

Bayesian two-step estimation in differential equation models

Prithwish Bhaumik

*Department of Statistics and Data Sciences
The University of Texas at Austin
2317 Speedway D9800
Austin, TX 78712-1823
e-mail: prithwish.bhaumik@utexas.edu*

and

Subhashis Ghosal

*Department of Statistics
North Carolina State University
4276 SAS Hall, 2311 Stinson Drive
Raleigh, NC 27695-8203
e-mail: sghosal@ncsu.edu*

Abstract: Ordinary differential equations (ODEs) are used to model dynamic systems appearing in engineering, physics, biomedical sciences and many other fields. These equations contain an unknown vector of parameters of physical significance, say θ which has to be estimated from the noisy data. Often there is no closed form analytic solution of the equations and hence we cannot use the usual non-linear least squares technique to estimate the unknown parameters. The two-step approach to solve this problem involves fitting the data nonparametrically and then estimating the parameter by minimizing the distance between the nonparametrically estimated derivative and the derivative suggested by the system of ODEs. The statistical aspects of this approach have been studied under the frequentist framework. We consider this two-step estimation under the Bayesian framework. The response variable is allowed to be multidimensional and the true mean function of it is not assumed to be in the model. We induce a prior on the regression function using a random series based on the B-spline basis functions. We establish the Bernstein-von Mises theorem for the posterior distribution of the parameter of interest. Interestingly, even though the posterior distribution of the regression function based on splines converges at a rate slower than $n^{-1/2}$, the parameter vector θ is nevertheless estimated at $n^{-1/2}$ rate.

MSC 2010 subject classifications: 62J02, 62G08, 62G20, 62F15.

Keywords and phrases: Ordinary differential equation, Bayesian inference, spline smoothing, Bernstein-von Mises theorem.

Received August 2014.

Contents

1	Introduction	3125
2	Notations, model assumption and prior specification	3128

3 Main results 3131
 4 Extensions 3133
 5 Simulation study 3134
 6 Real life data 3136
 7 Proofs 3137
 Appendix 3146
 References 3152

1. Introduction

Suppose that we have a regression model $\mathbf{Y} = \mathbf{f}_\theta(t) + \varepsilon$, $\theta \in \Theta \subseteq \mathbb{R}^p$. The explicit form of $\mathbf{f}_\theta(\cdot)$ may not be known, but the function is assumed to satisfy the system of ordinary differential equations (ODEs) given by

$$\frac{d\mathbf{f}_\theta(t)}{dt} = \mathbf{F}(t, \mathbf{f}_\theta(t), \theta), t \in [0, 1]; \tag{1.1}$$

here \mathbf{F} is a known appropriately smooth vector-valued function and θ is a parameter vector controlling the regression function. Equations of this type are encountered in various branches of science such as in genetics (Chen et al., 1999), viral dynamics of infectious diseases (Anderson and May (1992), Nowak and May (2000)). There are numerous applications in the fields of pharmacokinetics and pharmacodynamics (PKPD) as well. There are a lot of instances where no closed form solution exist. Such an example can be found in the feedback system (Gabrielsson and Weiner, 2006, page 332) modeled by the ODEs

$$\begin{aligned} \frac{dR(t)}{dt} &= \frac{k_{in}}{M(t)} - k_{out}R(t), \\ \frac{dM(t)}{dt} &= k_{tol}(R(t) - M(t)), \end{aligned}$$

where $R(t)$ and $M(t)$ stand for loss of response and modulator at time t respectively. Here k_{in}, k_{out} and k_{tol} are unknown parameters which have to be estimated from the noisy observations given by

$$\begin{aligned} Y_R(t) &= R(t) + \varepsilon_R(t), \\ Y_M(t) &= M(t) + \varepsilon_M(t), \end{aligned}$$

$\varepsilon_R(t), \varepsilon_M(t)$ being the respective noises at time point t . Another popular example is the Lotka-Volterra equations, also known as predator-prey equations. The prey and predator populations change over time according to the equations

$$\begin{aligned} \frac{df_{1\theta}(t)}{dt} &= \theta_1 f_{1\theta}(t) - \theta_2 f_{1\theta}(t) f_{2\theta}(t), \\ \frac{df_{2\theta}(t)}{dt} &= -\theta_3 f_{2\theta}(t) + \theta_4 f_{1\theta}(t) f_{2\theta}(t), \end{aligned}$$

where $t \in [0, 1]$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ and $f_{1\boldsymbol{\theta}}(t)$ and $f_{2\boldsymbol{\theta}}(t)$ denote the prey and predator populations at time t respectively.

Another interesting model specification through differential equations appear in network analysis. Henderson and Michailidis (2014) considered an interesting situation where $\mathbf{F}(\cdot)$ is an unknown function of the solution of the ODE. They assumed an additive model for \mathbf{F} and used nonparametric techniques to fit \mathbf{F} . In this situation unlike the examples above the resulting model is not described by a finite set of parameters.

If the ODEs can be solved analytically, then the usual non linear least squares (NLS) (Levenberg (1944), Marquardt (1963)) can be used to estimate the unknown parameters. In many of the practical situations, such closed form solutions are not available as evidenced in the previous two examples. NLS was modified for this purpose by Bard (1974) and Domselaar and Hemker (1975). Hairer et al. (1993, page 134) and Mattheij and Molenaar (2002, page 53) used the 4-stage Runge-Kutta algorithm to solve (1.1) numerically. The NLS can be applied in the next step to estimate the parameters. The statistical properties of the corresponding estimator have been studied by Xue et al. (2010). The strong consistency, \sqrt{n} -consistency and asymptotic normality of the estimator were established in their work.

Ramsay et al. (2007) proposed the generalized profiling procedure where the solution is approximated by a linear combination of basis functions. The coefficients of the basis functions are estimated by solving a penalized optimization problem using an initial choice of the parameters of interest. A data-dependent fitting criterion is constructed which contains the estimated coefficients. Then $\boldsymbol{\theta}$ is estimated by the maximizer of this criterion. Qi and Zhao (2010) explored the statistical properties of this estimator including \sqrt{n} -consistency and asymptotic normality. Despite having desirable statistical properties, these approaches are computationally cumbersome especially for high-dimensional systems of ODEs as well as when $\boldsymbol{\theta}$ is high-dimensional.

Varah (1982) used a two-step method for estimating $\boldsymbol{\theta}$. In the first step each of the state variables is approximated by a cubic spline using the least squares technique. In the second step, the corresponding derivatives are estimated by differentiating the nonparametrically fitted curve and the estimator is obtained by minimizing the sum of squares of the difference between the derivatives of the fitted spline and the derivatives suggested by the ODEs at the design points. This method does not depend on the initial or boundary conditions of the state variables and is computationally very efficient irrespective of the complexity of the model. An example given in Voit and Almeida (2004) showed the computational superiority of the two-step approach over the usual least squares technique. Brunel (2008) replaced the sum of squares of the second step by a weighted integral of the squared deviation and proved \sqrt{n} -consistency as well as asymptotic normality of the resulting estimator. The order of the B-spline basis was determined by the smoothness of $\mathbf{F}(\cdot, \cdot, \cdot)$ with respect to its first two arguments. Gugushvili and Klaassen (2012) used the same approach but used kernel smoothing instead of spline. They also established \sqrt{n} -consistency of the estimator. Another modification has been made in the work of Wu et al. (2012). They

have used penalized smoothing spline in the first step and numerical derivatives instead of actual derivatives of the nonparametrically estimated functions. In another work Brunel et al. (2014) used nonparametric approximation of the true solution to (1.1) and then used a set of orthogonality conditions to estimate the parameters. The \sqrt{n} -consistency as well as the asymptotic normality of the estimator was also established in their work. In Dattner and Gugushvili (2015) the two-step estimator is used as a preliminary \sqrt{n} -consistent estimator and then Newton-Raphson technique is employed to obtain asymptotically efficient estimator. Although this approach is faster than the NLS technique, a numerical solution of the ODE is still required at the Newton-Raphson step.

In ODE models Bayesian estimation was considered in the works of Gelman et al. (1996), Rogers et al. (2007) and Girolami (2008). First they solved the ODEs numerically to approximate the expected response and hence constructed the likelihood. A prior was assigned on θ and MCMC technique was used to generate samples from the posterior. Computation cost might be an issue in this case as well. Campbell and Steele (2012) proposed the smooth functional tempering approach which is a population MCMC technique and it utilizes the generalized profiling approach (Ramsay et al., 2007) and the parallel tempering algorithm. Campbell (2007) and Jaeger (2012) also used Bayesian analog of the generalized profiling by putting prior on the coefficients of the basis functions. Chkrebtii et al. (2013) divided the time range into discrete grid points. They put a Gaussian process prior on the solution of the ODE and its derivative. The posterior distribution of the solution is used to draw the posterior sample of the parameter of interest. The theoretical aspects of Bayesian estimation methods have not been yet explored in the literature.

In this paper we consider a Bayesian analog of the approach of Brunel (2008) fitting a nonparametric regression model using B-spline basis. We assign priors on the coefficients of the basis functions. A posterior is then induced on θ using the posteriors of the coefficients of the basis functions. In this paper we study the asymptotic properties of the posterior distribution of θ and establish a Bernstein-von Mises theorem with $n^{-1/2}$ contraction rate. We allow the ODE model to be misspecified, that is, the true regression function may not be a solution of the ODE. The response variable is also allowed to be multidimensional with possibly correlated errors. Normal distribution is used as the working model for error distribution, but the true distribution of errors may be different. Interestingly, the original model is parametric but it is embedded in a nonparametric model, which is further approximated by high dimensional parametric models. Note that the slower rate of nonparametric estimation does not influence the convergence rate of the parameter in the original parametric model.

In the context of misspecification it is worthy to mention approximate Bayesian computation (ABC) which can be viewed as Bayesian inference using a misspecified likelihood for which a Bernstein-von Mises theorem with biased center and different scaling may hold (Dean and Singh, 2011). Thus our misspecified Bernstein-von Mises theorem has some formal similarity with results of this kind. See Kleijn and van der Vaart (2012) for a general approach to misspecified Bernstein-von Mises theorem.

The paper is organized as follows. Section 2 contains the description of the notations and the model as well as the priors used for the analysis. The main results are given in Section 3. We extend the results to more generalized setups in Section 4. In Section 5 we carry out a simulation study under different settings. We analyze a real life data in Section 6. Proofs of the theorems are given in Section 7. Appendix contains the proofs of some auxiliary lemmas.

2. Notations, model assumption and prior specification

We describe a set of notations to be used in this paper. Boldfaced letters are used to denote vectors and matrices. For a matrix \mathbf{A} , the symbols $\mathbf{A}_{i,\cdot}$ and $\mathbf{A}_{\cdot,j}$ stand for the i^{th} row and j^{th} column of \mathbf{A} respectively. The notation $((A_{i,j}))$ stands for a matrix with $(i,j)^{\text{th}}$ element being $A_{i,j}$. We use the notation $\text{rows}_r^s(\mathbf{A})$ with $r < s$ to denote the sub-matrix of \mathbf{A} consisting of r^{th} to s^{th} rows of \mathbf{A} . Similarly, we can define $\text{cols}_r^s(\mathbf{A})$ for columns. The notation $\mathbf{x}_{r:s}$ stands for the sub-vector consisting of r^{th} to s^{th} elements of a vector \mathbf{x} . By $\text{vec}(\mathbf{A})$, we mean the vector obtained by stacking the columns of the matrix \mathbf{A} one over another. For an $m \times n$ matrix \mathbf{A} and a $p \times q$ matrix \mathbf{B} , $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product between \mathbf{A} and \mathbf{B} ; see Steeb (2006) for the definition. The identity matrix of order p is denoted by \mathbf{I}_p . By the symbols $\text{maxeig}(\mathbf{A})$ and $\text{mineig}(\mathbf{A})$, we denote the maximum and minimum eigenvalues of the matrix \mathbf{A} respectively. For a vector $\mathbf{x} \in \mathbb{R}^p$, we denote $\|\mathbf{x}\| = (\sum_{i=1}^p x_i^2)^{1/2}$. We denote the r^{th} order derivative of a function $f(\cdot)$ by $f^{(r)}(\cdot)$, that is, $f^{(r)}(t) = \frac{d^r}{dt^r} f(t)$. The boldfaced symbol $\mathbf{f}(\cdot)$ stands for a vector valued function. For functions $\mathbf{f} : [0, 1] \rightarrow \mathbb{R}^p$ and $w : [0, 1] \rightarrow [0, \infty)$, we define $\|\mathbf{f}\|_w = (\int_0^1 \|\mathbf{f}(t)\|^2 w(t) dt)^{1/2}$. For a real-valued function $f : [0, 1] \rightarrow \mathbb{R}$ and a vector $\mathbf{x} \in \mathbb{R}^p$, we denote $f(\mathbf{x}) = (f(x_1), \dots, f(x_p))^T$. The notation $\langle \cdot, \cdot \rangle$ stands for an inner product. For numerical sequences a_n and b_n , by $a_n = o(b_n)$, we mean $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$. The notation $a_n = O(b_n)$ implies that a_n/b_n is bounded. We use the notation $a_n \asymp b_n$ to mean $a_n = O(b_n)$ and $b_n = O(a_n)$, while $a_n \lesssim b_n$ stands for $a_n = O(b_n)$. The symbol $a_n \gg b_n$ will mean $b_n = o(a_n)$. Similarly we can define $a_n \ll b_n$. The notation $o_P(1)$ is used to indicate a sequence of random variables which converges in probability to zero, whereas the expression $O_P(1)$ stands for a sequence of random variables bounded in probability. The boldfaced symbols $\mathbf{E}(\cdot)$ and $\mathbf{Var}(\cdot)$ stand for the mean vector and dispersion matrix respectively of a random vector. For the probability measures P and Q defined on \mathbb{R}^p , we define the total variation distance $\|P - Q\|_{TV} = \sup_{B \in \mathcal{B}^p} |P(B) - Q(B)|$, where \mathcal{B}^p denotes the Borel σ -field on \mathbb{R}^p . For an open set E , the symbol $C^m(E)$ stands for the collection of functions defined on E with first m continuous partial derivatives with respect to its arguments. Now we present the formal description of the model.

