Bernoulli **23**(4B), 2017, 3537–3570
DOI: 10.3150/16-BEJ856

Efficient Bayesian estimation and uncertainty quantification in ordinary differential equation models

PRITHWISH BHAUMIK[1] and SUBHASHIS GHOSAL[2]

[1]*Department of Statistics and Data Sciences, The University of Texas at Austin, GDC 7.502, 2317 Speedway D9800, Austin, TX 78712-1823, USA. E-mail: prithwish.bhaumik@utexas.edu*
[2]*Department of Statistics, North Carolina State University, 4276 SAS Hall, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, USA. E-mail: sghosal@ncsu.edu*

Often the regression function is specified by a system of ordinary differential equations (ODEs) involving some unknown parameters. Typically analytical solution of the ODEs is not available, and hence likelihood evaluation at many parameter values by numerical solution of equations may be computationally prohibitive. Bhaumik and Ghosal (*Electron. J. Stat.* **9** (2015) 3124–3154) considered a Bayesian two-step approach by embedding the model in a larger nonparametric regression model, where a prior is put through a random series based on B-spline basis functions. A posterior on the parameter is induced from the regression function by minimizing an integrated weighted squared distance between the derivative of the regression function and the derivative suggested by the ODEs. Although this approach is computationally fast, the Bayes estimator is not asymptotically efficient. In this paper, we suggest a modification of the two-step method by directly considering the distance between the function in the nonparametric model and that obtained from a four stage Runge–Kutta (RK4) method. We also study the asymptotic behavior of the posterior distribution of $\boldsymbol{\theta}$ based on an approximate likelihood obtained from an RK4 numerical solution of the ODEs. We establish a Bernstein–von Mises theorem for both methods which assures that Bayesian uncertainty quantification matches with the frequentist one and the Bayes estimator is asymptotically efficient.

Keywords: approximate likelihood; Bayesian inference; Bernstein–von Mises theorem; ordinary differential equation; Runge–Kutta method; spline smoothing

1. Introduction

Differential equations are encountered in various branches of science such as in genetics [6], viral dynamics of infectious diseases [1,19], pharmacokinetics and pharmacodynamics (PKPD) [8]. In many cases these equations do not lead to any explicit solution. A popular example is the Lotka–Volterra equations, also known as predator-prey equations. The rates of change of the prey and predator populations are given by the equations

$$\begin{aligned}\frac{df_{1\boldsymbol{\theta}}(t)}{dt} &= \theta_1 f_{1\boldsymbol{\theta}}(t) - \theta_2 f_{1\boldsymbol{\theta}}(t) f_{2\boldsymbol{\theta}}(t),\\ \frac{df_{2\boldsymbol{\theta}}(t)}{dt} &= -\theta_3 f_{2\boldsymbol{\theta}}(t) + \theta_4 f_{1\boldsymbol{\theta}}(t) f_{2\boldsymbol{\theta}}(t), \qquad t \in [0,1],\end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ and $f_{1\boldsymbol{\theta}}(t)$ and $f_{2\boldsymbol{\theta}}(t)$ denote the prey and predator populations at time t, respectively. These models can be put in a regression model $\mathbf{Y} = \mathbf{f}_{\boldsymbol{\theta}}(t) + \boldsymbol{\varepsilon}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, where $\mathbf{f}_{\boldsymbol{\theta}}(\cdot)$ satisfies the ODE

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

here $\mathbf{F}$ is a known appropriately smooth vector valued function and $\boldsymbol{\theta}$ is a parameter vector controlling the regression function.

The nonlinear least squares (NLS) is the usual way to estimate the unknown parameters provided that the analytical solution of the ODE is available, which is not the case in most real applications. The 4-stage Runge–Kutta algorithm (RK4) ([13], page 134 and [18], page 53) can solve (1.1) numerically. The parameters can be estimated by applying NLS in the next step. Xue, Miao and Wu [28] studied the asymptotic properties of the estimator. They used differential evolution method [24], scatter search method and sequential quadratic programming [22] method for the NLS part and established the strong consistency, $\sqrt{n}$-consistency and asymptotic normality of the estimator. The estimator turns out to be asymptotically efficient, but this approach is computationally intensive.

In the generalized profiling procedure [21], a linear combination of basis functions is used to obtain an approximate solution. A penalized optimization is used to estimate the coefficients of the basis functions. The estimated $\boldsymbol{\theta}$ is defined as the maximizer of a data dependent fitting criterion involving the estimated coefficients. The statistical properties of the estimator obtained from this approach were explored in the works of [20]. This method is also asymptotically efficient, but has a high computational cost.

Varah [26] used a two-step procedure where the state variables are approximated by cubic spline in the first step. In the second step, the parameters are estimated by minimizing the sum of squares of difference between the non-parametrically estimated derivative and the derivatives suggested by the ODEs at the design points. Thus, the ODE model is embedded in the nonparametric regression model. This method is very fast and independent of the initial or boundary conditions. Brunel [3] did a modification by replacing the sum of squares by a weighted integral and obtained the asymptotic normality of the estimator. Gugushvili and Klaassen [12] followed the approach of [3], but used kernel smoothing instead of spline and established $\sqrt{n}$-consistency of the estimator. Wu, Xue and Kumar [27] used penalized smoothing spline in the first step and numerical derivatives of the nonparametrically estimated functions. Brunel, Clairon and d'Alché-Buc [4] used nonparametric smoothing and a set of orthogonality conditions to estimate the parameters. But the major drawback of the two-step estimation methods is that these are not asymptotically efficient.

ODE models in Bayesian framework was considered in the works of [9,23] and [11]. They obtained an approximate likelihood by solving the ODEs numerically. Using the prior assigned on $\boldsymbol{\theta}$, MCMC technique was used to generate samples from the posterior. This method also has high computational complexity. Campbell and Steele [5] proposed the smooth functional tempering approach which utilizes the generalized profiling approach [21] and the parallel tempering algorithm. Jaeger [15] also used the generalized profiling in Bayesian framework. Bhaumik and Ghosal [2] considered the Bayesian analog of the two-step method suggested by [3], putting prior

on the coefficients of the B-spline basis functions and induced a posterior on Θ. They established a Bernstein–von Mises theorem for the posterior distribution of $\boldsymbol{\theta}$ with $n^{-1/2}$ contraction rate.

In this paper we propose two separate approaches. We use Gaussian distribution as the working model for error, although the true distribution may be different. The first approach involves assigning a direct prior on $\boldsymbol{\theta}$ and then constructing the posterior of $\boldsymbol{\theta}$ using an approximate likelihood function constructed using the approximate solution $\mathbf{f}_{\boldsymbol{\theta},r}(\cdot)$ obtained from RK4. Here r is the number of grid points used. When r is sufficiently large, the approximate likelihood is expected to behave like the actual likelihood. We call this method Runge–Kutta sieve Bayesian (RKSB) method. In the second approach we define $\boldsymbol{\theta}$ as $\arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} \int_0^1 \|\boldsymbol{\beta}^T\mathbf{N}(\cdot) - \mathbf{f}_{\boldsymbol{\eta},r}(\cdot)\|^2 w(t)\,dt$ for an appropriate weight function $w(\cdot)$ on $[0,1]$, where the posterior distribution of $\boldsymbol{\beta}$ is obtained in the nonparametric spline model and $\mathbf{N}(\cdot)$ is the B-spline basis vector. We call this approach Runge–Kutta two-step Bayesian (RKTB) method. Thus, this approach is similar in spirit to [2]. Similar to [2], prior is assigned on the coefficients of the B-spline basis and the posterior of $\boldsymbol{\theta}$ is induced from the posterior of the coefficients. But the main difference lies in the way of extending the definition of parameter. Instead of using deviation from the ODE, we consider the distance between function in the nonparametric model and RK4 approximation of the model. Ghosh and Goyal [10] considered Euler's approximation to construct the approximate likelihood and then drew posterior samples. In the same paper they also provided a non-Bayesian method by estimating $\boldsymbol{\theta}$ by minimizing the sum of squares of the difference between the spline fitting and the Euler approximation at the grid points. However they did not explore the theoretical aspects of those methods. We shall show both RKSB and RKTB lead to Bernstein–von Mises Theorem with dispersion matrix inverse of Fisher information and hence both the proposed Bayesian methods are asymptotically efficient. This was not the case for the two step-Bayesian approach [2]. Bernstein–von Mises Theorem implies that credible intervals have asymptotically correct frequentist coverage. The computational cost of the two-step Bayesian method [2] is the least, RKTB is more computationally involved and RKSB is the most computationally expensive.

