlled. Then using arguments similar to those in Section 3.2, it can be concluded that the estimator is also asymptotically normal with mean $\boldsymbol{\theta}_0$ and dispersion matrix given by $\boldsymbol{\Sigma}_n^{**}$. It is then an asymptotically efficient estimator of $\boldsymbol{\theta}$ in the well-specified setting. The computational method for NLS is given in Section 4.3. We compare this technique with the Bayesian techniques in the simulation study.

4. Algorithms

For a given data set, the algorithms associated with each method are briefly described below.

4.1. Algorithm for TSB method

Step 1: Draw a posterior sample for σ^2. Because of conjugacy σ^2 has an inverse gamma posterior with shape and scale parameters $(n+2a)/2$ and $b + \boldsymbol{Y}^\top\{\boldsymbol{I}_n - (\boldsymbol{X}_n^\top\boldsymbol{X}_n + n^{-2}k_n\boldsymbol{I}_{k_n+m-1})^{-1}\}\boldsymbol{Y}/2$, respectively.

Step 2: Draw a sample from the conditional posterior of $\boldsymbol{\beta}$ given σ^2 as mentioned in (5).

Step 3: Obtain a posterior sample of $\boldsymbol{\theta}$ from that of $\boldsymbol{\beta}$ using (6).

Step 4: Repeat the above steps until a sufficient number of samples has been collected. Obtain the Bayes estimate (pointwise posterior mean or median) and 95% credible interval of each component of $\boldsymbol{\theta}$.

4.2. Algorithm for RKTSB method

Step 1: Draw a posterior sample for σ^2 as in Algorithm 4.1.

Step 2: Draw a sample from the conditional posterior of $\boldsymbol{\beta}$ given σ^2 as mentioned in (5).

Step 3: Obtain the numerical solution of the ODE $f_{\boldsymbol{\eta},r_n}(t)$ for every t on a fine grid in $[0,1]$ and every $\boldsymbol{\eta}$ on a fine grid in Θ using the Runge–Kutta method.

Step 4: Obtain a posterior sample for $\boldsymbol{\theta}$ from that of $\boldsymbol{\beta}$ by solving the nonlinear optimization problem (9). As the initial choice of $\boldsymbol{\theta}$ we choose the estimate obtained from the two-step Bayesian method.

Step 5: Repeat the above steps until a sufficient number of samples has been collected and obtain the Bayes estimate and credible intervals as before.

4.3. Algorithm for NLS method

Step 1: Obtain the numerical solution of the ODE $f_{\eta,r_n}(t)$ at every point t on a fine grid in $[0,1]$ and every $\boldsymbol{\eta}$ on a fine grid in Θ using Runge–Kutta method. Find the function values at any intermediate t by spline interpolation.

Step 2: Estimate $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\eta}\in\Theta} \sum_{i=1}^{n} \{Y_i - f_{\boldsymbol{\eta},r_n}(t_i)\}^2.$$

Step 3: Obtain the 95% confidence interval of $\boldsymbol{\theta}$ using the fact that for large n, $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ follows approximately a normal distribution with mean zero and dispersion matrix $\sigma_0^2[\int_0^1 \{\dot{f}_{\boldsymbol{\theta}_0}(t)\}^\top \dot{f}_{\boldsymbol{\theta}_0}(t)dt]^{-1}$. This matrix is the inverse Fisher information as obtained in the proof of Corollary 1. To construct the confidence interval we estimate σ^2 by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \{Y_i - f_{\hat{\boldsymbol{\theta}},r_n}(t_i)\}^2$$

and plug-in for σ_0^2. Thus the approximate 95% confidence interval for θ_j is centered at $\hat{\theta}_j$ with length

$$2\frac{\tau_{0.025}}{\sqrt{n}}\hat{\sigma}\left[\left(\int_0^1 \{\dot{f}_{\hat{\boldsymbol{\theta}}}(t)\}^\top \dot{f}_{\hat{\boldsymbol{\theta}}}(t)dt\right)^{-1}\right]_{j,j}^{1/2}$$

for all $j \in \{1,\ldots,p\}$, $\tau_{0.025}$ being the 97.5% percentile of a standard normal distribution and $A_{j,j}$ stands for the jth diagonal element of the matrix $\boldsymbol{A}$.