We have a system of d ordinary differential equations given by

$$\frac{df_{j\theta}(t)}{dt} = F_j(t, \mathbf{f}\theta(t), \boldsymbol{\theta}), t \in [0, 1], j = 1, \dots, d, \quad (2.1)$$

where $\mathbf{f}_\theta(\cdot) = (f_{1\theta}(\cdot), \dots, f_{d\theta}(\cdot))^T$ and $\theta \in \Theta$, a compact subset of \mathbb{R}^p . Let us denote $\mathbf{F}(\cdot, \cdot, \cdot) = (F_1(\cdot, \cdot, \cdot), \dots, F_d(\cdot, \cdot, \cdot))^T$. We also assume that for a fixed θ , $\mathbf{F} \in C^{m-1}((0, 1), \mathbb{R}^d)$ for some integer $m \geq 1$. Then, by successive differentiation of the right hand side of (2.1), it follows that $\mathbf{f}_\theta \in C^m((0, 1))$. By the implied uniform continuity, the function and its several derivatives uniquely extend to continuous functions on $[0, 1]$.

Consider an $n \times d$ matrix of observations \mathbf{Y} with $Y_{i,j}$ denoting the measurement taken on the j^{th} response at the point x_i , $0 \leq x_i \leq 1$, $i = 1, \dots, n$; $j = 1, \dots, d$. We consider x_i 's to be deterministic covariates satisfying condition (2.6) below. If the covariates are random and sampled independently from a fixed continuous and positive density and $k_n \ll \sqrt{n}$, then the condition holds with probability tending to one since in view of Donsker's theorem the left hand size of (2.6) is $O_P(n^{-1/2}) = o_P(k_n^{-1})$. Therefore the results of this paper will also hold for random covariates. Denoting $\varepsilon = ((\varepsilon_{i,j}))$ as the corresponding matrix of errors, the proposed model is given by

$$Y_{i,j} = f_{j\theta}(x_i) + \varepsilon_{i,j}, \quad i = 1, \dots, n, \quad j = 1, \dots, d, \tag{2.2}$$

while the data is generated by the model

$$Y_{i,j} = f_{j0}(x_i) + \varepsilon_{i,j}, \quad i = 1, \dots, n, \quad j = 1, \dots, d, \tag{2.3}$$

where $\mathbf{f}_0(\cdot) = (f_{10}(\cdot), \dots, f_{d0}(\cdot))^T$ denotes the true mean vector which does not necessarily lie in $\{\mathbf{f}_\theta : \theta \in \Theta\}$. We assume that $\mathbf{f}_0 \in C^m([0, 1])$. Let $\varepsilon_{i,j} \stackrel{iid}{\sim} P_0$, which is a probability distribution with mean zero and finite variance σ_0^2 for $i = 1, \dots, n$; $j = 1, \dots, d$.

Since the expression of \mathbf{f}_θ is usually not available, the proposed model is embedded in nonparametric regression model

$$\mathbf{Y} = \mathbf{X}_n \mathbf{B}_n + \varepsilon, \tag{2.4}$$

where $\mathbf{X}_n = ((N_j(x_i)))_{1 \leq i \leq n, 1 \leq j \leq k_n+m-1}$, $\{N_j(\cdot)\}_{j=1}^{k_n+m-1}$ being the B-spline basis functions of order m with $k_n - 1$ interior knots $0 < \xi_1 < \xi_2 < \dots < \xi_{k_n-1} < 1$ chosen to satisfy the pseudo-uniformity criteria:

$$\begin{aligned} \max_{1 \leq i \leq k_n-1} |\xi_{i+1} - 2\xi_i + \xi_{i-1}| &= o(k_n^{-1}), \\ \max_{1 \leq i \leq k_n-1} |\xi_i - \xi_{i-1}| / \min_{1 \leq i \leq k_n-1} |\xi_i - \xi_{i-1}| &\leq M \end{aligned} \tag{2.5}$$

for some constant $M > 0$. Here ξ_0 and ξ_{k_n} are defined as 0 and 1 respectively. The criteria (2.5) is required to apply the asymptotic results obtained in Zhou et al. (1998) where they mention the similar criteria in equation (3) of that paper. Here we denote

$$\mathbf{B}_n = \left(\beta_1^{(k_n+m-1) \times 1}, \dots, \beta_d^{(k_n+m-1) \times 1} \right),$$

the matrix containing the coefficients of the basis functions. Also we consider P_0 to be unknown and use $N(0, \sigma^2)$ as the working distribution for the error where

σ may be treated as another unknown parameter. Denoting by Q_n , the empirical distribution function of x_i , $i = 1, \dots, n$, we assume that for some probability measure Q on $[0, 1]$ with positive and continuous density

$$\sup_{t \in [0, 1]} |Q_n(t) - Q(t)| = o(k_n^{-1}). \quad (2.6)$$

Let the prior distribution on the coefficients be given by

$$\beta_j \stackrel{iid}{\sim} N_{k_n+m-1}(\mathbf{0}, c^{-1}nk_n^{-1}(\mathbf{X}_n^T \mathbf{X}_n)^{-1}) \quad (2.7)$$

for some constant $c > 0$. Simple calculation yields the posterior distribution for β_j as

$$\beta_j | \mathbf{Y} \sim N_{k_n+m-1} \left(c_n^{-1} (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_{\cdot j}, c_n^{-1} \sigma^2 (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \right) \quad (2.8)$$

and the posterior distributions of β_j and $\beta_{j'}$ are mutually independent for $j \neq j'$; $j, j' = 1, \dots, d$, where $c_n = (1 + \sigma^2 ck_n/n)$. In the model (2.4), the expected response vector at a point $t \in [0, 1]$ is given by $\mathbf{B}_n^T \mathbf{N}(t)$, where $\mathbf{N}(\cdot) = (N_1(\cdot), \dots, N_{k_n+m-1}(\cdot))^T$.

Let $w(\cdot)$ be a continuous weight function with $w(0) = w(1) = 0$ and be positive on $(0, 1)$. We define

$$\begin{aligned} R_{\mathbf{f}}(\boldsymbol{\eta}) &= \left\{ \int_0^1 \|\mathbf{f}'(t) - \mathbf{F}(t, \mathbf{f}(t), \boldsymbol{\eta})\|^2 w(t) dt \right\}^{1/2}, \\ \boldsymbol{\psi}(\mathbf{f}) &= \arg \min_{\boldsymbol{\eta} \in \Theta} R_{\mathbf{f}}(\boldsymbol{\eta}). \end{aligned} \quad (2.9)$$

It is easy to check that $\boldsymbol{\psi}(\mathbf{f}_{\boldsymbol{\eta}}) = \boldsymbol{\eta}$ for all $\boldsymbol{\eta} \in \Theta$. Thus the map $\boldsymbol{\psi}$ extends the definition of the parameter $\boldsymbol{\theta}$ beyond the model. Let us define $\boldsymbol{\theta}_0 = \boldsymbol{\psi}(\mathbf{f}_0)$. Thus, $\boldsymbol{\theta}_0$ describes the projection of the true regression function on the parametric model. We assume that $\boldsymbol{\theta}_0$ lies in the interior of Θ . From now on, we shall write $\boldsymbol{\theta}$ for $\boldsymbol{\psi}(\mathbf{f})$ and treat it as the parameter of interest. A posterior is induced on Θ through the mapping $\boldsymbol{\psi}$ acting on $\mathbf{f}(\cdot) = \mathbf{B}_n^T \mathbf{N}(\cdot)$ and the posterior of \mathbf{B}_n given by (2.8).

Remark 1: Note that $\mathbf{f}_0(\cdot)$ need not be a solution of the ODE. In real life it is almost impossible to accurately describe a data generating mechanism in terms of a mathematical model. ODE is a useful tool to model many dynamic systems within a margin of error. This justifies the study of misspecified regression function in the context of ODE model. Often we are more interested in inferring on the parameters rather than the regression function described by the ODE model. Then the role of the true parameter is played by the parameter value which brings the ODE model closest to the true regression function. The definition of $\boldsymbol{\theta}_0$ reinforces this intuition.

3. Main results

Our objective is to study the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. The asymptotic representation of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is given by the next theorem under the assumption that

$$\text{for all } \epsilon > 0, \inf_{\boldsymbol{\eta}: \|\boldsymbol{\eta} - \boldsymbol{\theta}_0\| \geq \epsilon} R_{\mathbf{f}_0}(\boldsymbol{\eta}) > R_{\mathbf{f}_0}(\boldsymbol{\theta}_0). \tag{3.1}$$

We denote $D_{l,r,s}\mathbf{F}(t, \mathbf{f}, \boldsymbol{\theta}) = \partial^{l+r+s} / \partial \boldsymbol{\theta}^s \partial \mathbf{f}^r \partial t^l \mathbf{F}(t, \mathbf{f}(t), \boldsymbol{\theta})$. Since the posterior distributions of $\boldsymbol{\beta}_j$ are mutually independent when $\boldsymbol{\varepsilon}_{\cdot,j}$ are mutually independent for $j = 1, \dots, d$, we can assume $d = 1$ in Theorem 1 for the sake of simplicity in notation and write $f(\cdot), f_0(\cdot), F(\cdot, \cdot, \cdot), \boldsymbol{\beta}$ instead of $\mathbf{f}(\cdot), \mathbf{f}_0(\cdot), \mathbf{F}(\cdot, \cdot, \cdot)$ and \mathbf{B}_n respectively. Extension to d -dimensional case is straightforward as shown in Remark 5 after the statement of Theorem 1. We deal with the situation of correlated errors in Section 4.

Theorem 1. *Let the matrix*

$$\begin{aligned} \mathbf{J}_{\boldsymbol{\theta}_0} &= \int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0) w(t) dt \\ &\quad - \int_0^1 (D_{0,0,1}\mathbf{S}(t, f_0(t), \boldsymbol{\theta}_0)) w(t) dt \end{aligned}$$

be nonsingular, where

$$\mathbf{S}(t, f(t), \boldsymbol{\theta}) = (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0)).$$

Let m be an integer greater than or equal to 5 and $n^{1/2m} \ll k_n \ll n^{1/8}$. If $D_{0,2,1}F(t, y, \boldsymbol{\theta})$ and $D_{0,0,2}F(t, y, \boldsymbol{\theta})$ are continuous in their arguments, then under the assumption (3.1), there exists $E_n \subseteq C^m((0, 1)) \times \boldsymbol{\Theta}$ with $\Pi(E_n^c | \mathbf{Y}) = o_{P_0}(1)$, such that uniformly for $(f, \boldsymbol{\theta}) \in E_n$,

$$\|\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{J}_{\boldsymbol{\theta}_0}^{-1} \sqrt{n}(\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0))\| \rightarrow 0 \tag{3.2}$$

as $n \rightarrow \infty$, where

$$\begin{aligned} \boldsymbol{\Gamma}(z) &= \int_0^1 \left(-(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T D_{0,1,0}F(t, f_0(t), \boldsymbol{\theta}_0) w(t) \right. \\ &\quad \left. - \frac{d}{dt} [(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T w(t)] + (D_{0,1,0}\mathbf{S}(t, f_0(t), \boldsymbol{\theta}_0)) w(t) \right) z(t) dt. \end{aligned}$$

Remark 2: Condition (3.1) implies that $\boldsymbol{\theta}_0$ is the unique point of minimum of $R_{\mathbf{f}_0}(\cdot)$ and $\boldsymbol{\theta}_0$ should be a well-separated point of minimum.