The paper is organized as follows. Section 2 contains the description of the notations and some preliminaries of Runge–Kutta method. The model assumptions and prior specifications are given in Section 3. The main results are given in Section 4. In Section 5 we carry out a simulation study. Proofs of the main results are given in Section 6. Section 7 contains the proofs of the technical lemmas. The Appendix is provided in the last section.

2. Notations and preliminaries

We describe a set of notations to be used in this paper. Boldfaced letters are used to denote vectors and matrices. The identity matrix of order p is denoted by $\mathbf{I}_p$. We use the symbols maxeig($\mathbf{A}$) and mineig($\mathbf{A}$) to denote the maximum and minimum eigenvalues of the matrix $\mathbf{A}$, respectively. The L_2 norm of a vector $\mathbf{x}\in\mathbb{R}^p$ is given by $\|\mathbf{x}\| = (\sum_{i=1}^p x_i^2)^{1/2}$. The notation $f^{(r)}(\cdot)$ stands for the rth order derivative of a function $f(\cdot)$, that is, $f^{(r)}(t) = \frac{d^r}{dt^r} f(t)$. For the function $\boldsymbol{\theta}\mapsto f_{\boldsymbol{\theta}}(x)$, the notation $\dot{f}_{\boldsymbol{\theta}}(x)$ implies $\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x)$. Similarly, we denote $\ddot{f}_{\boldsymbol{\theta}}(x) = \frac{\partial^2}{\partial\boldsymbol{\theta}^2} f_{\boldsymbol{\theta}}(x)$. A vector valued function is represented by the boldfaced symbol $\mathbf{f}(\cdot)$. We use the notation $f(\mathbf{x})$ to denote the vector $(f(x_1),\ldots,f(x_p))^T$ for a real-valued function $f:[0,1]\to\mathbb{R}$ and a vector $\mathbf{x}\in\mathbb{R}^p$. Let

us define $\|\mathbf{f}\|_g = (\int_0^1 \|\mathbf{f}(t)\|^2 g(t)\,dt)^{1/2}$ for $\mathbf{f}:[0,1]\mapsto\mathbb{R}^p$ and $g:[0,1]\mapsto[0,\infty)$. The weighted inner product with the corresponding weight function $g(\cdot)$ is denoted by $\langle\cdot,\cdot\rangle_g$. For numerical sequences a_n and b_n, both $a_n = o(b_n)$ and $a_n \ll b_n$ mean $a_n/b_n \to 0$ as $n\to\infty$. Similarly, we define $a_n \gg b_n$. The notation $a_n = O(b_n)$ is used to indicate that a_n/b_n is bounded. The notation $a_n \asymp b_n$ stands for both $a_n = O(b_n)$ and $b_n = O(a_n)$, while $a_n \lesssim b_n$ means $a_n = O(b_n)$. The symbol $o_P(1)$ stands for a sequence of random vectors converging in P-probability to zero, whereas $O_P(1)$ stands for a stochastically bounded sequence of random vectors. Given a sample $\{X_i : i = 1,\dots,n\}$ and a measurable function $\psi(\cdot)$, we define $\mathbb{P}_n\psi = n^{-1}\sum_{i=1}^n \psi(X_i)$. The symbols $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$ stand for the mean and variance respectively of a random variable, or the mean vector and the variance-covariance matrix of a random vector. We use the notation $\mathbb{G}_n\psi$ to denote $\sqrt{n}(\mathbb{P}_n\psi - \mathrm{E}\psi)$. For a measure P, the notation $P^{(n)}$ implies the joint measure of a random sample $X_1,\dots,X_n$ coming from the distribution P. Similarly, we define p and $p^{(n)}$ for the corresponding densities. The total variation distance between the probability measures P and Q defined on $\mathbb{R}^p$ is given by $\|P-Q\|_{\mathrm{TV}} = \sup_{B\in\mathcal{R}^p}|P(B)-Q(B)|$, $\mathcal{R}^p$ being the Borel σ-field on $\mathbb{R}^p$. Given an open set E, the symbol $C^m(E)$ stands for the class of functions defined on E having first m continuous partial derivatives with respect to its arguments. For a set A, the notation $\mathbb{1}\{A\}$ stands for the indicator function for belonging to A. The symbol $:=$ means equality by definition. For two real numbers a and b, we use the notation $a\wedge b$ to denote the minimum of a and b. Similarly, we denote $a\vee b$ to be the maximum of a and b.

Given r equispaced grid points $a_1 = 0, a_2, \dots, a_r$ with common difference h and an initial condition $\mathbf{f}_{\boldsymbol{\theta}}(0) = \mathbf{y}_0$, Euler's method ([14], page 9) computes the approximate solution as $\mathbf{f}_{\boldsymbol{\theta},r}(a_{k+1}) = \mathbf{f}_{\boldsymbol{\theta},r}(a_k) + h\mathbf{F}(a_k, \mathbf{f}_{\boldsymbol{\theta},r}(a_k), \boldsymbol{\theta})$ for $k = 1,2,\dots,r-1$. The RK4 method ([14], page 68) is an improvement over Euler's method. Let us denote

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{F}\big(a_k, \mathbf{f}_{\boldsymbol{\theta},r}(a_k), \boldsymbol{\theta}\big),\\
\mathbf{k}_2 &= \mathbf{F}\big(a_k + h/2, \mathbf{f}_{\boldsymbol{\theta},r}(a_k) + h/2\mathbf{k}_1, \boldsymbol{\theta}\big),\\
\mathbf{k}_3 &= \mathbf{F}\big(a_k + h/2, \mathbf{f}_{\boldsymbol{\theta},r}(a_k) + h/2\mathbf{k}_2, \boldsymbol{\theta}\big),\\
\mathbf{k}_4 &= \mathbf{F}\big(a_k + h, \mathbf{f}_{\boldsymbol{\theta},r}(a_k) + h\mathbf{k}_3, \boldsymbol{\theta}\big).
\end{aligned}$$

Then we obtain $\mathbf{f}_{\boldsymbol{\theta},r}(a_{k+1})$ from $\mathbf{f}_{\boldsymbol{\theta},r}(a_k)$ as $\mathbf{f}_{\boldsymbol{\theta},r}(a_{k+1}) = \mathbf{f}_{\boldsymbol{\theta},r}(a_k) + h/6(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)$. By the proof of Theorem 3.3 of [14], page 124, we have

$$\sup_{t\in[0,1]} \big\|\mathbf{f}_{\boldsymbol{\theta}}(t) - \mathbf{f}_{\boldsymbol{\theta},r}(t)\big\| = O\big(r^{-4}\big), \qquad \sup_{t\in[0,1]} \big\|\dot{\mathbf{f}}_{\boldsymbol{\theta}}(t) - \dot{\mathbf{f}}_{\boldsymbol{\theta},r}(t)\big\| = O\big(r^{-4}\big). \tag{2.1}$$