5. Simulation study

We consider the van der Pol equation

$$\frac{d^2 f_\theta(t)}{dt^2} - \theta\{1 - f_\theta^2(t)\}\frac{df_\theta(t)}{dt} + f_\theta(t) = 0, \quad t \in [0,1]$$

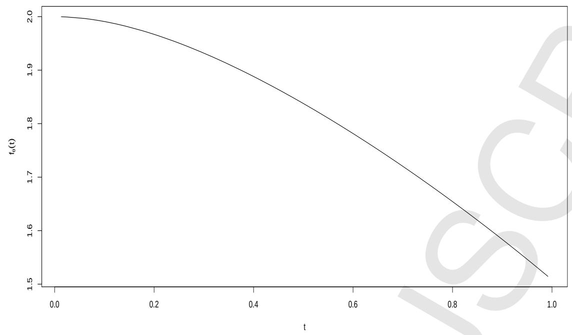

Figure 1: *A sample trajectory of van der Pol equation for* $\theta = 1$

with the initial conditions $f_\theta(0) = 2, f'_\theta(0) = 0$, to study the posterior distribution of θ.

The above system is not analytically solvable. This equation has been used to model a variety of physical and biological phenomena. For instance, in biology, the van der Pol equation is used to model coupled neurons in the gastric mill circuit of the stomatogastric ganglion. In seismology, this equation is often used to model the interaction of two plates in a geological fault.

We consider the situation where the true regression function belongs to the solution set. The true parameter is taken to be $\theta_0 = 1$. A sample trajectory is shown in Figure 1. For a sample of size n, the predictor variables $t_1, \ldots, t_n$ are drawn from the $\mathcal{U}(0,1)$ distribution. Samples of sizes 50, 100 and 500 are considered. We simulate 1000 replications for each case. Under each replication, a sample of size 1000 is drawn from the posterior distribution of θ using two-step Bayesian method (TSB) and RKTSB and then 95% equal tailed credible intervals are obtained. The simulation results are summarized in Tables 1–3.

We choose the number of grid points in the order of $n^{0.26}$ to construct the numerical solution of the ODE. We calculate the coverage and the average length of the corresponding credible interval over the 1000 replications. We also compare the two methods with the nonlinear least squares (NLS) technique based on exhaustive numerical solution of the ODE where we construct a 95% confidence interval using asymptotic normality. The estimated standard errors of the interval length and coverage are given in parentheses in the tables.

The true distribution of error is taken to be either $\mathcal{N}[0, (0.1)^2]$ or the scaled t distribution with 6 degrees of freedom, or a centered and scaled gamma distribution with both shape and scale parameters 1. The scaling and centering are done in order to make the mean 0 and standard deviation 0.1. We put an inverse gamma prior on σ^2 with shape and scale parameters being 99 and 1, respectively.

The choice of prior parameters is guided by the estimate of error variance from the data. This helps maintain reasonable size and coverage of credible intervals. But in larger sample sizes the effect of the prior diminishes and any sensible choice will lead to similar results. We take $m = 7$ and $m = 4$ for TSB and RKTSB respectively. We choose k_n as 3, 3 and 4 for $n = 50$, $n = 100$ and $n = 500$ respectively in TSB. In RKTSB the choices are 3, 3 and 4 for $n = 50$, $n = 100$ and $n = 500$ respectively. The choices of m and asymptotic orders of k_n are done in accordance with Lemma 1 and Theorem 2, respectively. The appropriate multiples to the chosen asymptotic orders are determined using cross validation. The weight function for TSB is chosen as $w(t) = t^2(1-t)^2$ which satisfies the constraints of having value zero and having first derivative zero at both 0 and 1.