Remark 3: The posterior distribution of $\boldsymbol{\Gamma}(\mathbf{f}) - \boldsymbol{\Gamma}(\mathbf{f}_0)$ contracts at $\mathbf{0}$ at the rate $n^{-1/2}$ as indicated by Lemma 4. Hence, the posterior distribution of $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ contracts at $\mathbf{0}$ at the rate $n^{-1/2}$ with high probability under the truth. We refer to Theorem 2 for a more refined version of this result.

Remark 4: We note that fifth order smoothness of the true mean function is good enough to ensure the contraction rate $n^{-1/2}$. We do not gain by assuming a higher order of smoothness. For $m = 5$, the required condition becomes $n^{1/10} \ll k_n \ll n^{1/8}$. Also, the knots are chosen deterministically and there is no need to assign a prior on the number of terms of the random series used. Hence, the issue of Bayesian adaptation, that is, improving convergence rate with higher smoothness without knowing the smoothness, does not arise in the present context.

Remark 5: When the response is a d -dimensional vector, (3.2) holds with the scalars being replaced by the corresponding d -dimensional vectors. Let $\mathbf{A}(t)$ stands for the $p \times d$ matrix

$$\begin{aligned} & \mathbf{J}_{\theta_0}^{-1} \{ -(D_{0,0,1} \mathbf{F}(t, \mathbf{f}_0(t), \theta_0))^T D_{0,1,0} \mathbf{F}(t, \mathbf{f}_0(t), \theta_0) w(t) \\ & - \frac{d}{dt} [(D_{0,0,1} \mathbf{F}(t, \mathbf{f}_0(t), \theta_0))^T w(t)] + (D_{0,1,0} \mathbf{S}(t, \mathbf{f}_0(t), \theta_0)) w(t) \}. \end{aligned}$$

Then we have

$$\mathbf{J}_{\theta_0}^{-1} \Gamma(\mathbf{f}) = \sum_{j=1}^d \int_0^1 \mathbf{A}_{\cdot,j}(t) \mathbf{N}^T(t) \beta_j dt = \sum_{j=1}^d \mathbf{G}_{n,j}^T \beta_j, \quad (3.3)$$

where $\mathbf{G}_{n,j}^T = \int_0^1 \mathbf{A}_{\cdot,j}(t) \mathbf{N}^T(t) dt$ which is a $p \times (k_n + m - 1)$ matrix for $j = 1, \dots, d$. Then in order to approximate the posterior distribution of θ , it suffices to study the asymptotic posterior distribution of the linear combination of β_j given by (3.3). The next theorem describes the approximate posterior distribution of $\sqrt{n}(\theta - \theta_0)$.

Theorem 2. Define

$$\begin{aligned} \mu_n &= \sqrt{n} \sum_{j=1}^d \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_{\cdot,j} - \sqrt{n} \mathbf{J}_{\theta_0}^{-1} \Gamma(\mathbf{f}_0), \\ \Sigma_n &= n \sum_{j=1}^d \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j} \end{aligned}$$

and $\mathbf{B}_j = ((\langle A_{k,j}(\cdot), A_{k',j}(\cdot) \rangle))_{k,k'=1,\dots,p}$ for $j = 1, \dots, d$. If \mathbf{B}_j is non-singular for all $j = 1, \dots, d$, then under the conditions of Theorem 1,

$$\|\Pi(\sqrt{n}(\theta - \theta_0) \in \cdot | \mathbf{Y}) - N(\mu_n, \sigma^2 \Sigma_n)\|_{TV} = o_{P_0}(1). \quad (3.4)$$

Inspecting the proof, we can conclude that (3.4) is uniform over σ^2 belonging to a compact subset of $(0, \infty)$. Also note that the scale of the approximating normal distribution involves the working variance σ^2 assuming that it is given, even though the convergence is studied under the true distribution P_0 with variance σ_0^2 , not necessarily equal to the given σ^2 . Thus, the distribution matches with the frequentist distribution of the estimator in Brunel (2008) only if σ is correctly specified as σ_0 . The next result assures that putting a prior on σ rectifies the problem.

Theorem 3. We assign independent $N(\mathbf{0}, nc^{-1}k_n^{-1}\sigma^2(\mathbf{X}_n^T X_n)^{-1})$ prior on β_j for $j = 1, \dots, d$ for some constant $c > 0$ and inverse gamma prior on σ^2 with shape and scale parameters a and b respectively. If the fourth order moment of the true error distribution is finite, then under the conditions of Theorem 1,

$$\|\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot | \mathbf{Y}) - N(\boldsymbol{\mu}_n, \sigma_0^2 \boldsymbol{\Sigma}_n)\|_{TV} = o_{P_0}(1). \tag{3.5}$$

4. Extensions

The results obtained so far can be extended for the case where $\varepsilon_{i,j}$ and $\varepsilon_{i,j'}$ are associated for $i = 1, \dots, n$ and $j \neq j'$; $j, j' = 1, \dots, d$. Let under the working model, ε_i , have the dispersion matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Omega}$ for $i = 1, \dots, n$, $\boldsymbol{\Omega}$ being a known positive definite matrix. Denoting $\boldsymbol{\Omega}^{-1/2} = ((\omega^{jk}))_{j,k=1}^d$, we have the following extension of Theorem 2.

Theorem 4. Define

$$\begin{aligned} \boldsymbol{\mu}_n^* &= \sqrt{n} \sum_{k=1}^d \text{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \left((\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \left(\boldsymbol{\Omega}^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) \right) \\ &\quad \times (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_{\cdot j} \omega^{jk} - \sqrt{n} \mathbf{J}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{\Gamma}(\mathbf{f}_0), \\ \boldsymbol{\Sigma}_n^* &= n \sum_{k=1}^d \text{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \left((\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \left(\boldsymbol{\Omega}^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) \right) \\ &\quad \times (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \\ &\quad \times \text{rows}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \left(\left(\boldsymbol{\Omega}^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) (\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T)^T \right). \end{aligned}$$

Then under the conditions of Theorem 1 and Theorem 2,

$$\|\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot | \mathbf{Y}) - N(\boldsymbol{\mu}_n^*, \sigma^2 \boldsymbol{\Sigma}_n^*)\|_{TV} = o_{P_0}(1). \tag{4.1}$$

If σ^2 is unknown and is given an inverse gamma prior, then under the conditions of Theorems 1 and 3,

$$\|\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot | \mathbf{Y}) - N(\boldsymbol{\mu}_n^*, \sigma_0^2 \boldsymbol{\Sigma}_n^*)\|_{TV} = o_{P_0}(1), \tag{4.2}$$

where σ_0^2 is the true value of σ^2 .

Remark 6: In many applications, the regression function is modeled as $\mathbf{h}_\theta(t) = \mathbf{g}(\mathbf{f}_\theta(t))$ instead of $\mathbf{f}_\theta(t)$, where \mathbf{g} is a known invertible function and $\mathbf{h}_\theta(t) \in \mathbb{R}^d$. It should be noted that

$$\begin{aligned} \frac{d\mathbf{h}_\theta(t)}{dt} &= \mathbf{g}'(\mathbf{f}_\theta(t)) \frac{d\mathbf{f}_\theta(t)}{dt} = \mathbf{g}'(\mathbf{g}^{-1}\mathbf{h}_\theta(t)) \mathbf{F}(t, \mathbf{g}^{-1}\mathbf{h}_\theta(t), \boldsymbol{\theta}) \\ &= \mathbf{H}(t, \mathbf{h}_\theta(t), \boldsymbol{\theta}), \end{aligned}$$

which is a known function of t, \mathbf{h}_θ and $\boldsymbol{\theta}$. Now we can carry out our analysis replacing \mathbf{F} and \mathbf{f}_θ in (1.1) by \mathbf{H} and \mathbf{h}_θ respectively.

Remark 7: Often we do not have the data on all the state variables. For the sake of simplicity let $d = 2$. Let the true regression function be $(f_{1\theta_0}(\cdot), f_{2\theta_0}(\cdot))^T$ and suppose that only the first component Y_1 of the response variable \mathbf{Y} is observable. Let the system of ODE be given by

$$\frac{d}{dt}f_{1\theta}(t) = F_1(t, f_{1\theta}(t), f_{2\theta}(t), \theta) \quad (4.3)$$

$$\frac{d}{dt}f_{2\theta}(t) = F_2(t, f_{1\theta}(t), f_{2\theta}(t), \theta). \quad (4.4)$$

Model $f_1(\cdot)$ by a spline series $\beta^T \mathbf{N}(\cdot)$, where β is a free parameter. We substitute this expression for $f_1(\cdot)$ and apply the four stage Runge-Kutta (RK4) method on (4.4) with meshwidth $h_n \ll n^{-1/8}$ to obtain the corresponding nonparametric regression model for Y_2 on t given by $f_2(t) = \phi_n(t, f_1(t), \theta)$. Now we define

$$\theta = \arg \min_{\eta \in \Theta} \int_0^1 (f_1'(t) - F_1(t, f_1(t), f_2(t), \eta))^2 w(t) dt. \quad (4.5)$$

Since the initial condition does not appear in the optimization in (4.5), the formulation of the two-step procedure is equivalent of treating the initial condition unknown, which can be absorbed in the vector of unknown parameters. Under the identifiability assumption that $f_{1\theta}(\cdot) \neq f_{1\theta'}(\cdot)$ whenever $\theta \neq \theta'$, the posterior distribution of θ induced from that of $f_1(\cdot)$ will satisfy the Bernstein-von Mises theorem with $n^{-1/2}$ rate of contraction towards θ_0 . The proof of this assertion is given later.

5. Simulation study

We consider the Lotka-Volterra equations to study the posterior distribution of θ . We consider the case when the true regression function is the solution of the ODE. For a sample of size n , the x_i 's are chosen as $x_i = (2i - 1)/2n$ for $i = 1, \dots, n$. This choice of design points satisfy (2.6) with $Q(t) = t$. Samples of sizes 50, 100 and 500 are considered. The weight function is chosen as $w(t) = t(1 - t)$, $t \in [0, 1]$. We simulate 1000 replications for each case. Under each replication a sample of size 1000 is directly drawn from the posterior distribution of θ and then 95% equal tailed credible interval is obtained. This method is abbreviated by "TS" in the tables. Each replication took around one minute. We calculate the coverage and the average length of the corresponding credible interval over these 1000 replications. The estimated standard errors of the interval length and coverage are given inside the parentheses in the tables. We also consider 1000 replications to construct the 95% equal tailed confidence interval based on asymptotic normality as obtained from the estimation method introduced by Varah (1982) and modified and studied by Brunel (2008). We abbreviate this method by "VB" in tables. The estimated standard errors of the interval length and coverage are given inside the parentheses in the tables. We also conduct a simulation using Bayesian nonlinear least squares method putting

TABLE 1
 Coverages and average lengths of the Bayesian credible intervals using TS and RK4 and confidence interval obtained from VB method for Gaussian error for Lotka-Volterra equations

n		TS		VB		RK4	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
50	θ_1	89.3 (0.04)	6.13 (1.04)	83.6 (0.05)	4.57 (0.84)	85.2 (0.05)	1.84 (0.23)
	θ_2	97.3 (0.02)	6.37 (1.27)	82.4 (0.05)	4.26 (0.80)	86.2 (0.05)	2.09 (0.27)
	θ_3	93.7 (0.03)	6.59 (1.21)	86.6 (0.05)	4.96 (0.98)	85.9 (0.05)	2.03 (0.34)
	θ_4	98.3 (0.02)	6.58 (1.40)	85.5 (0.05)	4.39 (0.88)	85.5 (0.05)	1.85 (0.31)
100	θ_1	93.0 (0.02)	4.24 (0.58)	88.6 (0.03)	3.38 (0.46)	91.6 (0.03)	1.57 (0.17)
	θ_2	97.3 (0.02)	4.21 (0.63)	89.0 (0.03)	3.15 (0.44)	92.0 (0.03)	1.78 (0.20)
	θ_3	93.3 (0.02)	4.49 (0.62)	88.9 (0.03)	3.60 (0.51)	92.6 (0.03)	1.73 (0.22)
	θ_4	97.3 (0.02)	4.27 (0.62)	87.4 (0.03)	3.19 (0.46)	91.7 (0.03)	1.57 (0.20)
500	θ_1	95.7 (0.01)	1.71 (0.11)	94.6 (0.01)	1.55 (0.09)	98.5 (0.00)	0.97 (0.08)
	θ_2	97.0 (0.01)	1.63 (0.10)	93.8 (0.01)	1.45 (0.09)	98.9 (0.00)	1.10 (0.09)
	θ_3	95.0 (0.01)	1.84 (0.13)	93.8 (0.01)	1.66 (0.10)	98.5 (0.00)	1.06 (0.10)
	θ_4	97.0 (0.01)	1.67 (0.11)	93.5 (0.01)	1.48 (0.09)	98.3 (0.00)	0.96 (0.09)

the same inverse Gamma prior on σ^2 and independent Gaussian priors on θ_j for $j = 1, \dots, 4$. The four stage Runge-Kutta (RK4) method is used to construct the numerical solution of the ODE with n equispaced grid points for a sample of size n . This numerical solution gives the likelihood based on the Gaussian working model and hence we construct posterior for θ . We abbreviated this approach by “RK4” in the tables. Each replication took around two hours in this method.