3. Model assumptions and prior specifications

Now we formally describe the model. For the sake of simplicity, we assume the response to be one dimensional. The extension to the multidimensional case is straight forward. The proposed model is given by

$$Y_i = f_{\boldsymbol{\theta}}(X_i) + \varepsilon_i, \qquad i = 1,\dots,n, \tag{3.1}$$

where $\boldsymbol{\theta} \subseteq \Theta$, which is a compact subset of $\mathbb{R}^p$. The function $f_{\boldsymbol{\theta}}(\cdot)$ satisfies the ODE given by

$$\frac{df_{\boldsymbol{\theta}}(t)}{dt} = F\big(t, f_{\boldsymbol{\theta}}(t), \boldsymbol{\theta}\big), \qquad t \in [0,1]. \tag{3.2}$$

Let for a fixed $\boldsymbol{\theta}$, $F \in C^{m-1}((0,1), \mathbb{R})$ for some integer $m \geq 1$. Then, by successive differentiation we have $f_{\boldsymbol{\theta}} \in C^m((0,1))$. By the implied uniform continuity, the function and its several derivatives can be uniquely extended to continuous functions on $[0,1]$. We also assume that $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(x)$ is continuous in $\boldsymbol{\theta}$. The true regression function $f_0(\cdot)$ does not necessarily lie in $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. We assume that $f_0 \in C^m([0,1])$. Let ε_i are identically and independently distributed with mean zero and finite moment generating function for $i = 1, \ldots, n$. Let the common variance be σ_0^2. We use $N(0, \sigma^2)$ as the working model for the error, which may be different from the true distribution. We treat σ^2 as an unknown parameter and assign an inverse gamma prior on σ^2 with shape and scale parameters a and b, respectively. Additionally it is assumed that $X_i \overset{\text{i.i.d.}}{\sim} G$ with density g. The approximate solution to (1.1) is given by $f_{\theta,r}$, where $r = r_n$ is the number of grid points, which is chosen so that

$$r_n \gg n^{1/8}. \tag{3.3}$$

Let us denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (X_1, \ldots, X_n)^T$. The true joint distribution of (X_i, ε_i) is denoted by P_0. Now we describe the two different approaches of inference on $\boldsymbol{\theta}$ used in this paper.

3.1. Runge–Kutta sieve Bayesian method (RKSB)

For RKSB, we denote $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \sigma^2)$. The approximate likelihood of the sample $\{(X_i, Y_{i,}) : i = 1, \ldots, n\}$ is given by $L_n^*(\boldsymbol{\gamma}) = \prod_{i=1}^n p_{\boldsymbol{\gamma},n}(X_i, Y_{i,})$, where

$$p_{\boldsymbol{\gamma},n}(X_i, Y_i) = (\sqrt{2\pi}\sigma)^{-1} \exp\big\{-\big(2\sigma^2\big)^{-1}\big|Y_i - f_{\boldsymbol{\theta},r_n}(X_i)\big|^2\big\} g(X_i). \tag{3.4}$$

We also denote

$$p_{\boldsymbol{\gamma}}(X_i, Y_i) = (\sqrt{2\pi}\sigma)^{-1} \exp\big\{-\big(2\sigma^2\big)^{-1}\big|Y_i - f_{\boldsymbol{\theta}}(X_i)\big|^2\big\} g(X_i). \tag{3.5}$$

The true parameter $\boldsymbol{\gamma}_0 := (\boldsymbol{\theta}_0, \sigma_*^2)$ is defined as

$$\boldsymbol{\gamma}_0 = \arg\max_{\boldsymbol{\gamma}} P_0 \log p_{\boldsymbol{\gamma}},$$

which takes into account the situation when $f_{\boldsymbol{\theta}_0}$ is the true regression function, $\boldsymbol{\theta}_0$ being the true parameter. We denote by $\ell_{\boldsymbol{\gamma}}$ and $\ell_{\boldsymbol{\gamma},n}$ the log-likelihoods with respect to (3.5) and (3.4), respectively. We make the following assumptions.

(A1) The parameter vector $\boldsymbol{\gamma}_0$ is the unique maximizer of the right-hand side above.

(A2) The sub-matrix of the Hessian matrix of $-P_0 \log p_{\boldsymbol{\gamma}}$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ given by

$$\int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial \boldsymbol{\theta}} \big(\dot{f}_{\boldsymbol{\theta}}^T(t) \big(f_0(t) - f_{\boldsymbol{\theta}_0}(t) \big) \big) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) g(t)\,dt \tag{3.6}$$

is positive definite.

(A3) The prior measure on Θ has a Lebesgue-density continuous and positive on a neighborhood of $\boldsymbol{\theta}_0$.

(A4) The prior distribution of $\boldsymbol{\theta}$ is independent of that of σ^2.

By (A1) we get

$$\begin{gathered} \int_0^1 \dot{f}_{\boldsymbol{\theta}_0}^T(t) \big(f_0(t) - f_{\boldsymbol{\theta}_0}(t) \big) g(t)\,dt = \mathbf{0}, \\ \sigma_*^2 = \sigma_0^2 + \int_0^1 \big| f_0(t) - f_{\boldsymbol{\theta}_0}(t) \big|^2 g(t)\,dt. \end{gathered} \tag{3.7}$$

The joint prior measure of $\boldsymbol{\gamma}$ is denoted by Π with corresponding density π. We obtain the posterior of $\boldsymbol{\gamma}$ using the approximate likelihood given by (3.4).

Remark 3.1. In the RKSB method, the space of densities induced by the RK4 numerical solution of the ODEs approaches the space of actual densities as the sample size n goes to infinity. This justifies the use of the term "sieve" in "RKSB."

Remark 3.2. The assumptions (A1) and (A2) are necessary to prove the convergence of the Bayes estimator of $\boldsymbol{\gamma}$ to the true value $\boldsymbol{\gamma}_0$. These are usually satisfied in most practical situations for example the Lotka–Volterra equations considered in the simulation study. When the true regression function is the solution of the ODE, the Hessian matrix becomes $\int_0^1 \dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt$ which is positive definite unless the components of $\dot{f}_{\boldsymbol{\theta}_0}^T$ as is the case in our simulation study.

3.2. Runge–Kutta two-step Bayesian method (RKTB)

In the RKTB approach, the proposed model is embedded in nonparametric regression model

$$\mathbf{Y} = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.8}$$

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

$$\begin{gathered} \max_{1\le i\le k_n-1} |\xi_{i+1} - 2\xi_i + \xi_{i-1}| = o\big(k_n^{-1}\big), \\ \max_{1\le i\le k_n-1} |\xi_i - \xi_{i-1}| / \min_{1\le i\le k_n-1} |\xi_i - \xi_{i-1}| \le M \end{gathered} \tag{3.9}$$

for some constant $M > 0$. Here ξ_0 and ξ_{k_n} are defined as 0 and 1, respectively. The criteria (3.9) is required to apply the asymptotic results obtained in [29] where they mention the similar criteria in equation (3) of that paper. We assume for a given σ^2

$$\boldsymbol{\beta} \sim N_{k_n+m-1}\bigl(\mathbf{0}, \sigma^2 n^2 k_n^{-1} I_{k_n+m-1}\bigr). \tag{3.10}$$

Simple calculation yields the conditional posterior distribution for $\boldsymbol{\beta}$ given σ^2 as

$$N_{k_n+m-1}\left(\left(\mathbf{X}_n^T\mathbf{X}_n + \frac{k_n}{n^2} I_{k_n+m-1}\right)^{-1}\mathbf{X}_n^T\mathbf{Y}, \sigma^2\left(\mathbf{X}_n^T\mathbf{X}_n + \frac{k_n}{n^2} I_{k_n+m-1}\right)^{-1}\right). \tag{3.11}$$