Note the similarity in the outputs corresponding to RKTSB and NLS for large n because of asymptotic efficiency while TSB intervals are much wider. The asymptotic variance for NLS is derived through the delta method which usually underestimates the true variance for small sample sizes. This explains the low coverage of the NLS confidence intervals for $n = 50$. However, TSB is computationally much faster. Each replication in TSB took around 30 seconds. In contrast, each replication in RKTSB took around 60 seconds and each replication in NLS took around 2 seconds. But in the latter case the runtime is not comparable to those of the Bayesian methods since here we just obtain an estimate of θ in each replication whereas in Bayesian methods we need to draw an entire posterior sample at each replication. We also plot the densities

Table 1: *Coverages and average lengths of the Bayesian credible intervals and confidence intervals for Gaussian error*

n		RKTSB		TSB		NLS	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
50	θ	93.2 (0.04)	0.55 (0.32)	94.6 (0.03)	4.4 (1.98)	87.2 (0.05)	0.58 (0.52)
100	θ	95.1 (0.02)	0.34 (0.05)	95.8 (0.02)	2.55 (1.11)	95.1 (0.02)	0.32 (0.03)
500	θ	95.5 (0.01)	0.14 (0.01)	97.0 (0.01)	0.85 (0.17)	95.1 (0.01)	0.14 (0.01)

corresponding to one posterior sample of θ obtained from each Bayesian method in Figures 2 and 3, respectively, for the Gaussian error.

6. Proofs

260 We use the operators E_0 and var_0 to denote expectation and variance with respect to P_0.

Proof (Lemma 1). By the definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ we have

$$\int_0^1 \boldsymbol{G}\{t, \boldsymbol{h}(t), \boldsymbol{\theta}\} w(t) dt = \mathbf{0}, \quad \int_0^1 \boldsymbol{G}\{t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0\} w(t) dt = \mathbf{0}.$$

Subtracting the second equation from the first and applying the mean-value theorem, we get

$$\begin{aligned}\int_0^1 & D_{\boldsymbol{\theta}_0}\{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\} w(t) dt (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ & + \int_0^1 D_{\boldsymbol{h}_0}\{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\} w(t) \{\boldsymbol{h}(t) - \boldsymbol{h}_0(t)\} dt \\ & \qquad + O\left(\sup_{t\in[0,1]} \|\boldsymbol{h}(t) - \boldsymbol{h}_0(t)\|^2\right) + O(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2) = \mathbf{0}.\end{aligned}$$

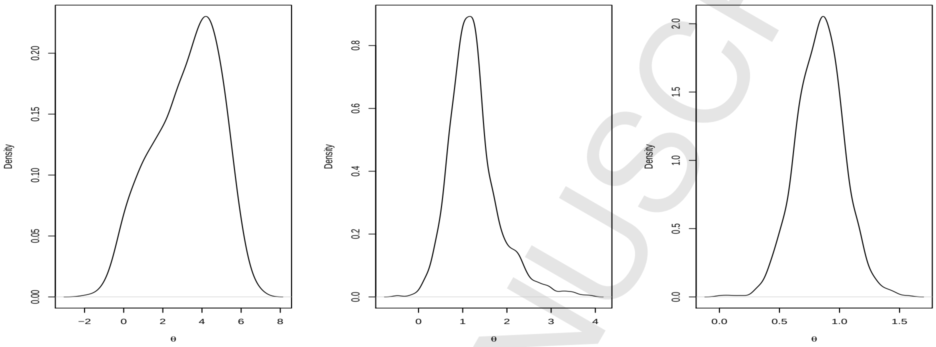

Figure 2: *Density curve corresponding to posterior sample of θ obtained by TSB method for $n = 50$, $n = 100$ and $n = 500$, respectively, for Gaussian error*

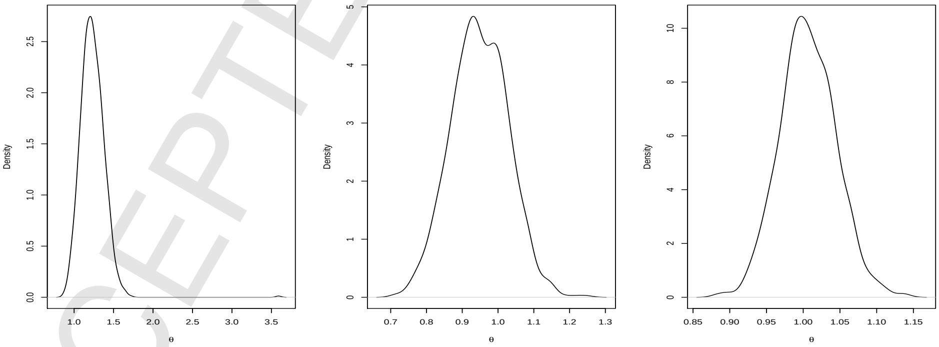

Figure 3: *Density curve corresponding to posterior sample of θ obtained by RKTSB for $n = 50$, $n = 100$ and $n = 500$, respectively, for Gaussian error*