Thus we have $p = 4, d = 2$ and the ODE’s are given by

$$\begin{aligned} F_1(t, \mathbf{f}_\theta(t), \theta) &= \theta_1 f_{1\theta}(t) - \theta_2 f_{1\theta}(t) f_{2\theta}(t), \\ F_2(t, \mathbf{f}_\theta(t), \theta) &= -\theta_3 f_{2\theta}(t) + \theta_4 f_{1\theta}(t) f_{2\theta}(t) \end{aligned}$$

for $t \in [0, 1]$ with initial condition $f_{1\theta}(0) = 1, f_{2\theta}(0) = 0.5$. The above system is not analytically solvable. We take $\theta_0 = (10, 10, 10, 10)^T$. The true regression function is $\mathbf{f}_{\theta_0}(\cdot) = (f_{1\theta_0}(\cdot), f_{2\theta_0}(\cdot))^T$.

The true distribution of error is taken either $N(0, (0.2)^2)$ or a scaled t -distribution with 6 degrees of freedom, where scaling is done in order to make the standard deviation 0.2. We put an inverse gamma prior on σ^2 with shape and scale parameters being 99 and 1 respectively and independent Gaussian priors on β_1 and β_2 with mean vector $\mathbf{0}$ and dispersion matrix $nc^{-1}k_n^{-1}\sigma^2(\mathbf{X}_n^T \mathbf{X}_n)^{-1}$ with

TABLE 2
 Coverages and average lengths of the Bayesian credible intervals using TS and RK4 and confidence interval obtained from VB method for scaled t_6 error for Lotka-Volterra equations

n		TS		VB		RK4	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
50	θ_1	88.9	6.13	83.6	4.55	85.5	1.83
		(0.04)	(1.02)	(0.05)	(0.93)	(0.05)	(0.25)
	θ_2	97.0	6.36	81.1	4.24	86.3	2.08
		(0.02)	(1.24)	(0.06)	(0.87)	(0.05)	(0.30)
	θ_3	92.0	6.57	85.1	4.93	87.0	2.03
		(0.04)	(1.23)	(0.05)	(1.06)	(0.05)	(0.32)
	θ_4	97.7	6.54	84.3	4.36	86.5	1.84
		(0.02)	(1.42)	(0.05)	(0.94)	(0.05)	(0.29)
100	θ_1	92.0	4.22	87.6	3.34	90.7	1.56
		(0.03)	(0.57)	(0.03)	(0.47)	(0.03)	(0.18)
	θ_2	98.3	4.19	88.1	3.13	90.8	1.78
		(0.01)	(0.62)	(0.03)	(0.45)	(0.03)	(0.20)
	θ_3	94.7	4.53	89.7	3.61	90.0	1.73
		(0.02)	(0.65)	(0.03)	(0.56)	(0.03)	(0.23)
	θ_4	98.7	4.29	89.4	3.21	90.1	1.57
		(0.01)	(0.67)	(0.03)	(0.50)	(0.03)	(0.21)
500	θ_1	95.3	1.72	93.8	1.55	97.6	0.96
		(0.01)	(0.12)	(0.01)	(0.10)	(0.01)	(0.08)
	θ_2	96.3	1.64	93.2	1.45	97.8	1.09
		(0.00)	(0.12)	(0.01)	(0.10)	(0.01)	(0.09)
	θ_3	94.7	1.84	93.8	1.66	98.1	1.06
		(0.01)	(0.13)	(0.01)	(0.12)	(0.01)	(0.10)
	θ_4	97.7	1.67	94.3	1.48	98.1	0.96
		(0.00)	(0.12)	(0.01)	(0.10)	(0.01)	(0.09)

$c = 3.5$ for TS. We choose $k_n - 1$ equispaced interior knots $\frac{1}{k_n}, \frac{2}{k_n}, \dots, \frac{k_n-1}{k_n}$. This specific choice of knots satisfies the quasi-uniformity criteria (2.5) with $M = 1$. As far as choosing k_n is concerned, we take $k_n = 16, 17, 20$ for $n = 50, 100$ and 500 respectively by taking the order of k_n as $n^{1/9}$ as suggested by Theorem 1. The constant multiplier to the chosen asymptotic order is selected through cross validation. For RK4 we put the same inverse Gamma prior on σ^2 and independent $N(6, 4^2)$ priors on θ_j for $j = 1, \dots, 4$. The simulation results are summarized in the Tables 1 and 2. Not surprisingly asymptotic normality based confidence intervals obtained from VB method are shorter but too optimistic, failing to give adequate coverage for finite sample sizes since delta method is known to underestimate variation. The RK4 credible intervals are shorter because of asymptotic efficiency. But the corresponding coverages are much lower than 95% for small samples. Also RK4 method is more computationally expensive.

6. Real life data

(Barnes' problem) We consider the Barnes' problem given by the chemical reaction equations

$$\frac{df_{1\theta}(t)}{dt} = \theta_1 f_{1\theta}(t) - \theta_2 f_{1\theta}(t) f_{2\theta}(t)$$

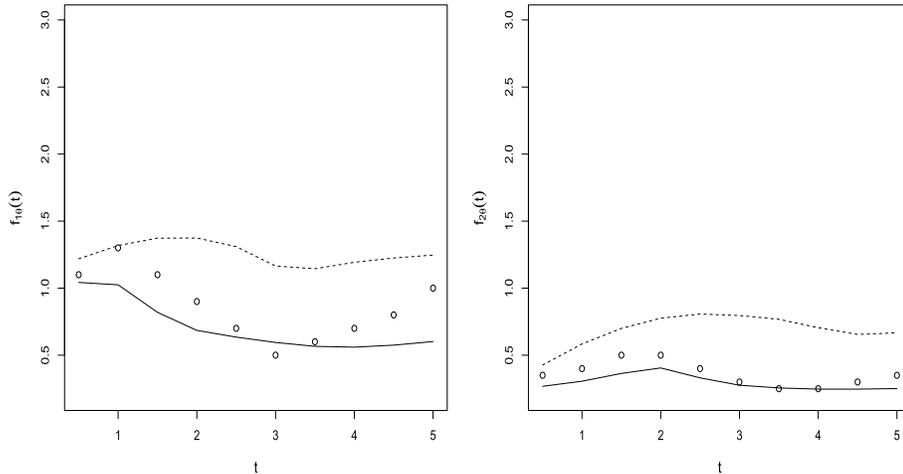


FIG 1. Observed values and the 95% posterior predictive intervals of $f_{1\theta}$ and $f_{2\theta}$ over the 10 data points.

$$\frac{df_{2\theta}(t)}{dt} = \theta_2 f_{1\theta}(t) f_{2\theta}(t) - \theta_3 f_{2\theta}(t).$$

The data can be found in Varah (1982) where we have 10 data points with initial value (1,0.3). We use B-spline basis of order 7 with $k_n = 2$, where we take the order of k_n as $n^{1/9}$ as suggested in Theorem 1. Again we select the constant multiplier to the chosen asymptotic order using cross validation. We put an inverse gamma prior on σ^2 with shape and scale parameters 1000 and 1 respectively and use $w(t) = t^{0.3}(1 - t)^{0.3}$. Conditional on σ^2 we put Gaussian prior on β with mean vector $\mathbf{0}$ and dispersion matrix $nc^{-1}k_n^{-1}\sigma^2(\mathbf{X}_n^T \mathbf{X}_n)^{-1}$ with $c = 100$. Samples of size 1000 are drawn from the posterior distributions of θ_1, θ_2 and θ_3 . The 95% posterior predictive interval of $f_{1\theta}(\cdot)$ and $f_{2\theta}(\cdot)$ at the 10 data points is superimposed on the data in Figure 1. The figures show that most of the observed values are falling within the predictive intervals. For $f_{2\theta}$ some data points are very close to the lower bound which may be attributed to the small sample size of 10.

7. Proofs

Proof of Theorem 1. The structure of the proof follows that of Proposition 3.1 of Brunel (2008) and Proposition 3.3 of Gugushvili and Klaassen (2012), but differs substantially in detail since we address posterior variation and also allow misspecification. First note that $\mathbf{\Gamma}(f) - \mathbf{\Gamma}(f_0)$ can be expressed as

$$\int_0^1 (-(D_{0,0,1}F(t, f_0(t), \theta_0))^T D_{0,1,0}F(t, f_0(t), \theta_0)w(t) \tag{7.1}$$

$$-\frac{d}{dt}[(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T w(t)] + (D_{0,1,0}\mathbf{S}(t, f_0(t), \boldsymbol{\theta}_0)) w(t) \\ \times (f(t) - f_0(t))dt.$$

Interchanging the orders of differentiation and integration and using the definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$,

$$\int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T (f'(t) - F(t, f(t), \boldsymbol{\theta}))w(t)dt = \mathbf{0}, \quad (7.2)$$

$$\int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0))w(t)dt = \mathbf{0}. \quad (7.3)$$

Taking difference, we get

$$\int_0^1 ((D_{0,0,1}F(t, f(t), \boldsymbol{\theta}) - D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0))^T (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0))) w(t)dt \\ + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0) - D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T \\ \times (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0))w(t)dt \\ + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T (f'(t) - f'_0(t) \\ + F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0))w(t)dt \\ + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}) - D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0))^T (f'(t) - f'_0(t) \\ + F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0))w(t)dt \\ + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T (F(t, f(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta})) w(t)dt = \mathbf{0}.$$

Replacing the difference between the values of a function at two different values of an argument by the integral of the corresponding partial derivative, we get

$$\mathbf{M}(f, \boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ = \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0) - D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T \\ \times (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0))w(t)dt \\ + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0))^T (f'(t) - f'_0(t) \\ + F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0))w(t)dt,$$

where $\mathbf{M}(f, \boldsymbol{\theta})$ is given by

$$\int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T \left\{ \int_0^1 D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0 + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}_0))d\lambda \right\} w(t)dt$$

$$\begin{aligned}
 & - \int_0^1 \left\{ \int_0^1 (D_{0,0,1} \mathbf{S}(t, f(t), \boldsymbol{\theta}_0 + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}_0))) d\lambda \right\} w(t) dt \\
 & - \int_0^1 \left\{ \int_0^1 (D_{0,0,2} \mathbf{F}(t, f(t), \boldsymbol{\theta}_0 + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}_0))) d\lambda \right\} (f'(t) - f'_0(t)) \\
 & \quad + F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0) w(t) dt.
 \end{aligned}$$

Note that $\mathbf{M}(f_0, \boldsymbol{\theta}_0) = \mathbf{J}_{\boldsymbol{\theta}_0}$. We also define

$$E_n = \{(f, \boldsymbol{\theta}) : \sup_{t \in [0,1]} |f(t) - f_0(t)| \leq \epsilon_n, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \epsilon_n\},$$

where $\epsilon_n \rightarrow 0$. By Lemmas 2 and 3, there exists such a sequence $\{\epsilon_n\}$ so that $\Pi(E_n^c | \mathbf{Y}) = o_{P_0}(1)$. Then, $\mathbf{M}(f, \boldsymbol{\theta})$ is invertible and the eigenvalues of $[\mathbf{M}(f, \boldsymbol{\theta})]^{-1}$ are bounded away from 0 and ∞ for sufficiently large n and

$$\|(\mathbf{M}(f, \boldsymbol{\theta}))^{-1} - \mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\| = o(1)$$

for $(f, \boldsymbol{\theta}) \in E_n$. Hence, on E_n

$$\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = (\mathbf{J}_{\boldsymbol{\theta}_0}^{-1} + o(1)) \sqrt{n}(\mathbf{T}_{1n} + \mathbf{T}_{2n} + \mathbf{T}_{3n}),$$

for sufficiently large n , where

$$\begin{aligned}
 \mathbf{T}_{1n} &= \int_0^1 (D_{0,0,1} F(t, f(t), \boldsymbol{\theta}_0) - D_{0,0,1} F(t, f_0(t), \boldsymbol{\theta}_0))^T \\
 & \quad \times (f'_0(t) - F(t, f_0(t), \boldsymbol{\theta}_0)) w(t) dt, \\
 \mathbf{T}_{2n} &= \int_0^1 (D_{0,0,1} F(t, f(t), \boldsymbol{\theta}_0))^T (f'(t) - f'_0(t)) w(t) dt, \\
 \mathbf{T}_{3n} &= \int_0^1 (D_{0,0,1} F(t, f(t), \boldsymbol{\theta}_0))^T (F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0)) w(t) dt.
 \end{aligned}$$