By model (3.8), the expected response at a point $t \in [0,1]$ is given by $\boldsymbol{\beta}^T\mathbf{N}(t)$, where $\mathbf{N}(\cdot) = (N_1(\cdot), \ldots, N_{k_n+m-1}(\cdot))^T$. Let us denote for a given parameter $\boldsymbol{\eta}$

$$R_{f,n}(\boldsymbol{\eta}) = \left\{\int_0^1 \bigl|f(t) - f_{\eta,r_n}(t)\bigr|^2 g(t)\,dt\right\}^{1/2},$$
$$R_{f_0}(\boldsymbol{\eta}) = \left\{\int_0^1 \bigl|f_0(t) - f_\eta(t)\bigr|^2 g(t)\,dt\right\}^{1/2},$$

where $f(t) = \boldsymbol{\beta}^T\mathbf{N}(t)$. Now we define $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\Theta} R_{f,n}(\boldsymbol{\eta})$ and induce posterior distribution on Θ through the posterior of $\boldsymbol{\beta}$ given by (3.11). Also let us define $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\eta}\in\Theta} R_{f_0}(\boldsymbol{\eta})$. Note that this definition of $\boldsymbol{\theta}_0$ takes into account the case when $f_{\boldsymbol{\theta}_0}$ is the true regression function with corresponding true parameter $\boldsymbol{\theta}_0$. We use the following standard assumptions.

(A5) For all $\epsilon > 0$,

$$\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} R_{f_0}(\boldsymbol{\eta}) > R_{f_0}(\boldsymbol{\theta}_0). \tag{3.12}$$

(A6) The matrix

$$\mathbf{J}_{\boldsymbol{\theta}_0} = -\int_0^1 \ddot{f}_{\boldsymbol{\theta}_0}(t)\bigl(f_0(t) - f_{\theta_0}(t)\bigr)g(t)\,dt + \int_0^1 \bigl(\dot{f}_{\boldsymbol{\theta}_0}(t)\bigr)^T\bigl(\dot{f}_{\boldsymbol{\theta}_0}(t)\bigr)g(t)\,dt$$

is nonsingular.

Remark 3.3. The assumption (A5) implies that $\boldsymbol{\theta}_0$ is a well-separated point of minima of $R_{f_0}(\cdot)$ which is needed to prove the convergence of the posterior distribution of $\boldsymbol{\theta}$ to the true value $\boldsymbol{\theta}_0$. A similar looking assumption appears in the argmax theorem used to show consistency of M-estimators and is a stronger version of the condition of uniqueness of the location of minimum.

Remark 3.4. The matrix $\mathbf{J}_{\boldsymbol{\theta}_0}$ is usually non-singular specially in the case when the true regression function satisfies the ODE since then the expression of $\mathbf{J}_{\boldsymbol{\theta}_0}$ becomes $\int_0^1 \dot{f}^T_{\boldsymbol{\theta}_0}(t)\,\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt$ which is usually positive definite.

4. Main results

Our main results are given by Theorems 4.1 and 4.2.

Theorem 4.1. *Let the posterior probability measure related to RKSB be denoted by* Π_n. *Then posterior of* $\boldsymbol{\gamma}$ *contracts at* $\boldsymbol{\gamma}_0$ *at the rate* $n^{-1/2}$ *and*

$$\big\|\Pi_n\big(\sqrt{n}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\in\cdot|\mathbf{X},\mathbf{Y}\big)-\mathbf{N}\big(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0},\sigma_*^2\mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}\big)\big\|_{\mathrm{TV}}=o_{P_0}(1),$$

where $\mathbf{V}_{\boldsymbol{\gamma}_0}=\left(\begin{smallmatrix}\sigma_*^{-2}\mathbf{V}_{\theta_0} & \mathbf{0}\\ \mathbf{0} & \sigma_*^{-4}/2\end{smallmatrix}\right)$ *with*

$$\mathbf{V}_{\boldsymbol{\theta}_0}=\int_0^1\left(\dot f_{\boldsymbol{\theta}_0}^T(t)\dot f_{\boldsymbol{\theta}_0}(t)-\frac{\partial}{\partial\boldsymbol{\theta}}\big(\dot f_{\boldsymbol{\theta}}^T(t)\big(f_0(t)-f_{\boldsymbol{\theta}_0}(t)\big)\big)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)g(t)\,dt$$

and $\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}=\mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}\mathbb{G}_n\dot\ell_{\boldsymbol{\gamma}_0,n}$.

Since $\boldsymbol{\theta}$ is a sub-vector of $\boldsymbol{\gamma}$, we get Bernstein–von Mises theorem for the posterior distribution of $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$, the mean and dispersion matrix of the limiting Gaussian distribution being the corresponding sub-vector and sub-matrix of $\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}$ and $\sigma_*^2\mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}$ respectively. We also get the following important corollary.

Corollary 1. *When the regression model* (3.1) *is correctly specified and also the error is Gaussian, the Bayes estimator based on* Π_n *is asymptotically efficient.*

Let us denote $\mathbf{C}(t)=\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}(\dot f_{\boldsymbol{\theta}_0}(t))^T$ and $\mathbf{H}_n^T=\int_0^1\mathbf{C}(t)\mathbf{N}^T(t)g(t)\,dt$. Note that $\mathbf{C}(t)$ is a p-component vector. Also, we denote the posterior probability measure of RKTB by Π_n^*. Now we have the following result.

Theorem 4.2. *Let*

$$\boldsymbol{\mu}_n=\sqrt{n}\mathbf{H}_n^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\mathbf{X}_n^T\mathbf{Y}-\sqrt{n}\int_0^1\mathbf{C}(t)f_0(t)g(t)\,dt,$$

$$\boldsymbol{\Sigma}_n=n\mathbf{H}_n^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\mathbf{H}_n$$

and $\mathbf{B}=((\langle C_k(\cdot),C_{k'}(\cdot)\rangle_g))_{k,k'=1,\ldots,p}$. *If* $\mathbf{B}$ *is non-singular, then for* $m\geq 2$ *and* $n^{1/(2m)}\ll k_n\ll n^{1/2}$,

$$\big\|\Pi_n^*\big(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|Y\big)-N\big(\boldsymbol{\mu}_n,\sigma_0^2\boldsymbol{\Sigma}_n\big)\big\|_{\mathrm{TV}}=o_{P_0}(1).\tag{4.1}$$

Remark 4.1. It will be proved later in Lemma 10 that both $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are stochastically bounded. Hence, with high true probability the posterior distribution of $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ contracts at $\mathbf{0}$ at $n^{-1/2}$ rate.

Remark 4.2. The previous theorem indicates that second order smoothness of the true mean function is sufficient to ensure the contraction rate $n^{-1/2}$. The issue of Bayesian adaptation does not arise in this case. For $m = 2$, the required condition becomes $n^{1/4} \ll k_n \ll n^{1/2}$. Since the knots are chosen deterministically, we do not need to assign a prior on the number of terms of the random series used.

We also get the following important corollary.

Corollary 2. *When the regression model* (3.1) *is correctly specified and the true distribution of error is Gaussian, the Bayes estimator based on* Π_n^* *is asymptotically efficient.*

Remark 4.3. The Bayesian two-step approach [2] considers the distance between the derivative of the function in the nonparametric model and the derivative given by the ODE. On the other hand, RKTB approach deals directly with the distance between the function in the nonparametric model and the parametric nonlinear regression model through the RK4 approximate solution of the ODE. Direct distance in the latter approach produces the efficient linearization giving rise to efficient concentration of the posterior distribution which can be traced back to efficiency properties of minimum distance estimation methods depending on the nature of the distance.