Table 2: *Coverages and average lengths of the Bayesian credible intervals and confidence intervals for scaled t_6 error*

n		RKTSB		TSB		NLS	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
50	θ	92.9 (0.04)	0.54 (0.25)	92.9 (0.04)	4.41 (2.02)	86.1 (0.05)	0.58 (0.52)
100	θ	94.1 (0.02)	0.34 (0.05)	95.4 (0.02)	2.53 (1.14)	93.5 (0.02)	0.32 (0.05)
500	θ	94.9 (0.01)	0.14 (0.01)	97.4 (0.01)	0.85 (0.18)	94.7 (0.01)	0.14 (0.01)

Now we shall show that the second summand is a linear functional of $f - f_0$. Note that this term can be written as

$$\sum_{r=0}^{q} \int_0^1 \left[D_{\boldsymbol{h}_0}\{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\}w(t)\right]_{,r} \{f^{(r)}(t, \boldsymbol{\beta}) - f_0^{(r)}(t)\}dt.$$

We shall show that every term of this sum is a linear functional of $f - f_0$. We observe that for each $r \in \{0, \dots, q\}$,

$$\begin{aligned}\int_0^1 \left[D_{\boldsymbol{h}_0}\{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\}w(t)\right]_{,r} \{f^{(r)}(t, \boldsymbol{\beta}) - f_0^{(r)}(t)\}dt \\ = (-1)^r \int_0^1 \frac{d^r}{dt^r} \left[D_{\boldsymbol{h}_0}\{\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\}w(t)\right]_{,r} \{f(t, \boldsymbol{\beta}) - f_0(t)\}dt\end{aligned}$$

using integration by parts and the fact that the function w and its first $q-1$ derivatives vanish at 0 and 1. Proceeding this way we get

$$\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \boldsymbol{\Gamma}(f - f_0) + O\left(\sup_{t\in[0,1]} \|\boldsymbol{h}(t) - \boldsymbol{h}_0(t)\|^2\right) + O(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2) = \mathbf{0}.$$

Using the steps of the proofs of Lemmas 2 and 3 of Bhaumik and Ghosal [4], we can prove the posterior consistency of $\boldsymbol{\theta}$. Let $\epsilon_n > 0$ and define $E_n = \{(\boldsymbol{h}, \boldsymbol{\theta}) : \sup_{t\in[0,1]} \|\boldsymbol{h}(t) - \boldsymbol{h}_0(t)\| \leq \epsilon_n, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \epsilon_n\}$. Hence for some sequence $\epsilon_n \to 0$,

Table 3: *Coverages and average lengths of the Bayesian credible intervals and confidence intervals for centered and scaled Gamma(1,1) error*

n		RKTSB		TSB		NLS	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
50	θ	92.5 (0.04)	0.55 (0.29)	92.5 (0.04)	4.36 (2.00)	86.3 (0.05)	0.6 (0.56)
100	θ	95.4 (0.02)	0.34 (0.05)	95.4 (0.02)	2.54 (1.13)	94.1 (0.02)	0.32 (0.07)
500	θ	94.0 (0.01)	0.14 (0.01)	96.0 (0.01)	0.85 (0.19)	93.4 (0.01)	0.14 (0.01)

$\Pi(E_n^c|\boldsymbol{t},\boldsymbol{Y}) = o_{P_0}(1)$ and on E_n

$$\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = [\{\boldsymbol{M}(\boldsymbol{h}_0,\boldsymbol{\theta}_0)\}^{-1}+o(1)]\sqrt{n}\boldsymbol{\Gamma}(f-f_0) + \sqrt{n}\sup_{t\in[0,1]}\left\|\boldsymbol{h}(t)-\boldsymbol{h}_0(t)\right\|^2 O(1).$$

As in Lemma 4 of Bhaumik and Ghosal [4], $\sqrt{n}\,\boldsymbol{\Gamma}(f-f_0)$ assigns most of its mass inside a large compact set. Now by proceeding as in Lemma 2 of Bhaumik and Ghosal [4], we can assert that on E_n, the second term on the display is $o(1)$ and the conclusion follows. □