In view of Lemmas 2 and 4, on a set in the sample space with high true probability, the posterior distribution of $\mathbf{J}_{\boldsymbol{\theta}_0}^{-1} \sqrt{n}(\mathbf{T}_{1n} + \mathbf{T}_{2n} + \mathbf{T}_{3n})$ assigns most of its mass inside a large compact set. Thus, we can assert that inside the set E_n , the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is given by that of

$$\mathbf{J}_{\boldsymbol{\theta}_0}^{-1} \sqrt{n}(\mathbf{T}_{1n} + \mathbf{T}_{2n} + \mathbf{T}_{3n}). \tag{7.4}$$

We shall extract $\sqrt{n} \mathbf{J}_{\boldsymbol{\theta}_0}^{-1} (\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0))$ from (7.4) and show that the remainder term goes to zero. First write

$$\begin{aligned} \mathbf{T}_{2n} &= - \int_0^1 \left(\frac{d}{dt} [(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T w(t)] \right) (f(t) - f_0(t)) dt \\ &\quad + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0) - D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T (f'(t) - f'_0(t)) w(t) dt, \end{aligned}$$

which follows by integration by parts and the fact that $w(0) = w(1) = 0$. Note that the first integral of the above equation appears in (7.1). The norm of the second integral is bounded by $\sup_{t \in [0,1]} |f(t) - f_0(t)|^2 + \sup_{t \in [0,1]} |f'(t) - f'_0(t)|^2$ up to a constant multiple using the continuity of $D_{0,1,1}F(t, y, \boldsymbol{\theta})$. Now we consider \mathbf{T}_{3n} in (7.4). Then,

$$\begin{aligned} \mathbf{T}_{3n} &= \int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T (F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0)) w(t) dt \\ &\quad + \int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}_0) - D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T \\ &\quad \quad \quad \times (F(t, f_0(t), \boldsymbol{\theta}_0) - F(t, f(t), \boldsymbol{\theta}_0)) w(t) dt. \end{aligned} \quad (7.5)$$

The first integral on the right hand side of (7.5) can be written as

$$\begin{aligned} &- \int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T D_{0,1,0}F(t, f_0(t), \boldsymbol{\theta}_0) (f(t) - f_0(t)) w(t) dt \\ &- \int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T \\ &\quad \times \left\{ \int_0^1 [D_{0,1,0}F(t, f_0(t) + \lambda(f - f_0)(t), \boldsymbol{\theta}_0) - D_{0,1,0}F(t, f_0(t), \boldsymbol{\theta}_0)] d\lambda \right\} \\ &\quad \quad \quad \times (f(t) - f_0(t)) w(t) dt \\ &= \mathbf{T}_{31n} + \mathbf{T}_{32n}, \end{aligned}$$

say. Now \mathbf{T}_{31n} appears in (7.1). By the continuity of $D_{0,2,0}F(t, y, \boldsymbol{\theta})$, $\|\mathbf{T}_{32n}\|$ can be bounded above by a constant multiple of $\sup_{t \in [0,1]} |f(t) - f_0(t)|^2$. We apply the Cauchy-Schwarz inequality and the continuity of $D_{0,1,1}F(t, y, \boldsymbol{\theta})$ to bound the second integral on the right hand side of (7.5) by a constant multiple of $\sup\{|f(t) - f_0(t)|^2 : t \in [0, 1]\}$. The term \mathbf{T}_{1n} inside the bracket of (7.4) is given by

$$\begin{aligned} &\int_0^1 (D_{0,1,0}\mathbf{S}(t, f_0(t), \boldsymbol{\theta}_0)) (f(t) - f_0(t)) w(t) dt \\ &\quad + \int_0^1 \left\{ \int_0^1 (D_{0,1,0}\mathbf{S}(t, f_0(t) + \lambda(f - f_0)(t), \boldsymbol{\theta}_0) - D_{0,1,0}\mathbf{S}(t, f_0(t), \boldsymbol{\theta}_0)) d\lambda \right\} \\ &\quad \quad \quad \times (f(t) - f_0(t)) w(t) dt. \end{aligned}$$

The first integral appears in (7.1). The norm of the second integral of the above display can be bounded by a multiple of $\sup\{|f(t) - f_0(t)|^2 : t \in [0, 1]\}$ using the continuity of $D_{0,2,1}F(t, y, \boldsymbol{\theta})$ with respect to its arguments. Combining these,

we find that the norm of $\mathbf{J}_{\theta_0}^{-1} \sqrt{n}(\mathbf{T}_{1n} + \mathbf{T}_{2n} + \mathbf{T}_{3n}) - \mathbf{J}_{\theta_0}^{-1} \sqrt{n}(\Gamma(f) - \Gamma(f_0))$ is bounded above by a multiple of $\sqrt{n} \sup_{t \in [0,1]} |f(t) - f_0(t)|^2 + \sqrt{n} \sup_{t \in [0,1]} |f'(t) - f'_0(t)|^2$. Now applying Lemma 2, we get the desired result. \square

Proof of Theorem 2. By Theorem 1 and (3.3), it suffices to show that

$$\left\| \Pi \left(\sqrt{n} \sum_{j=1}^d \mathbf{G}_{n,j}^T \beta_j - \sqrt{n} \mathbf{J}_{\theta_0}^{-1} \Gamma(\mathbf{f}_0) \in \cdot \mid \mathbf{Y} \right) - N(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n) \right\|_{TV} = o_{P_0}(1). \tag{7.6}$$

Note that the posterior distribution of $\mathbf{G}_{n,j}^T \beta_j$ is a normal with mean vector $(1 + \sigma^2 ck_n/n)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j$ and dispersion matrix given by $\sigma^2(1 + \sigma^2 ck_n/n)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$. We calculate the Kullback-Leibler divergence between two Gaussian distributions to prove the assertion. Alternatively, we can also follow the approach given in Theorem 1 and Corollary 1 of Bontemps (2011). The Kulback-Leibler divergence between the distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_2)$ is given by

$$\frac{1}{2} (\text{tr}(\boldsymbol{\Omega}_1^{-1} \boldsymbol{\Omega}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Omega}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - p - \log(\det(\boldsymbol{\Omega}_1^{-1} \boldsymbol{\Omega}_2))).$$

In the present context $\boldsymbol{\mu}_1 = (1 + \sigma^2 ck_n/n)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j$, $\boldsymbol{\mu}_2 = \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j$, $\boldsymbol{\Omega}_1 = \sigma^2(1 + \sigma^2 ck_n/n)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$ and $\boldsymbol{\Omega}_2 = \sigma^2 \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$. Note that $\text{tr}(\boldsymbol{\Omega}_1^{-1} \boldsymbol{\Omega}_2) = p + o(1)$ and $\log(\det(\boldsymbol{\Omega}_1^{-1} \boldsymbol{\Omega}_2)) = p \log(1 + c\sigma^2 k_n/n) \asymp k_n/n = o(1)$. From the proof of Lemma 4, it follows that

$$\begin{aligned} & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Omega}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ & \asymp n \frac{k_n^2}{n^2} \mathbf{Y}_j^T \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j \\ & \lesssim n \frac{k_n^2}{n^2} \frac{1}{k_n} \frac{k_n^2}{n^2} \frac{n}{k_n} \mathbf{Y}_j^T \mathbf{Y}_j = o_{P_0}(1). \end{aligned}$$

Hence, the total variation distance between the posterior distribution of $\mathbf{G}_{n,j}^T \beta_j$ and a Gaussian distribution with mean $\mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j$ and dispersion matrix $\sigma^2 \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$ converges in P_0 -probability to zero for $j = 1, \dots, d$. Since the posterior distributions of β_j and $\beta_{j'}$ are mutually independent for $j \neq j'$; $j, j' = 1, \dots, d$, we can assert that the posterior distribution of $\sqrt{n} \sum_{j=1}^d \mathbf{G}_{n,j}^T \beta_j - \sqrt{n} \mathbf{J}_{\theta_0}^{-1} \Gamma(\mathbf{f}_0)$ can be approximated in total variation by $N(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n)$. \square

Proof of Theorem 3. The marginal posterior of σ^2 is also inverse gamma with parameters $(dn + 2a)/2$ and $b + \sum_{j=1}^d \mathbf{Y}_j^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n} (1 + c(k_n/n))^{-1}) \mathbf{Y}_j / 2$, where $\mathbf{P}_{\mathbf{X}_n} = \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T$. Straightforward calculations show that

$$\begin{aligned} \mathbb{E}(\sigma^2|\mathbf{Y}) &= \frac{\frac{1}{2} \sum_{j=1}^d \{ \mathbf{Y}_{\cdot j}^T \mathbf{Y}_{\cdot j} - \mathbf{Y}_{\cdot j}^T \mathbf{P}_{\mathbf{X}_n} \mathbf{Y}_{\cdot j} (1 + ck_n n^{-1})^{-1} \} + b}{\frac{1}{2} dn + a - 1}, \\ \text{Var}(\sigma^2|\mathbf{Y}) &= \frac{(\mathbb{E}(\sigma^2|\mathbf{Y}))^2}{\frac{1}{2} dn + a - 2}. \end{aligned}$$

By Chebyshev's inequality, $|\mathbb{E}(\sigma^2|\mathbf{Y}) - \sigma_0^2| = O_{P_0}(n^{-1/2})$ and $\text{Var}(\sigma^2|\mathbf{Y}) = O_{P_0}(n^{-1})$. In particular, the marginal posterior distribution of σ^2 is consistent at the true value of error variance. Let \mathcal{N} be an arbitrary neighborhood of σ_0 . Then, $\Pi(\mathcal{N}^c|\mathbf{Y}) = o_{P_0}(1)$. We observe that

$$\begin{aligned} & \sup_{B \in \mathcal{R}^p} |\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in B|\mathbf{Y}) - \Phi(B; \boldsymbol{\mu}_n, \sigma_0^2 \boldsymbol{\Sigma}_n)| \\ & \leq \int \sup_{B \in \mathcal{R}^p} |\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in B|\mathbf{Y}, \sigma) - \Phi(B; \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n)| d\Pi(\sigma|\mathbf{Y}) \\ & \quad + \int \sup_{B \in \mathcal{R}^p} |\Phi(B; \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n) - \Phi(B; \boldsymbol{\mu}_n, \sigma_0^2 \boldsymbol{\Sigma}_n)| d\Pi(\sigma|\mathbf{Y}) \\ & \leq \sup_{\sigma \in \mathcal{N}} \sup_{B \in \mathcal{R}^p} |\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in B|\mathbf{Y}, \sigma) - \Phi(B; \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n)| \\ & \quad + \sup_{\sigma \in \mathcal{N}, B \in \mathcal{R}^p} |\Phi(B; \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n) - \Phi(B; \boldsymbol{\mu}_n, \sigma_0^2 \boldsymbol{\Sigma}_n)| + 2\Pi(\mathcal{N}^c|\mathbf{Y}). \end{aligned}$$

The total variation distance between the two normal distributions appearing in the second term of the previous display is bounded by a constant multiple of $|\sigma - \sigma_0|$, and hence the term can be made arbitrarily small by choosing \mathcal{N} appropriately. The first term converges in probability to zero by (3.4). The third term converges in probability to zero by the posterior consistency. Hence, we get the desired result. \square

Proof of Theorem 4. According to the fitted model, $\mathbf{Y}_i^{1 \times d} \sim N_d((\mathbf{X}_n)_i, \mathbf{B}_n, \boldsymbol{\Sigma}^{d \times d})$ for $i = 1, \dots, n$. The logarithm of the posterior probability density function (p.d.f.) is negative half times

$$\sum_{i=1}^n ((\mathbf{X}_n)_i, \mathbf{B}_n - \mathbf{Y}_i) \boldsymbol{\Sigma}^{-1} (\mathbf{B}_n^T (\mathbf{X}_n^T)_{\cdot i} - \mathbf{Y}_i^T) + \sum_{j=1}^d \boldsymbol{\beta}_j^T \frac{\mathbf{X}_n^T \mathbf{X}_n}{nc^{-1}k_n^{-1}} \boldsymbol{\beta}_j, \quad (7.7)$$

where $\mathbf{B}_n = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d)$. The quadratic term in $\boldsymbol{\beta}_j$ above for $j = 1, \dots, d$, can be consolidated to

$$\text{tr} \left(\left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right) \mathbf{B}_n^T \mathbf{X}_n^T \mathbf{X}_n \mathbf{B}_n \right). \quad (7.8)$$