Remark 4.4. RKSB is the Bayesian analog of estimating $\boldsymbol{\theta}$ as

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

Similarly, RKTB is the Bayesian analog of $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\eta}\in\Theta} \int_0^1 (\hat{f}(t) - f_{\boldsymbol{\eta},r_n})^2 g(t)\,dt$, where $\hat{f}(\cdot)$ stands for the nonparametric estimate of f based on B-splines. Arguments similar to ours should be able to establish analogous convergence results for these estimators.

5. Simulation study

We consider the Lotka–Volterra equations to study the posterior distribution of $\boldsymbol{\theta}$. We consider two cases. In case 1, the true regression function belongs to the solution set and in case 2 it does not. Thus, we have $p = 4$, $d = 2$ and the ODE's are given by

$$\begin{aligned} F_1\big(t, \mathbf{f}_{\boldsymbol{\theta}}(t), \boldsymbol{\theta}\big) &= \theta_1 f_{1\boldsymbol{\theta}}(t) - \theta_2 f_{1\boldsymbol{\theta}}(t) f_{2\boldsymbol{\theta}}(t), \\ F_2\big(t, \mathbf{f}_{\boldsymbol{\theta}}(t), \boldsymbol{\theta}\big) &= -\theta_3 f_{2\boldsymbol{\theta}}(t) + \theta_4 f_{1\boldsymbol{\theta}}(t) f_{2\boldsymbol{\theta}}(t), \qquad t \in [0,1], \end{aligned}$$

with initial condition $f_{1\boldsymbol{\theta}}(0) = 1$, $f_{2\boldsymbol{\theta}}(0) = 0.5$. The above system is not analytically solvable.

Case 1 (well-specified case): The true regression function is $\mathbf{f}_0(t) = (f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t))^T$ where $\boldsymbol{\theta}_0 = (10, 10, 10, 10)^T$.

Table 1. Coverages and average lengths of the Bayesian credible intervals for the three methods in case of well-specified regression model

		RKSB		RKTB		TS	
n		Coverage (se)	Length (se)	Coverage (se)	Length (se)	Coverage (se)	Length (se)
100	θ_1	100.0	2.25	100.0	2.17	100.0	6.93
		(0.00)	(0.29)	(0.00)	(0.65)	(0.00)	(4.95)
	θ_2	100.0	2.57	100.0	2.48	100.0	6.67
		(0.00)	(0.33)	(0.00)	(0.74)	(0.00)	(4.90)
	θ_3	99.9	2.50	100.0	2.44	100.0	7.12
		(0.00)	(0.34)	(0.00)	(1.44)	(0.00)	(4.92)
	θ_4	100.0	2.27	100.0	2.20	100.0	6.59
		(0.00)	(0.32)	(0.00)	(1.19)	(0.00)	(4.77)
500	θ_1	100.0	0.75	99.4	0.56	99.2	1.09
		(0.00)	(0.06)	(0.00)	(0.02)	(0.00)	(0.05)
	θ_2	100.0	0.85	99.4	0.64	98.8	1.02
		(0.00)	(0.07)	(0.00)	(0.02)	(0.00)	(0.05)
	θ_3	100.0	0.82	99.3	0.61	99.0	1.16
		(0.00)	(0.07)	(0.00)	(0.02)	(0.00)	(0.05)
	θ_4	99.9	0.74	99.3	0.56	99.0	1.04
		(0.00)	(0.06)	(0.00)	(0.02)	(0.00)	(0.05)

Case 2 (misspecified case): The true regression function is $\mathbf{f}_0(t) = (f_{1\boldsymbol{\tau}_0}(t) + \frac{t^2+t-c_1}{6}, f_{2\boldsymbol{\tau}_0}(t) + \frac{t^2+t-c_2}{6})^T$ where $\boldsymbol{\tau}_0 = (10, 10, 10, 10)^T$ and c_1 and c_2 are chosen so that

$$\int_0^1 f_{1\boldsymbol{\tau}_0}(t)\big(t^2+t-c_1\big) = \int_0^1 f_{2\boldsymbol{\tau}_0}(t)\big(t^2+t-c_2\big) = 0.$$

For a sample of size n, the X_i's are drawn from Uniform(0, 1) distribution for $i = 1, \ldots, n$. Samples of sizes 100 and 500 are considered. We simulate 900 replications for each case. The output are displayed in Tables 1 and 2, respectively. Under each replication, a sample of size 1000 is drawn from the posterior distribution of $\boldsymbol{\theta}$ using RKSB, RKTB and Bayesian two-step [2] methods and then 95% equal tailed credible intervals are obtained. For case 2, we do not consider the Bayesian two-step method since there is no existing result on asymptotic efficiency under misspecification of the regression function and hence it is not comparable with the numerical solution based methods. The Bayesian two-step method is abbreviated as TS in Table 1. We calculate the coverage and the average length of the corresponding credible intervals over these 900 replications. The estimated standard errors of the interval length and coverage are given inside the parentheses in the tables.

The true distribution of error is taken $N(0, (0.1)^2)$. We put an inverse gamma prior on σ^2 with shape and scale parameters being 30 and 5, respectively. For RKSB, the prior for each θ_j is chosen as independent Gaussian distribution with mean 6 and variance 16 for $j = 1, \ldots, 4$. We take n grid points to obtain the numerical solution of the ODE by RK4 for a sample of size n.

Table 2. Coverages and average lengths of the Bayesian credible intervals for RKSB and RKTB in case of misspecified regression model

		RKSB		RKTB	
n		Coverage (se)	Length (se)	Coverage (se)	Length (se)
100	θ_1	99.4	2.32	100.0	2.21
		(0.01)	(0.31)	(0.00)	(0.83)
	θ_2	99.1	2.64	100.0	2.46
		(0.01)	(0.36)	(0.00)	(0.79)
	θ_3	99.4	2.78	100.0	2.7
		(0.01)	(0.46)	(0.00)	(1.73)
	θ_4	99.3	2.5	100.0	2.38
		(0.01)	(0.43)	(0.00)	(1.43)
500	θ_1	98.8	0.89	99.3	0.56
		(0.00)	(0.07)	(0.00)	(0.02)
	θ_2	98.9	1	99.5	0.62
		(0.00)	(0.13)	(0.00)	(0.02)
	θ_3	99.0	1.05	99.2	0.66
		(0.00)	(0.09)	(0.00)	(0.03)
	θ_4	99.2	0.94	99.1	0.59
		(0.00)	(0.08)	(0.00)	(0.02)

According to the requirements of Theorem 4.2 of this paper and Theorem 1 of [2], we take $m = 3$ and $m = 5$ for RKTB and Bayesian two-step method, respectively. In both cases, we choose $k_n - 1$ equispaced interior knots $\frac{1}{k_n}, \frac{2}{k_n}, \ldots, \frac{k_n-1}{k_n}$. This specific choice of knots satisfies the pseudo-uniformity criteria (3.9) with $M = 1$. Looking at the order of k_n suggested by Theorem 4.2, k_n is chosen in the order of $n^{1/5}$ giving the values of k_n as 13 and 18 for $n = 100$ and $n = 500$ respectively in RKTB. In Bayesian two-step method, the values of k_n are 17 and 20 for $n = 100$ and $n = 500$, respectively by choosing k_n in the order of $n^{1/9}$ following the suggestion given in Theorem 1 of [2]. In all the cases, the constant multiplier to the chosen asymptotic order is selected through cross-validation.

We separately analyze the output given in Table 1 since it deals with asymptotic efficiency. Not surprisingly the first two methods perform much better compared to the third one because of asymptotic efficiency obtained from Corollaries 1 and 2, respectively. For RKSB, a single replication took about one hour and four hours for samples of sizes 100 and 500, respectively. For RKTB, these times are around one hour and two and half hours, respectively. In Bayesian two-step method, each replication took about one and two minutes for $n = 100$ and 500, respectively. Thus from the computational point of view Bayesian two-step method is preferable than the numerical solution based approaches. We also show the trace plots of RKSB in Figures 1 and 2. We used 50 000 MCMC iterations with 2000 burn in and thinning of lag 48. It seems that the mixing is reasonable and it should be possible to improve by running a longer chain at the expense of more computing time.