Proof (Theorem 1). By Lemma 1 and (10), it suffices to show that for any σ^2 in a neighborhood of σ_0^2,

$$\left\|\Pi_n^*[\sqrt{n}\boldsymbol{H}_n^\top\boldsymbol{\beta}-\sqrt{n}\{\boldsymbol{M}(\boldsymbol{h}_0,\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Gamma}(f_0)\in\cdot|\boldsymbol{t},\boldsymbol{Y},\sigma^2]-\mathcal{N}(\boldsymbol{\mu}_n^*,\sigma^2\boldsymbol{\Sigma}_n^*)\right\|_{TV}$$

is $o_{P_0}(1)$. Note that the conditional posterior distribution of $\boldsymbol{H}_n^\top\boldsymbol{\beta}$ is a normal distribution with mean vector

$$\boldsymbol{H}_n^\top\left(\boldsymbol{X}_n^\top\boldsymbol{X}_n+n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{X}_n^\top\boldsymbol{Y}$$

and dispersion matrix

$$\sigma^2\boldsymbol{H}_n^\top\left(\boldsymbol{X}_n^\top\boldsymbol{X}_n+n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{H}_n,$$

respectively. We compute the Kullback–Leibler divergence between two Gaussian distributions and show that it converges in P_0-probability to 0 to prove the assertion. The rest of the proof is similar to that of Theorem 3 of Bhaumik and Ghosal [4]. □

Proof (Theorem 2). Note that $\int_0^1 \boldsymbol{C}(t)\boldsymbol{\beta}^\top \boldsymbol{N}(t)g(t)dt = \boldsymbol{G}_n^\top \boldsymbol{\beta}$, where

$$\boldsymbol{G}_n^\top = \int_0^1 \boldsymbol{C}(t)\boldsymbol{N}^\top(t)g(t)dt$$

which is a matrix of order $p \times (k_n + m - 1)$. We can derive the posterior consistency of σ^2 as in Lemma 11 of Bhaumik and Ghosal [5]. Proceeding as in Lemma 9 of Bhaumik and Ghosal [5], it can be shown that on a set with high posterior probability

$$\left\| \sqrt{n}\,(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \left[\sqrt{n}\boldsymbol{G}_n^\top \boldsymbol{\beta} - \sqrt{n}\{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1} \int_0^1 \{\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\}^\top f_0(t)g(t) \right] \right\| \to 0$$

as $n \to \infty$. Then it suffices to show that for any neighborhood $\mathscr{N}$ of σ_0^2, uniformly in $\sigma^2 \in \mathscr{N}$, the total variation distance between

$$\Pi_n^{**} \left[\sqrt{n}\,\boldsymbol{G}_n^\top \boldsymbol{\beta} - \sqrt{n}\{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1} \int_0^1 \{\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\}^\top f_0(t)g(t) \in \cdot | \boldsymbol{t}, \boldsymbol{Y}, \sigma^2 \right]$$

and $\mathcal{N}(\boldsymbol{\mu}_n^{**}, \sigma^2 \boldsymbol{\Sigma}_n^{**})$ is $o_{P_0}(1)$. The rest of the proof follows from that of Theorem 4.2 of Bhaumik and Ghosal [5]. □

Proof (Corollary 1). The log-likelihood of the correctly specified model is given by

$$\ell_{\boldsymbol{\theta}_0}(t, Y) \quad = \quad -\ln \sigma_0 - \frac{1}{2\sigma_0^2}|Y - f_{\boldsymbol{\theta}_0}(t)|^2 + \ln g(t).$$

Thus $\dot{\ell}_{\boldsymbol{\theta}_0}(t, Y) = -\sigma_0^{-2}\{\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\}^\top \{Y - f_{\boldsymbol{\theta}_0}(t)\}$ and the Fisher information is given by $\boldsymbol{I}(\boldsymbol{\theta}_0) = \sigma_0^{-2} \int_0^1 \{\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\}^\top \dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)g(t)dt$. Following the proof of Lemma 10 of Bhaumik and Ghosal [5] we get

$$\sigma_0^2 \boldsymbol{\Sigma}_n^{**} \xrightarrow{P_0} \sigma_0^2 \{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1} \int_0^1 \{\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\}^\top \dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)g(t)dt\,[\{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1}]^\top.$$