The term in (7.7) which is linear in $\boldsymbol{\beta}_j$, $j = 1, \dots, d$, is given by

$$\sum_{i=1}^n (\mathbf{X}_n)_i, (\boldsymbol{\beta}_1 \dots \boldsymbol{\beta}_d) \boldsymbol{\Sigma}^{-1} \mathbf{Y}_i^T = \text{tr} (\mathbf{X}_n \mathbf{B}_n \boldsymbol{\Sigma}^{-1} \mathbf{Y}^T) = \text{tr} (\boldsymbol{\Sigma}^{-1} \mathbf{Y}^T \mathbf{X}_n \mathbf{B}_n).$$

A completing square argument gives the posterior density to be proportional to

$$\exp \left\{ -\frac{1}{2} \text{tr} \left[\left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right) \left(\mathbf{B}_n - (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{-1} \right)^T \right. \right. \\ \left. \left. \mathbf{X}_n^T \mathbf{X}_n \left(\mathbf{B}_n - (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{-1} \right) \right] \right\},$$

which can be identified with the pdf of a matrix normal distribution. More precisely,

$$\text{vec}(\mathbf{B}_n) | \mathbf{Y} \sim N \left(\text{vec} \left((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{-1} \right), \right. \\ \left. \left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{-1} \otimes (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \right).$$

For $j = 1, \dots, d$, the posterior mean of β_j is a weighted sum of $(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_{j'}$ for $j' = 1, \dots, d$. The weight attached with $(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j$ is of the order of 1, whereas for $j' \neq j$, the contribution from $(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_{j'}$ is of the order of k_n/n which goes to zero as n goes to infinity. Thus, the results of Lemmas 1 to 4 can be shown to hold under this setup. We are interested in the limiting distribution of $\mathbf{J}_{\theta_0}^{-1} \boldsymbol{\Gamma}(\mathbf{f}) = \sum_{j=1}^d \mathbf{G}_{n,j}^T \beta_j = (\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \text{vec}(\mathbf{B}_n)$. We note that the posterior distribution of $\left((\boldsymbol{\Sigma}^{-1} + ck_n \mathbf{I}_d/n)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) \text{vec}(\mathbf{B}_n)$ is a $(k_n + m - 1)d$ -dimensional normal distribution with mean vector and dispersion matrix being $\text{vec} \left((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + ck_n \mathbf{I}_d/n \right)^{-1/2} \right)$ and $\mathbf{I}_d \otimes (\mathbf{X}_n^T \mathbf{X}_n)^{-1}$ respectively, since by the properties of Kronecker product, for the matrices \mathbf{A} , \mathbf{B} and \mathbf{D} of appropriate orders $(\mathbf{B}^T \otimes \mathbf{A}) \text{vec}(\mathbf{D}) = \text{vec}(\mathbf{A} \mathbf{D} \mathbf{B})$. Let us consider the mean vector of the posterior distribution of the vector $\left((\boldsymbol{\Sigma}^{-1} + ck_n \mathbf{I}_d/n)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) \text{vec}(\mathbf{B}_n)$. We observe that

$$\begin{aligned} & (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{-1/2} \\ &= (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T (\mathbf{Y}_{,1} \dots \mathbf{Y}_{,d}) \left(\boldsymbol{\Sigma} + \frac{ck_n \boldsymbol{\Sigma}^2}{n} \right)^{-1/2} \\ &= (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \left(\sum_{j=1}^d \mathbf{Y}_{,j} c_{j1} \dots \sum_{j=1}^d \mathbf{Y}_{,j} c_{jd} \right), \end{aligned}$$

where $\mathbf{C}_n = ((c_{jk})) = (\boldsymbol{\Sigma} + ck_n \boldsymbol{\Sigma}^2/n)^{-1/2}$. For $k = 1, \dots, d$, we define \mathbf{Z}_k to be the sub-vector consisting of $[(k-1)(k_n + m - 1) + 1]^{th}$ to $[k(k_n + m - 1)]^{th}$ elements of the vector $\left((\boldsymbol{\Sigma}^{-1} + \frac{ck_n \mathbf{I}_d}{n})^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right) \text{vec}(\mathbf{B}_n)$. Then $\mathbf{Z}_k | \mathbf{Y} \sim$

$N_{k_n+m-1} \left((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j c_{jk}, (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \right)$. Also, the posterior distributions of \mathbf{Z}_k and $\mathbf{Z}_{k'}$ are mutually independent for $k \neq k'; k, k' = 1, \dots, d$. Now we show that the total variation distance between the posterior distribution of \mathbf{Z}_k and $N \left((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j \sigma^{jk}, (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \right)$ converges in P_0 -probability to zero for $k = 1, \dots, d$, where $\Sigma^{-1/2} = ((\sigma^{jk}))$. The total variation distance between two multivariate normal distributions with equal dispersion matrix $(\mathbf{X}_n^T \mathbf{X}_n)^{-1}$ and mean vectors $(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j c_{jk}$ and $(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j \sigma^{jk}$ is bounded by $\sum_{j=1}^d \|(\mathbf{X}_n^T \mathbf{X}_n)^{-1/2} \mathbf{X}_n^T \mathbf{Y}_j (c_{jk} - \sigma^{jk})\|$. Fixing k , for $j = 1, \dots, d$, we have that

$$\begin{aligned} \|(\mathbf{X}_n^T \mathbf{X}_n)^{-1/2} \mathbf{X}_n^T \mathbf{Y}_j (c_{jk} - \sigma^{jk})\| &= |c_{jk} - \sigma^{jk}| \left(\mathbf{Y}_j^T \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j \right)^{1/2} \\ &\leq |c_{jk} - \sigma^{jk}| \left(\mathbf{Y}_j^T \mathbf{Y}_j \right), \end{aligned}$$

since the eigenvalues of $\mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T$ are either zero or 1. Since clearly \mathbf{C}_n converges to $\Sigma^{-1/2}$ at the rate k_n/n , we have for $j = 1, \dots, d$,

$$\|(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j (c_{jk} - \sigma^{jk})\| \lesssim \frac{k_n}{n} O_{P_0}(\sqrt{n}) = o_{P_0}(1). \quad (7.9)$$

The total variation distance between $N((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j c_{jk}, (\mathbf{X}_n^T \mathbf{X}_n)^{-1})$ and $N((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j \sigma^{jk}, (\mathbf{X}_n^T \mathbf{X}_n)^{-1})$ therefore converges to zero in P_0 -probability. Note that we can write $(\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \text{vec}(\mathbf{B}_n)$ in terms of \mathbf{Z}_k as

$$\begin{aligned} &\sum_{k=1}^d \text{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \\ &\left((\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \left(\left(\Sigma^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right)^{-1} \right) \mathbf{Z}_k. \end{aligned}$$

Since the posterior distributions of \mathbf{Z}_k , $k = 1, \dots, d$ are independent, we therefore obtain

$$\|(\sqrt{n} (\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \text{vec}(\mathbf{B}_n) - \sqrt{n} \mathbf{J}_{\theta_0}^{-1}(\mathbf{f}_0)) - N(\boldsymbol{\mu}_n^{**}, \boldsymbol{\Sigma}_n^{**})\|_{TV} = o_{P_0}(1),$$

where $\boldsymbol{\mu}_n^{**}$ is given by

$$\begin{aligned} &\sqrt{n} \sum_{k=1}^d \text{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \\ &\left((\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \left(\left(\Sigma^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right)^{-1} \right) \\ &\times (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \sum_{j=1}^d \mathbf{Y}_j \sigma^{jk} - \mathbf{J}_{\theta_0}^{-1} \sqrt{n} \boldsymbol{\Gamma}(\mathbf{f}_0), \end{aligned}$$

and Σ_n^{**} is given by

$$\begin{aligned}
 & n \sum_{k=1}^d \text{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \\
 & \left((\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \left(\left(\Sigma^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right)^{-1} \right) \\
 & \times (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \times \text{rows}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)} \\
 & \left(\left(\left(\Sigma^{-1} + \frac{ck_n \mathbf{I}_d}{n} \right)^{1/2} \otimes \mathbf{I}_{k_n+m-1} \right)^{-1} (\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T)^T \right).
 \end{aligned}$$

Following the steps of the proof of Lemma 4, it can be shown that the eigenvalues of the matrix Σ_n^* mentioned in the statement of Theorem 4 are bounded away from zero and infinity. We can show that the Kullback-Leibler divergence of $N(\boldsymbol{\mu}_n^{**}, \Sigma_n^{**})$ from $N(\boldsymbol{\mu}_n^*, \sigma^2 \Sigma_n^*)$ converges in probability to zero by going through some routine matrix manipulations. Hence,

$$\left\| \left(\sqrt{n} (\mathbf{G}_{n,1}^T \dots \mathbf{G}_{n,d}^T) \text{vec}(\mathbf{B}_n) - \sqrt{n} \mathbf{J}_{\boldsymbol{\theta}_0}^{-1}(\mathbf{f}_0) \right) - N(\boldsymbol{\mu}_n^*, \sigma^2 \Sigma_n^*) \right\|_{TV} = o_{P_0}(1).$$

The above expression is equivalent to (7.6) of the proof of Theorem 2. Following steps similar to those of Theorem 2, we get (4.1). We obtain (4.2) by following the proof of Theorem 3. \square

Proof of Remark 7. By the definition of $\boldsymbol{\theta}$ in (4.5), both the expressions

$$\int_0^1 (D_{0,0,0,1} F_1(t, f_1(t), f_2(t), \boldsymbol{\theta}))^T (f_1'(t) - F_1(t, f_1(t), f_2(t), \boldsymbol{\theta})) w(t) dt$$

and

$$\int_0^1 (D_{0,0,0,1} F_1(t, f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}_0))^T (f_{1\boldsymbol{\theta}_0}'(t) - F_1(t, f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}_0)) w(t) dt$$

are zero vectors, and hence the difference

$$\begin{aligned}
 & \left(\int_0^1 (D_{0,0,0,1} F_1(t, f_1(t), f_2(t), \boldsymbol{\theta}))^T (f_1'(t) - F_1(t, f_1(t), f_2(t), \boldsymbol{\theta})) w(t) dt \right. \\
 & \quad - \int_0^1 (D_{0,0,0,1} F_1(t, f_1(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}))^T (f_1'(t) - F_1(t, f_1(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta})) w(t) dt \\
 & \quad + \left(\int_0^1 (D_{0,0,0,1} F_1(t, f_1(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}))^T (f_1'(t) - F_1(t, f_1(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta})) w(t) dt \right. \\
 & \quad \quad - \int_0^1 (D_{0,0,0,1} F_1(t, f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}_0))^T (f_{1\boldsymbol{\theta}_0}'(t) \\
 & \quad \quad \left. - F_1(t, f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t), \boldsymbol{\theta}_0)) w(t) dt \right)
 \end{aligned}$$

is the zero vector as well. Since $f_{2\theta_0}(t)$ is a known function of t , it can be absorbed in the first argument of F_1 which then becomes a function of three arguments. Then the second part of the left side above can be analyzed as in Theorem 1. To deal with the first part of left side it is sufficient to study the difference $f_2(\cdot) - f_{2\theta_0}(\cdot)$. Note that $f_2(t)$ can be written as

$$\begin{aligned} & \phi_n(t, f_{1\theta_0}(t), \theta_0) + (f_1(t) - f_{1\theta_0}(t))D_{0,1,0}\phi_n(t, f_{1\theta_0}(t), \theta_0) \\ & + (\theta - \theta_0)^T D_{0,0,1}\phi_n(t, f_{1\theta_0}(t), \theta_0) + O((f_1(t) - f_{1\theta_0}(t))^2) + O(\|\theta - \theta_0\|^2). \end{aligned}$$

By the accuracy of the RK4 method the difference $\sup_{t \in [0,1]} |\phi_n(t, f_{1\theta_0}(t), \theta_0) - f_{2\theta_0}(t)|$ is of the order h_n^4 . Now using Lemmas 2 to 4, we can conclude that

$$\|\sqrt{n}(\theta - \theta_0) - \mathbf{J}_{\theta_0}^{-1}\sqrt{n}(\Gamma(f_1) - \Gamma(f_{10}))\| \rightarrow 0$$

as $n \rightarrow \infty$. Now we can prove the Bernstein-von Mises theorem as before. \square

Appendix

A few lemmas presented below are instrumental in proving the main results. We denote by $E_0(\cdot)$ and $\text{Var}_0(\cdot)$ the expectation and variance operators respectively with respect to P_0 -probability. The following lemma helps to estimate the bias of the Bayes estimator.

Lemma 1. For $m \geq 2$ and k_n satisfying $n^{1/2m} \ll k_n \ll n$, for $r = 0, 1$, $\sup_{t \in [0,1]} |E_0(E(f^{(r)}(t)|\mathbf{Y})) - f_0^{(r)}(t)| = o(k_n^{r+1/2}/\sqrt{n})$.