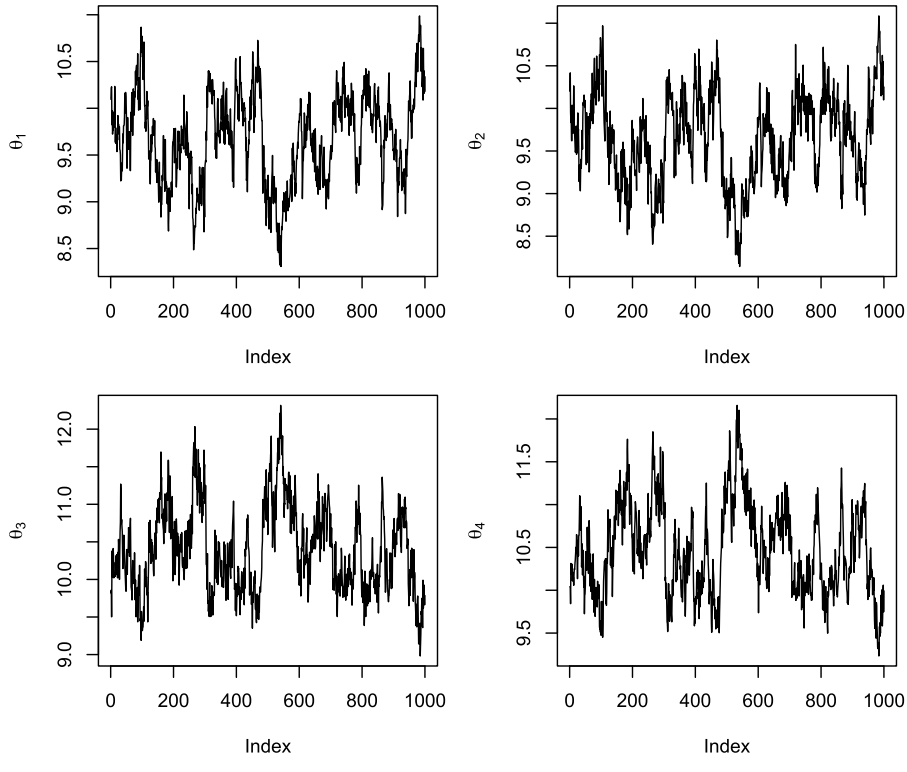

Figure 1. Trace plots of θ_1, θ_2, θ_3 and θ_4 in RKSB for $n = 100$.

We also consider another study for the well-specified case for true parameter $\boldsymbol{\theta}_0 = (5, 5, 5, 5)^T$. The output are given in Table 3 which shows that the coverages are robust with respect to the choice of the true parameter vector.

6. Proofs

We use the operators $\mathrm{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ to denote expectation and variance with respect to P_0.

Proof of Theorem 4.1. From Lemma 1 below, we know that there exists a compact subset U of $(0, \infty)$ such that $\Pi_n(\sigma^2 \in U|\mathbf{X}, \mathbf{Y}) \xrightarrow{P_0} 1$. Let $\Pi_{U,n}(\cdot|\mathbf{X}, \mathbf{Y})$ be the posterior distribution conditioned on $\sigma^2 \in U$. By Theorem 2.1 of [17] if we can ensure that there exist stochastically bounded random variables $\boldsymbol{\Delta}_{n,\gamma_0}$ and a positive definite matrix $\mathbf{V}_{\boldsymbol{\gamma}_0}$ such that for every compact set $K \subset \mathbb{R}^{p+1}$,

$$\sup_{h \in K} \left| \log \frac{p^{(n)}_{\boldsymbol{\gamma}_0 + \mathbf{h}/\sqrt{n}, n}}{p^{(n)}_{\boldsymbol{\gamma}_0, n}}(\mathbf{X}, Y) - \mathbf{h}^T \mathbf{V}_{\boldsymbol{\gamma}_0} \boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0} + \frac{1}{2} \mathbf{h}^T \mathbf{V}_{\boldsymbol{\gamma}_0} \mathbf{h} \right| \to 0, \tag{6.1}$$

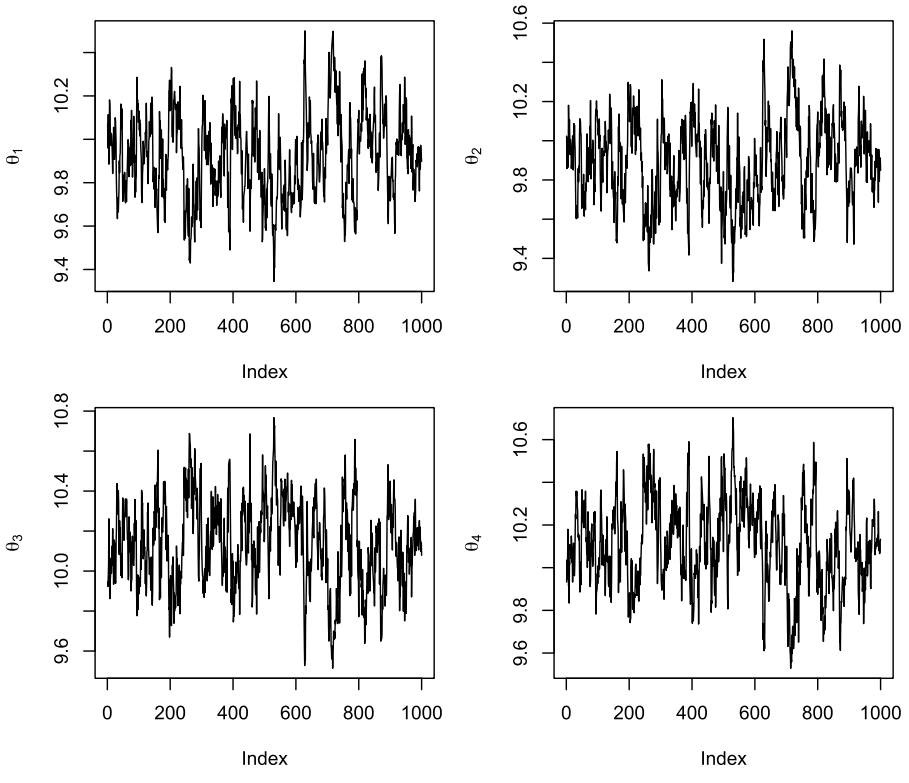

Figure 2. Trace plots of θ_1, θ_2, θ_3 and θ_4 in RKSB for $n = 500$.

in (outer) $P_0^{(n)}$ probability and that for every sequence of constants $M_n \to \infty$, we have

$$P_0^{(n)} \Pi_{U,n}\big(\sqrt{n}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| > M_n | \mathbf{X}, Y\big) \to 0, \tag{6.2}$$

then

$$\big\| \Pi_{U,n}\big(\sqrt{n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \in \cdot | \mathbf{X}, Y\big) - \mathbf{N}\big(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}, \mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}\big) \big\|_{\mathrm{TV}} = o_{P_0}(1).$$