This limit is equal to $\{\boldsymbol{I}(\boldsymbol{\theta}_0)\}^{-1}$ if the regression function and the error distribution are correctly specified. □

Appendix: Runge–Kutta method for higher order ODE

Often the differential equation has the form

$$\begin{aligned} F\left(t, f_{\boldsymbol{\theta}}(t), f_{\boldsymbol{\theta}}^{(1)}(t), \ldots, f_{\boldsymbol{\theta}}^{(q)}(t), \boldsymbol{\theta}\right) & \\ = f_{\boldsymbol{\theta}}^{(q)}(t) - H\left(t, f_{\boldsymbol{\theta}}(t), f_{\boldsymbol{\theta}}^{(1)}(t), \ldots, f_{\boldsymbol{\theta}}^{(q-1)}(t), \boldsymbol{\theta}\right) = 0 & \end{aligned}$$

with initial conditions $f_{\boldsymbol{\theta}}^{(\nu)}(0) = c_\nu$ for all $\nu \in \{0, \ldots, q-1\}$, and H is a known function. Note that t can be treated as a state variable $\chi(t) = t$ which satisfies the qth order ODE $\chi^{(q)}(t) = 0$ with initial conditions $\chi(0) = 0$, $\chi^{(1)}(0) = 1$ and $\chi^{(j)}(0) = 0$ for all $j \in \{2, \ldots, q-1\}$. Denoting $\boldsymbol{\psi}_{\boldsymbol{\theta}}(\cdot) = (f_{\boldsymbol{\theta}}(\cdot), \chi(\cdot))$, we can rewrite the ODE as

$$\boldsymbol{\psi}_{\boldsymbol{\theta}}^{(q)}(t) = \boldsymbol{H}\{\boldsymbol{\psi}_{\boldsymbol{\theta}}(t), \ldots, \boldsymbol{\psi}_{\boldsymbol{\theta}}^{(q-1)}(t)\},$$

where $\boldsymbol{H} = (H(\cdot), 0)$. Given r_n equispaced grid points $a_1 = 0, a_2, \ldots, a_{r_n}$ with common difference r_n^{-1}, the approximate solution to (1) is given by $\boldsymbol{\psi}_{\theta, r_n}(\cdot) = (f_{\boldsymbol{\theta}, r_n}(\cdot), \chi_{r_n}(\cdot))$, where r_n is chosen so that $r_n \gg n^{1/8}$; here n denotes the number of observations. Let $\boldsymbol{z}_k = (\boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}(a_k), \boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}^{(1)}(a_k), \ldots, \boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}^{(q-1)}(a_k))$ stand for the vector formed by the function $\psi_{\boldsymbol{\theta}, r_n}$ and its $q-1$ derivatives at the kth grid point a_k for all $k \in \{1, \ldots, r_n\}$. For each $\nu \in \{1, \ldots, q-1\}$, we define

$$\begin{aligned} \boldsymbol{T}^{\nu}(a_k, \boldsymbol{z}_k, r_n) &= \boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}^{(\nu)}(a_k) + \frac{1}{2! r_n} \boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}^{(\nu+1)}(a_k) + \cdots \\ &\quad + \frac{1}{r_n^{(q-\nu-1)}(q-\nu)!} \boldsymbol{\psi}_{\boldsymbol{\theta}, r_n}^{(q-1)}(a_k), \\ \boldsymbol{T}^{q}(a_k, \boldsymbol{z}_k, r_n) &= \mathbf{0}. \end{aligned}$$

Now let $\boldsymbol{k}_\rho(a_k) = \boldsymbol{H}(\boldsymbol{U}^1, \ldots, \boldsymbol{U}^q)$ with $\boldsymbol{U}^1, \ldots, \boldsymbol{U}^q$ as given in Table 4.