Proof. We note that $f^{(r)}(t) = (\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}$ for $r = 0, 1$ with $\mathbf{N}^{(r)}(\cdot)$ standing for the r^{th} order derivative of $\mathbf{N}(\cdot)$. By (2.8),

$$E(f^{(r)}(t)|\mathbf{Y}) = \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} (\mathbf{N}^{(r)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}. \quad (\text{A.10})$$

Zhou and Wolfe (2000) showed that

$$(\mathbf{N}^{(r)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{N}^{(r)}(t) \asymp \frac{k_n^{2r+1}}{n}. \quad (\text{A.11})$$

Since $f_0^{(r)} \in C^{(m-r)}$, there exists a $\boldsymbol{\beta}^*$ (De Boor, 1978, Theorem XII.4, page 178) such that

$$\sup_{t \in [0,1]} |f_0^{(r)}(t) - (\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}^*| = O(k_n^{-(m-r)}). \quad (\text{A.12})$$

For any $t \in [0, 1]$, we can bound the absolute bias of $E(f_0^{(r)}(t)|\mathbf{Y})$ multiplied with $\sqrt{n}k_n^{-r-1/2}$ by

$$\sqrt{n}k_n^{-r-1/2} \sup_{t \in [0,1]} |E_0(E(f^{(r)}(t)|\mathbf{Y})) - f_0^{(r)}(t)|$$

$$\begin{aligned} &\leq \sqrt{n}k_n^{-r-1/2} \sup_{t \in [0,1]} \left| \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} (\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}^* - (\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}^* \right| \\ &\quad + \sqrt{n}k_n^{-r-1/2} \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} \\ &\quad \times \sup_{t \in [0,1]} |(\mathbf{N}^{(r)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T (f_0(\mathbf{x}) - \mathbf{X}_n \boldsymbol{\beta}^*)| \\ &\quad + \sqrt{n}k_n^{-r-1/2} \sup_{t \in [0,1]} |f_0^{(r)}(t) - (\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}^*|. \end{aligned}$$

Using the fact that $\sup_{t \in [0,1]} |(\mathbf{N}^{(r)}(t))^T \boldsymbol{\beta}^*| = O(1)$, first term on the right hand side of the previous inequality is of the order of $k_n^{-r+1/2}/\sqrt{n}$. Using the Cauchy-Schwarz inequality, (A.11) and (A.12), we can bound the second term up to a constant multiple by $\sqrt{n}k_n^{-m}$. The third term has the order of $\sqrt{n}k_n^{-m-1/2}$ as a result of (A.12). By the assumed conditions on m and k_n , the assertion holds. \square

The following lemma controls posterior variability.

Lemma 2. *If $m \geq 5$ and $n^{1/2m} \ll k_n \ll n^{1/8}$, then for $r = 0, 1$ and for all $\epsilon > 0$, $\Pi \left(\sqrt{n} \sup_{t \in [0,1]} |f^{(r)}(t) - f_0^{(r)}(t)|^2 > \epsilon | \mathbf{Y} \right) = o_{P_0}(1)$.*

Proof. By Markov’s inequality and the fact that $|a + b|^2 \leq 2(|a|^2 + |b|^2)$ for two real numbers a and b , we can bound $\Pi \left(\sup_{t \in [0,1]} \sqrt{n} |f^{(r)}(t) - f_0^{(r)}(t)|^2 > \epsilon | \mathbf{Y} \right)$ by

$$\begin{aligned} &2 \frac{\sqrt{n}}{\epsilon} \left\{ \sup_{t \in [0,1]} \left| \mathbb{E}(f^{(r)}(t) | \mathbf{Y}) - f_0^{(r)}(t) \right|^2 \right\} \\ &\quad + \mathbb{E} \left[\sup_{t \in [0,1]} \left| f^{(r)}(t) - \mathbb{E}(f^{(r)}(t) | \mathbf{Y}) \right|^2 | \mathbf{Y} \right] \}. \end{aligned} \tag{A.13}$$

Now we obtain the asymptotic orders of the expectations of the two terms inside the bracket above. We can bound the expectation of the first term by

$$\begin{aligned} &2 \sup_{t \in [0,1]} \left| \mathbb{E}_0(\mathbb{E}(f^{(r)}(t) | \mathbf{Y})) - f_0^{(r)}(t) \right|^2 \\ &\quad + 2\mathbb{E}_0 \left[\sup_{t \in [0,1]} \left| \mathbb{E}(f^{(r)}(t) | \mathbf{Y}) - \mathbb{E}_0(\mathbb{E}[f^{(r)}(t) | \mathbf{Y}]) \right|^2 \right]. \end{aligned} \tag{A.14}$$

Using (A.10), $\sup_{t \in [0,1]} |\mathbb{E}(f^{(r)}(t) | \mathbf{Y}) - \mathbb{E}_0(\mathbb{E}[f^{(r)}(t) | \mathbf{Y}])|$ can be bounded up to a constant multiple by

$$\max_{1 \leq k \leq n} \left| (\mathbf{N}^{(r)}(s_k))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\varepsilon} \right|$$

$$+ \sup_{t, t': |t-t'| \leq n^{-1}} \left| (\mathbf{N}^{(r)}(t) - \mathbf{N}^{(r)}(t'))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\varepsilon} \right|,$$

where $s_k = k/n$ for $k = 1, \dots, n$. Applying the mean value theorem to the second term of the above sum, we can bound the expression inside the E_0 expectation in the second term of (A.14) by a constant multiple of

$$\begin{aligned} & \max_{1 \leq k \leq n} \left| (\mathbf{N}^{(r)}(s_k))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\varepsilon} \right|^2 \\ & + \sup_{t \in [0,1]} \frac{1}{n^2} \left| (\mathbf{N}^{(r+1)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\varepsilon} \right|^2. \end{aligned} \quad (\text{A.15})$$

By the spectral decomposition, we can write $\mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T = \mathbf{P}^T \mathbf{D} \mathbf{P}$, where \mathbf{P} is an orthogonal matrix and \mathbf{D} is a diagonal matrix with $k_n + m - 1$ ones and $n - k_n - m + 1$ zeros in the diagonal. Now using the Cauchy-Schwarz inequality, we get

$$\begin{aligned} & E_0 \left(\max_{1 \leq k \leq n} \left| (\mathbf{N}^{(r)}(s_k))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\varepsilon} \right|^2 \right) \\ & \leq \max_{1 \leq k \leq n} \left\{ (\mathbf{N}^{(r)}(s_k))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{N}^{(r)}(s_k) \right\} E_0 (\boldsymbol{\varepsilon}^T \mathbf{P}^T \mathbf{D} \mathbf{P} \boldsymbol{\varepsilon}). \end{aligned}$$

By Lemma 5.4 of Zhou and Wolfe (2000) and the fact that $\mathbf{Var}_0(\mathbf{P}\boldsymbol{\varepsilon}) = \mathbf{Var}_0(\boldsymbol{\varepsilon})$, we can conclude that the expectation of the first term of (A.15) is $O(k_n^{2r+2}/n)$. Again applying the Cauchy-Schwarz inequality, the second term of (A.15) is bounded by

$$\sup_{t \in [0,1]} \left\{ \frac{1}{n^2} (\mathbf{N}^{(r+1)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{N}^{(r+1)}(t) \right\} (\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}),$$

whose expectation is of the order $n(k_n^{2r+3}/n^3) = k_n^{2r+3}/n^2$, using Lemma 5.4 of Zhou and Wolfe (2000). Thus, the expectation of the bound given by (A.15) is of the order k_n^{2r+2}/n . Combining it with (A.14) and Lemma 1, we get

$$E_0 \left[\sup_{t \in [0,1]} \left| E(f^{(r)}(t) | \mathbf{Y}) - f_0^{(r)}(t) \right|^2 \right] = O\left(\frac{k_n^{2r+2}}{n}\right). \quad (\text{A.16})$$

Define $\boldsymbol{\varepsilon}^* := (\mathbf{X}_n^T \mathbf{X}_n)^{1/2} \boldsymbol{\beta} - \left(1 + \frac{\sigma^2 c k_n}{n}\right)^{-1} (\mathbf{X}_n^T \mathbf{X}_n)^{-1/2} \mathbf{X}_n^T \mathbf{Y}$. Note that $\boldsymbol{\varepsilon}^* | \mathbf{Y} \sim N(\mathbf{0}, (\sigma^{-2} + c k_n/n)^{-1} \mathbf{I}_{k_n+m-1})$. Expressing $\sup_{t \in [0,1]} |f^{(r)}(t) - E[f^{(r)}(t) | \mathbf{Y}]|$ as $\sup_{t \in [0,1]} \left| (\mathbf{N}^{(r)}(t))^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1/2} \boldsymbol{\varepsilon}^* \right|$ and using the Cauchy-Schwarz inequality and Lemma 5.4 of Zhou and Wolfe (2000), the second term inside the bracket in (A.13) is seen to be $O(k_n^{2r+2}/n)$. Combining it with (A.13) and (A.16) and using $2r + 2 \leq 4$, we have the desired assertion. \square

Lemmas 1 and 2 can be used to prove the posterior consistency of $\boldsymbol{\theta}$ as shown in the next lemma.

Lemma 3. *If $m \geq 5$ and $n^{1/2m} \ll k_n \ll n^{1/8}$, then for all $\epsilon > 0$, $\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon | \mathbf{Y}) = o_{P_0}(1)$.*

Proof. By the triangle inequality, using the definition in (2.9),

$$\begin{aligned} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| &\leq \|f'(\cdot) - f'_0(\cdot)\|_w + \|F(\cdot, f(\cdot), \boldsymbol{\eta}) - F(\cdot, f_0(\cdot), \boldsymbol{\eta})\|_w \\ &\leq c_1 \sup_{t \in [0,1]} |f'(t) - f'_0(t)| + c_2 \sup_{t \in [0,1]} |f(t) - f_0(t)|, \end{aligned}$$

for appropriately chosen constants c_1 and c_2 . We denote the set $T_n = \{f : \sup_{t \in [0,1]} |f(t) - f_0(t)| \leq \tau_n, \sup_{t \in [0,1]} |f'(t) - f'_0(t)| \leq \tau_n\}$ for some $\tau_n \rightarrow 0$. By Lemma 2, there exists such a sequence $\{\tau_n\}$ so that $\Pi(T_n^c | \mathbf{Y}) = o_{P_0}(1)$. Hence for $f \in T_n$,

$$\sup_{\boldsymbol{\eta} \in \Theta} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| \leq (c_1 + c_2)\tau_n = o(1)$$

Therefore, for any $\delta > 0$, $\Pi(\sup_{\boldsymbol{\eta} \in \Theta} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta | \mathbf{Y}) = o_{P_0}(1)$. By assumption (3.1), for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon$ there exists a $\delta > 0$ such that

$$\begin{aligned} \delta < R_{f_0}(\boldsymbol{\theta}) - R_{f_0}(\boldsymbol{\theta}_0) &\leq R_{f_0}(\boldsymbol{\theta}) - R_f(\boldsymbol{\theta}) + R_f(\boldsymbol{\theta}_0) - R_{f_0}(\boldsymbol{\theta}_0) \\ &\leq 2 \sup_{\boldsymbol{\eta} \in \Theta} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})|, \end{aligned}$$

since $R_f(\boldsymbol{\theta}) \leq R_f(\boldsymbol{\theta}_0)$. Consequently,

$$\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon | \mathbf{Y}) \leq \Pi\left(\sup_{\boldsymbol{\eta} \in \Theta} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta/2 | \mathbf{Y}\right) = o_{P_0}(1). \quad \square$$

The asymptotic behavior of the posterior distribution of $\sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{\Gamma}(\mathbf{f}) - \boldsymbol{\Gamma}(\mathbf{f}_0))$ is given by the next lemma.