We show that the conditions (6.1) and (6.2) hold in Lemmas 1 to 5. Lemma 2 gives that $\mathbf{V}_{\boldsymbol{\gamma}_0} = \begin{pmatrix} \sigma_*^{-2}\mathbf{V}_{\theta_0} & \mathbf{0} \\ \mathbf{0} & \sigma_*^{-4}/2 \end{pmatrix}$ with

$$\mathbf{V}_{\boldsymbol{\theta}_0} = \int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}^T(t) \dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial \boldsymbol{\theta}} \big(\dot{f}_{\boldsymbol{\theta}}^T(t) \big(f_0(t) - f_{\boldsymbol{\theta}_0}(t) \big) \big) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} \right) g(t)\, dt$$

and $\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0} = \mathbf{V}_{\boldsymbol{\gamma}_0}^{-1} \mathbb{G}_n \dot{\ell}_{\boldsymbol{\gamma}_0,n}$. Since $\|\Pi_n - \Pi_{U,n}\|_{\mathrm{TV}} = o_{P_0}(1)$, we get

$$\big\| \Pi_n\big(\sqrt{n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \in \cdot | \mathbf{X}, Y\big) - \mathbf{N}\big(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}, \mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}\big) \big\|_{\mathrm{TV}} = o_{P_0}(1).$$

Hence, we get the desired result. □

Table 3. Coverages and average lengths of the Bayesian credible intervals for the three methods in case of well-specified regression model with $\boldsymbol{\theta}_0 = (5, 5, 5, 5)^T$

		RKSB		RKTB		TS	
n		Coverage (se)	Length (se)	Coverage (se)	Length (se)	Coverage (se)	Length (se)
100	θ_1	100.0 (0.00)	1.16 (0.1)	100.0 (0.00)	1.01 (0.14)	100.0 (0.00)	6.24 (5.44)
	θ_2	100.0 (0.00)	1.35 (0.12)	100.0 (0.00)	1.18 (0.16)	100.0 (0.00)	5.25 (4.64)
	θ_3	100.0 (0.00)	1.96 (0.22)	100.0 (0.00)	1.82 (0.66)	100.0 (0.00)	4.98 (4.35)
	θ_4	100.0 (0.00)	1.58 (0.17)	100.0 (0.00)	1.45 (0.45)	100.0 (0.00)	4.55 (4.07)
500	θ_1	100.0 (0.00)	0.36 (0.02)	99.0 (0.00)	0.26 (0.01)	99.8 (0.00)	0.77 (0.04)
	θ_2	100.0 (0.00)	0.42 (0.02)	99.2 (0.00)	0.31 (0.01)	99.6 (0.00)	0.64 (0.03)
	θ_3	100.0 (0.00)	0.6 (0.04)	99.2 (0.00)	0.44 (0.02)	99.4 (0.00)	0.64 (0.03)
	θ_4	100.0 (0.00)	0.48 (0.03)	99.0 (0.00)	0.36 (0.02)	99.4 (0.00)	0.57 (0.02)

Proof of Corollary 1. The log-likelihood of the correctly specified model with Gaussian error is given by

$$\ell_{\boldsymbol{\gamma}_0}(X, Y) = -\log \sigma_0 - \frac{1}{2\sigma_0^2}\big|Y - f_{\boldsymbol{\theta}_0}(X)\big|^2 + \log g(X).$$

Thus, $\frac{\partial}{\partial \boldsymbol{\theta}_0}\ell_{\boldsymbol{\gamma}_0}(X, Y) = \sigma_0^{-2}(\dot{f}_{\boldsymbol{\theta}_0}(X))^T(Y - f_{\boldsymbol{\theta}_0}(X))$ and $\frac{\partial}{\partial \sigma_0^2}\ell_{\boldsymbol{\gamma}_0}(X, Y) = -\frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}|Y - f_{\boldsymbol{\theta}_0}(X)|^2$. Hence, the Fisher information is given by

$$\mathbf{I}(\boldsymbol{\gamma}_0) = \begin{pmatrix} \sigma_0^{-2}\displaystyle\int_0^1 \dot{f}^T_{\boldsymbol{\theta}_0}(t)\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt & \mathbf{0} \\ \mathbf{0} & \sigma_0^{-4}/2 \end{pmatrix}.$$

Looking at the form of $\mathbf{V}_{\boldsymbol{\gamma}_0}$ in Theorem 4.1, we get $\mathbf{V}^{-1}_{\boldsymbol{\gamma}_0} = (\mathbf{I}(\boldsymbol{\gamma}_0))^{-1}$ if the regression function is correctly specified and the true error distribution is $N(0, \sigma_0^2)$. □

Proof of Theorem 4.2. We have for $f(\cdot) = \boldsymbol{\beta}^T\mathbf{N}(\cdot)$

$$\mathbf{J}^{-1}_{\boldsymbol{\theta}_0}\boldsymbol{\Gamma}(f) = \int_0^1 \mathbf{C}(t)\boldsymbol{\beta}^T\mathbf{N}(t)g(t)\,dt = \mathbf{H}_n^T\boldsymbol{\beta}, \tag{6.3}$$

where $\mathbf{H}_n^T = \int_0^1 \mathbf{C}(t)\mathbf{N}^T(t)g(t)\,dt$ which is a matrix of order $p \times (k_n + m - 1)$. Consequently, the asymptotic variance of the conditional posterior distribution of $\mathbf{H}_n^T\boldsymbol{\beta}$ is $\sigma^2\mathbf{H}_n^T(\mathbf{X}_n^T\mathbf{X}_n + \frac{k_n}{n^2}\mathbf{I})^{-1}\mathbf{H}_n$. By Lemma 9 and the posterior consistency of the σ^2 given by Lemma 11, it suffices to show that for any neighborhood $\mathcal{N}$ of σ_0^2,

$$\sup_{\sigma^2\in\mathcal{N}} \left\| \Pi_n^*\left(\sqrt{n}\mathbf{H}_n^T\boldsymbol{\beta} - \sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}(f_0) \in \cdot|\mathbf{X},\mathbf{Y},\sigma^2\right) - N\left(\boldsymbol{\mu}_n,\sigma^2\boldsymbol{\Sigma}_n\right)\right\|_{\mathrm{TV}} = o_{P_0}(1). \quad (6.4)$$

Note that $\Pi(\mathcal{N}^c|\mathbf{X},\mathbf{Y}) = o_{P_0}(1)$. It is straightforward to verify that the Kullback–Leibler divergence between $N((\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{Y}, \sigma^2(\mathbf{X}_n^T\mathbf{X}_n)^{-1})$ and the distribution given by (3.11) converges in P_0-probability to zero uniformly over $\sigma^2 \in \mathcal{N}$ and hence, so is the total variation distance. By linear transformation, (6.4) follows. Note that

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

Using the fact that $\boldsymbol{\Sigma}_n$ is stochastically bounded given by Lemma 10, the total variation distance between the two normal distributions appearing in the second term of the above display is bounded by a constant multiple of $|\sigma^2 - \sigma_0^2|$, and hence can be made arbitrarily small by choosing $\mathcal{N}$ accordingly. The first term converges in probability to zero by (6.4). The third term converges in probability to zero by the posterior consistency. □

Proof of Corollary 2. The log-likelihood of the correctly specified model is given by

$$\ell_{\boldsymbol{\theta}_0}(X,Y) = -\log\sigma_0 - \frac{1}{2\sigma_0^2}\left|Y - f_{\boldsymbol{\theta}_0}(X)\right|^2 + \log g(X).$$

Thus $\dot{\ell}_{\boldsymbol{\theta}_0}(X,Y) = -\sigma_0^{-2}(\dot{f}_{\boldsymbol{\theta}_0}(X))^T(Y - f_{\boldsymbol{\theta}_0}(X))$ and the Fisher information is given by $\mathbf{I}(\boldsymbol{\theta}_0) = \sigma_0^{-2}\int_0^1(\dot{f}_{\boldsymbol{\theta}_0}(X))^T\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt$. In the proof of Lemma 10 we obtained that $\sigma_0^2\boldsymbol{\Sigma}_n \xrightarrow{P_0} \sigma_0^2\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\int_0^1(\dot{f}_{\boldsymbol{\theta}_0}(t))^T\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}$. This limit is equal to $(\mathbf{I}(\boldsymbol{\theta}_0))^{-1}$ under the correct specification of the regression function as well as the likelihood. □

7. Proofs of technical lemmas

The first five lemmas in this section are related to RKSB. The rest are for RKTB. The first lemma shows that the posterior of σ^2 lies inside a compact set with high probability.