“Table 4 here”

Following Eq. (4.16) of Henrici [7, p. 169] we define

$$\boldsymbol{\Phi}^{\nu}(a_k, \boldsymbol{z}_k, r_n) = \boldsymbol{T}^{\nu}(a_k, \boldsymbol{z}_k, r_n) + \frac{1}{r_n^{(q-\nu)}(q-\nu+1)!} \sum_{\rho=1}^{4} \gamma_{\nu\rho} \boldsymbol{k}_\rho(a_k),$$

where the coefficients $\gamma_{\nu\rho}$ are given by

$$\begin{aligned}\gamma_{\nu 1} &= \frac{(q-\nu+1)^2}{(q-\nu+2)(q-\nu+3)},\\ \gamma_{\nu 2} = \gamma_{\nu 3} &= \frac{2(q-\nu+1)}{(q-\nu+2)(q-\nu+3)},\\ \gamma_{\nu 4} &= \frac{1-q+\nu}{(q-\nu+2)(q-\nu+3)}\end{aligned}$$

for all $\nu \in \{1,\dots,q\}$. Then the sequence $z_1,\dots,z_{r_n}$ can be constructed by the recurrence relation

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + r_n^{-1}\left(\boldsymbol{\Phi}^1(a_k,\boldsymbol{z}_k,r_n),\dots,\boldsymbol{\Phi}^q(a_k,\boldsymbol{z}_k,r_n)\right)^\top.$$

By the proof of Theorem 4.2 of Henrici [7, p. 174], we have

$$\sup_{t\in[0,1]} |f_{\boldsymbol{\theta}}(t) - f_{\boldsymbol{\theta},r_n}(t)| = O(r_n^{-4}), \sup_{t\in[0,1]} |\dot{f}_{\boldsymbol{\theta}}(t) - \dot{f}_{\boldsymbol{\theta},r_n}(t)| = O(r_n^{-4}).$$

References

[1] A.R. Bergstrom, Gaussian estimation of structural parameters in higher order continuous time dynamic models, Econometrica 51 (1983) 117–152.

[2] A.R. Bergstrom, The estimation of parameters in nonstationary higher order continuous-time dynamic models, Econometric Theory 1 (1985) 369–385.

[3] A.R. Bergstrom, The estimation of open higher-order continuous time dynamic models with mixed stock and flow data, Econometric Theory 2 (1986) 350–373.

[4] P. Bhaumik, S. Ghosal, Bayesian two-step estimation in differential equation models, Electronic J. Statist. 2 (2015) 3124–3154.

[5] P. Bhaumik, S. Ghosal, Efficient Bayesian estimation and uncertainty quantification in ordinary differential equation models, Bernoulli (2016).

[6] C. De Boor, A Practical Guide to Splines, Springer, New York, 1978.

[7] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York, 1962.

[8] S. Karlin, H.M. Taylor, A Second Course in Stochastic Processes, Academic Press, New York, 1981.

[9] X. Qi, H. Zhao, Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations, Ann. Statist. 38 (2010) 435–481.

[10] H. Xue, H. Miao, H. Wu, Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error, Ann. Statist. 38 (2010) 2351–2387.

Table 4: *Arguments of* $\boldsymbol{H}$

ρ	1	2	3	4
$\boldsymbol{U}^1$	$\boldsymbol{\psi}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^2}\boldsymbol{\psi}^{(2)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n^2}\boldsymbol{\psi}^{(2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^3}\boldsymbol{\psi}^{(3)}_{\boldsymbol{\theta},r_n}(a_k)$
$\boldsymbol{U}^2$	$\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{\psi}^{(2)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{\psi}^{(2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^2}\boldsymbol{\psi}^{(3)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(1)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{\psi}^{(2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n^2}\boldsymbol{\psi}^{(3)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^3}\boldsymbol{\psi}^{(4)}_{\boldsymbol{\theta},r_n}(a_k)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{U}^{q-3}$	$\boldsymbol{\psi}^{(q-4)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-4)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-4)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^2}\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-4)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n^2}\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^3}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$
$\boldsymbol{U}^{q-2}$	$\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^2}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-3)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n^2}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{4r_n^3}\boldsymbol{k}_1$
$\boldsymbol{U}^{q-1}$	$\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{4r_n^2}\boldsymbol{k}_1$	$\boldsymbol{\psi}^{(q-2)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n^2}\boldsymbol{k}_2$
$\boldsymbol{U}^{q}$	$\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$	$\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)$ $+\frac{1}{2r_n}\boldsymbol{k}_1$	$\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{2r_n}\boldsymbol{k}_2$	$\boldsymbol{\psi}^{(q-1)}_{\boldsymbol{\theta},r_n}(a_k)+\frac{1}{r_n}\boldsymbol{k}_3$