Lemma 4. *Under the conditions of Theorem 2, on a set in the sample space with high true probability, the posterior distribution of $\sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{\Gamma}(\mathbf{f}) - \boldsymbol{\Gamma}(\mathbf{f}_0))$ assigns most of its mass inside a large compact set.*

Proof. Note that

$$\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}(\mathbf{f}) = \sum_{j=1}^d \mathbf{G}_{n,j}^T \boldsymbol{\beta}_j, \quad \mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}(\mathbf{f}_0) = \sum_{j=1}^d \int_0^1 \mathbf{A}_{,j}(t) f_{j0}(t) dt,$$

where $\mathbf{A}_{,j}(t)$ denotes the j^{th} column of the matrix $\mathbf{A}(t)$ as defined in Remark 5 for $j = 1, \dots, d$. In order to prove the assertion, we show that $\mathbf{Var}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y})$ and $\mathbf{Var}_0(\mathbf{E}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y}))$ have all eigenvalues of the order n^{-1} and

$$\max_{1 \leq k \leq p} \left| [\mathbf{E}_0(\mathbf{E}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y}))]_k - \int_0^1 A_{k,j}(t) f_{j0}(t) dt \right| = o(n^{-1/2}),$$

for $k = 1, \dots, p$, $j = 1, \dots, d$, where $A_{k,j}(t)$ is the $(k, j)^{th}$ element of the matrix $\mathbf{A}(t)$. Let us fix $j \in \{1, \dots, d\}$. We note that

$$\mathbf{E}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y}) = \left(1 + \frac{ck_n \sigma^2}{n}\right)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_j.$$

Hence,

$$\mathbf{Var}_0(\mathbf{E}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y})) = \sigma_0^2 \left(1 + \frac{\sigma^2 ck_n}{n}\right)^{-2} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}.$$

Also note that

$$\mathbf{Var}(\mathbf{G}_{n,j}^T \boldsymbol{\beta}_j | \mathbf{Y}) = \sigma^2 \left(1 + \frac{\sigma^2 ck_n}{n}\right)^{-1} \mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}.$$

If $A_{k,j}(\cdot) \in C^{m^*}((0, 1))$ for some $1 \leq m^* < m$, then by equation (2) of De Boor (1978, page 167), we have $\sup\{|A_{k,j}(t) - \tilde{A}_{k,j}(t)| : t \in [0, 1]\} = O(k_n^{-1})$, where $\tilde{A}_{k,j}(\cdot) = \boldsymbol{\alpha}_{k,j}^T \mathbf{N}(\cdot)$ and $\boldsymbol{\alpha}_{k,j}^T = (A_{k,j}(t_1^*), \dots, A_{k,j}(t_{k_n+m-1}^*))$ with appropriately chosen $t_1^*, \dots, t_{k_n+m-1}^*$. We can express $\mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$ as

$$\begin{aligned} & (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j}) + \tilde{\mathbf{G}}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j}) \\ & + \tilde{\mathbf{G}}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \tilde{\mathbf{G}}_{n,j} + (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \tilde{\mathbf{G}}_{n,j} \end{aligned}$$

where $[\tilde{\mathbf{G}}_{n,j}^T]_k = \int_0^1 \tilde{A}_{k,j}(t) (\mathbf{N}(t))^T dt$ for $k = 1, \dots, p$. Let $\tilde{\mathbf{A}} = ((\tilde{A}_{k,j}))$. We study the asymptotic orders of the eigenvalues of the matrices $\tilde{\mathbf{G}}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \tilde{\mathbf{G}}_{n,j}$ and $(\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})$. Note that

$$\boldsymbol{\alpha}_{k,j}^T \int_0^1 \mathbf{N}(t) \mathbf{N}^T(t) dt \boldsymbol{\alpha}_{k,j} = \int_0^1 \tilde{A}_{k,j}^2(t) dt \asymp \|\boldsymbol{\alpha}_{k,j}\|^2 k_n^{-1},$$

by Lemma 6.1 of Zhou et al. (1998) implying that the eigenvalues of the matrix $\int_0^1 \mathbf{N}(t) (\mathbf{N}(t))^T dt$ are of order k_n^{-1} . Since the eigenvalues of $(\mathbf{X}_n^T \mathbf{X}_n/n)$ are of the order k_n^{-1} (Zhou et al., 1998), we have

$$\begin{aligned} & \text{maxeig} \left(\tilde{\mathbf{G}}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \tilde{\mathbf{G}}_{n,j} \right) \\ & \lesssim \frac{k_n}{n} \text{maxeig} \left(\tilde{\mathbf{G}}_{n,j}^T \tilde{\mathbf{G}}_{n,j} \right) \\ & = \frac{k_n}{n} \text{maxeig} \left(\int_0^1 \tilde{\mathbf{A}}_j(t) \mathbf{N}^T(t) dt \int_0^1 \mathbf{N}(t) (\tilde{\mathbf{A}}_j(t))^T dt \right) \\ & = \frac{k_n}{n} \text{maxeig} \left((\boldsymbol{\alpha}_{1,j} \cdots \boldsymbol{\alpha}_{p,j})^T \left(\int_0^1 \mathbf{N}(t) \mathbf{N}^T(t) dt \right)^2 (\boldsymbol{\alpha}_{1,j} \cdots \boldsymbol{\alpha}_{p,j}) \right) \\ & \lesssim \frac{1}{nk_n} \text{maxeig}((\boldsymbol{\alpha}_{k,j}^T \boldsymbol{\alpha}_{l,j})) \end{aligned}$$

$$\begin{aligned} &\asymp \frac{1}{n} \text{maxeig}(\langle\langle A_{k,j}(\cdot), A_{l,j}(\cdot) \rangle\rangle) \\ &= \frac{1}{n} \text{maxeig}(\mathbf{B}_j) \asymp \frac{1}{n}. \end{aligned}$$

Similarly, it can be shown that $\text{mineig}(\tilde{\mathbf{G}}_{n,j}^T(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \tilde{\mathbf{G}}_{n,j}) \gtrsim n^{-1}$. Let us denote by $\mathbf{1}_{k_n+m-1}$ the $k_n + m - 1$ -component vector with all elements 1. Then for $k = 1, \dots, p$,

$$\begin{aligned} &\left[(\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j}) \right]_{k,k} \\ &= \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) (\mathbf{N}(t))^T dt (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \\ &\quad \times \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) (\mathbf{N}(t)) dt \\ &= \frac{1}{n} \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) (\mathbf{N}(t))^T dt (\mathbf{X}_n^T \mathbf{X}_n / n)^{-1} \\ &\quad \times \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) \mathbf{N}(t) dt \\ &\asymp \frac{k_n}{n} \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) (\mathbf{N}(t))^T dt \int_0^1 (A_{k,j}(t) - \tilde{A}_{k,j}(t)) \mathbf{N}(t) dt \\ &\lesssim \frac{1}{nk_n}, \end{aligned}$$

the last step following by the application of the Cauchy-Schwarz inequality and the facts that $\sup\{|A_{k,j}(t) - \tilde{A}_{k,j}(t)| : t \in [0, 1]\} = O(k_n^{-1})$ and $\int_0^1 \|\mathbf{N}(t)\|^2 dt \leq 1$. Thus, the eigenvalues of $(\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} (\mathbf{G}_{n,j} - \tilde{\mathbf{G}}_{n,j})$ are of the order $(nk_n)^{-1}$ or less. Hence, the eigenvalues of $\mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{G}_{n,j}$ are of the order n^{-1} .

As in the proof of Lemma 1, we can write for the β_j^* given in (A.12),

$$\begin{aligned} &\sqrt{n} \left| [\mathbf{E}_0(\mathbf{E}(\mathbf{G}_{n,j}^T \beta_j | \mathbf{Y}))]_k - \int_0^1 A_{k,j}(t) f_{j0}(t) dt \right| \\ &\leq \sqrt{n} \left| \left(1 + \frac{ck_n \sigma^2}{n} \right)^{-1} [\mathbf{G}_{n,j}^T \beta_j^*]_k - [\mathbf{G}_{n,j}^T \beta_j^*]_k \right| \\ &\quad + \sqrt{n} \left(1 + \frac{ck_n \sigma^2}{n} \right)^{-1} \left| [\mathbf{G}_{n,j}^T (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T (f_{j0}(\mathbf{x}) - \mathbf{X}_n \beta_j^*)]_k \right| \\ &\quad + \sqrt{n} \left| \int_0^1 A_{k,j}(t) f_{j0}(t) dt - [\mathbf{G}_{n,j}^T \beta_j^*]_k \right|, \end{aligned}$$

where $[\mathbf{G}_{n,j}^T \beta_j^*]_k = \int_0^1 A_{k,j}(t) f_j^*(t) dt$ and $f_j^*(t) = \mathbf{N}^T(t) \beta_j^*$ for $k = 1, \dots, p$. Proceeding in the same way as in the proof of Lemma 1, we can show that each term on the right hand side of the above equation converges to zero. Hence, the proof. \square

References

- R. M. ANDERSON and R. M. MAY (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.
- Y. BARD (1974). *Nonlinear Parameter Estimation*. Academic Press New York. [MR0326870](#)
- D. BONTEMPS (2011). Bernstein-von mises theorems for gaussian regression with increasing number of regressors. *The Annals of Statistics* **39**, 2557–2584. [MR2906878](#)
- N. J. BRUNEL (2008). Parameter estimation of ode’s via nonparametric estimators. *Electronic Journal of Statistics* **2**, 1242–1267. [MR2471285](#)
- N. J. BRUNEL, Q. CLAIRON, and F. D’ALCHÉ BUC (2014). Parametric estimation of ordinary differential equations with orthogonality conditions. *Journal of the American Statistical Association* **109**, 173–185. [MR3180555](#)
- D. CAMPBELL and R. J. STEELE (2012). Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing* **22**, 429–443. [MR2865027](#)
- D. A. CAMPBELL (2007). *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models*. ProQuest. [MR2711737](#)
- T. CHEN, H. L. HE, G. M. CHURCH, et al. (1999). Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, Volume **4**, pp. 4.
- O. CHKREBTII, D. A. CAMPBELL, M. A. GIROLAMI, and B. CALDERHEAD (2013). Bayesian uncertainty quantification for differential equations. *arXiv preprint arXiv:1306.2365*.
- I. DATNER and S. GUGUSHVILI (2015). Accelerated least squares estimation for systems of ordinary differential equations. *arXiv preprint arXiv:1503.07973*.
- C. DE BOOR (1978). *A Practical Guide to Splines*, Volume **27**. Springer-Verlag, New York. [MR0507062](#)
- T. A. DEAN and S. S. SINGH (2011). Asymptotic behaviour of approximate bayesian estimators. *arXiv preprint arXiv:1105.3655*.
- V. B. DOMSELAAR and P. HEMKER (1975). Nonlinear parameter estimation in initial value problems. *Stichting Mathematisch Centrum. Numerieke Wiskunde*, 1–49.
- J. GABRIELSSON and D. WEINER (2006). *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*. Swedish Pharmaceutical Press.
- A. GELMAN, F. BOIS, and J. JIANG (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**, 1400–1412.
- M. GIROLAMI (2008). Bayesian inference for differential equations. *Theoretical Computer Science* **408**, 4–16. [MR2460604](#)

- S. GUGUSHVILI and C. A. KLAASSEN (2012). \sqrt{n} -consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing *Bernoulli* **18**, 1061–1098. [MR2948913](#)
- E. HAIRER, S. NØRSETT, and G. WANNER (1993). *Solving Ordinary Differential Equations 1: Nonstiff Problems*. Springer-Verlag, New York, Inc. [MR1227985](#)
- J. HENDERSON and G. MICHAILIDIS (2014). Network reconstruction using non-parametric additive ode models. *PloS one* **9**(4), 94003.
- J. JAEGER (2012). *Functional estimation in systems defined by differential equations using Bayesian smoothing methods*. Ph. D. thesis, Université Catholique de Louvain.
- B. KLEIJN and A. VAN DER VAART (2012). The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics* **6**, 354–381. [MR2988412](#)
- K. LEVENBERG (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics* **2**, 164–168. [MR0010666](#)
- D. W. MARQUARDT (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics* **11**, 431–441. [MR0153071](#)
- R. M. MATTHEIJ and J. MOLENAAR (2002). Ordinary differential equations in theory and practice. Reprint of (1996) original. *Classics in Applied Mathematics*. [MR1946758](#)
- M. NOWAK and R. M. MAY (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press. [MR2009143](#)
- X. QI and H. ZHAO (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics* **38**, 435–481. [MR2589327](#)
- J. O. RAMSAY, G. HOOKER, D. CAMPBELL, and J. CAO (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 741–796. [MR2368570](#)
- S. ROGERS, R. KHANIN, and M. GIROLAMI (2007). Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics* **8**(Suppl 2), S2.
- W.-H. STEEB (2006). *Problems and Solutions in Introductory and Advanced Matrix Calculus*. World Scientific. [MR2311921](#)
- J. VARAH (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing* **3**, 28–46. [MR0651865](#)
- E. O. VOIT and J. ALMEIDA (2004). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* **20**, 1670–1681.
- H. WU, H. XUE, and A. KUMAR (2012). Numerical discretization-based estimation methods for ordinary differential equation models via penalized spline smoothing with applications in biomedical research. *Biometrics* **68**, 344–352. [MR2959600](#)
- H. XUE, H. MIAO, and H. WU (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by con-

- sidering both numerical error and measurement error. *The Annals of Statistics* **38**, 2351. [MR2676892](#)
- S. ZHOU, X. SHEN, and D. WOLFE (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26**, 1760–1782. [MR1673277](#)
- S. ZHOU and D. A. WOLFE (2000). On derivative estimation in spline regression. *Statistica Sinica* **10**, 93–108. [MR1742102](#)