Lemma 1. *There exists a compact set U independent of $\boldsymbol{\theta}$ and n such that $\Pi_n(\sigma^2 \in U|\mathbf{X}, \mathbf{Y}) \xrightarrow{P_0} 1$.*

Proof. Given $\boldsymbol{\theta}$, the conditional posterior of σ^2 is an inverse gamma distribution with shape and scale parameters $n/2+a$ and $2^{-1}\sum_{i=1}^n (Y_i - f_{\boldsymbol{\theta}}(X_i))^2 + b$, respectively. Clearly $\mathrm{E}(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n (Y_i - f_{\boldsymbol{\theta}}(X_i))^2 + o(1)$ a.s. Hence, it is easy to show using the weak law of large numbers that the mean of the conditional posterior of σ^2 converges in P_0-probability to $\sigma_{\boldsymbol{\theta}}^2 := \sigma_0^2 + \int_0^1 (f_0(t) - f_{\boldsymbol{\theta}}(t))^2 g(t)\,dt$. Then it follows that for any $\epsilon > 0$, $\Pi_n(\sigma^2 \in [\sigma_{\boldsymbol{\theta}}^2 - \epsilon, \sigma_{\boldsymbol{\theta}}^2 + \epsilon]|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})$ converges in P_0-probability to 1. Since Θ is compact and $\sigma_{\boldsymbol{\theta}}^2$ is continuous in $\boldsymbol{\theta}$, there exists a compact set U such that $U \supseteq [\sigma_{\boldsymbol{\theta}}^2 - \epsilon, \sigma_{\boldsymbol{\theta}}^2 + \epsilon]$ for all $\boldsymbol{\theta}$. Now $\Pi_n(\sigma^2 \notin U|\mathbf{X}, \mathbf{Y})$ is bounded above by

$$\begin{aligned}&\int_\Theta \Pi_n\big(\big|\sigma^2 - \sigma_{\boldsymbol{\theta}}^2\big| > \epsilon|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big)\,d\Pi_n(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})\\ &\quad\le \epsilon^{-2}\int_\Theta \big(\big(\mathrm{E}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big) - \sigma_{\boldsymbol{\theta}}^2\big)^2 + \mathrm{Var}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big)\big)\,d\Pi_n(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}).\end{aligned}$$

It suffices to prove that

$$\sup_{\boldsymbol{\theta}\in\Theta}\big|\mathrm{E}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big) - \sigma_{\boldsymbol{\theta}}^2\big| = o_{P_0}(1) \quad \text{and} \quad \sup_{\boldsymbol{\theta}\in\Theta}\mathrm{Var}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big) = o_{P_0}(1).$$

Using the facts that $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(x)$ is Lipschitz continuous and other smoothness criteria of $f_{\boldsymbol{\theta}}(x)$ and $f_0(x)$ and applying Theorem 19.4 and Example 19.7 of [25], it follows that $\{(Y - f_{\boldsymbol{\theta}}(X))^2 : \boldsymbol{\theta} \in \Theta\}$ is P_0-Glivenko–Cantelli and hence

$$\sup_{\boldsymbol{\theta}\in\Theta}\big|\mathrm{E}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big) - \mathrm{E}_0\big(\mathrm{E}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big)\big)\big| = o_{P_0}(1).$$

Also, it can be easily shown that the quantity $\sup_{\boldsymbol{\theta}\in\Theta}|\mathrm{E}_0(\mathrm{E}(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})) - \sigma_{\boldsymbol{\theta}}^2| \to 0$ as $n \to \infty$ since

$$\begin{aligned}\mathrm{E}_0\big(\mathrm{E}\big(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}\big)\big) &= \sigma_0^2 + \int_0^1 \big(f_0(t) - f_{\boldsymbol{\theta}}(t)\big)^2 g(t)\,dt\\ &\quad - \frac{2(a-1)(\sigma_0^2 + \int_0^1 (f_0(t) - f_{\boldsymbol{\theta}}(t))^2 g(t)\,dt)}{n + 2a - 2} + \frac{2b}{n + 2a - 2}\end{aligned}$$

and the parameter space Θ is compact and the mapping $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\cdot)$ is continuous. This gives the first assertion. To see the second assertion, observe that $\mathrm{Var}(\sigma^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) = O(n^{-1})$ a.s. by the previous assertion and the fact that the conditional posterior of σ^2 given $\boldsymbol{\theta}$ is inverse gamma. □

In view of the previous lemma, we choose the parameter space for $\boldsymbol{\gamma}$ to be $\Theta \times U$ from now onwards. We show that the condition (6.1) holds by the following lemma.

Lemma 2. *For the model induced by Runge–Kutta method as described in Section 3, we have*

$$\sup_{\mathbf{h}\in K}\left|\log\frac{\prod_{i=1}^n p_{\boldsymbol{\gamma}_0+\mathbf{h}/\sqrt{n},n}(X_i,Y_i)}{\prod_{i=1}^n p_{\boldsymbol{\gamma}_0,n}(X_i,Y_i)} - \mathbf{h}^T\mathbf{V}_{\boldsymbol{\gamma}_0}\boldsymbol{\Delta}_{n,\gamma_0} + \frac{1}{2}\mathbf{h}^T\mathbf{V}_{\gamma_0}\mathbf{h}\right| \to 0,$$

in (outer) $P_0^{(n)}$*-probability for every compact set* $\mathbf{K}\subset\mathbb{R}^{p+1}$*, where*

$$\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0} = \mathbf{V}_{\boldsymbol{\gamma}_0}^{-1}\mathbb{G}_n\dot{\ell}_{\boldsymbol{\gamma}_0,n}$$

and $\mathbf{V}_{\boldsymbol{\gamma}_0} = \begin{pmatrix}\sigma_*^{-2}\mathbf{V}_{\theta_0} & \mathbf{0}\\ \mathbf{0} & \sigma_*^{-4}/2\end{pmatrix}$ *with*

$$\mathbf{V}_{\boldsymbol{\theta}_0} = \int_0^1\left(\dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial\boldsymbol{\theta}}\big(\dot{f}_{\boldsymbol{\theta}}^T(t)\big(f_0(t)-f_{\boldsymbol{\theta}_0}(t)\big)\big)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)g(t)\,dt.$$

Proof. Let G be an open neighborhood containing $\boldsymbol{\gamma}_0$. For $\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in G$, we have

$$\big|\log\big(p_{\boldsymbol{\gamma}_1}(X_1,Y_1)/p_{\boldsymbol{\gamma}_2}(X_1,Y_1)\big)\big| \le m(X_1,Y_1)\|\gamma_1-\gamma_2\|,$$

where $m(X_1,Y_1)$ is

$$\sup\left\{\frac{|Y_1-f_{\boldsymbol{\theta}}(X_1)|}{\sigma^2}\big\|\dot{f}_{\boldsymbol{\theta}}(X_1)\big\| + \frac{(Y_1-f_{\boldsymbol{\theta}}(X_1))^2}{2\sigma^4} + \frac{1}{2\sigma^2} : \big(\boldsymbol{\theta},\sigma^2\big)\in G\right\},$$

which is square integrable. Therefore, by Lemma 19.31 of [25], for any sequence $\{\mathbf{h}_n\}$ bounded in P_0-probability,

$$\mathbb{G}_n\big(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\mathbf{h}_n/\sqrt{n})}-\ell_{\boldsymbol{\gamma}_0}) - \mathbf{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}\big) = o_{P_0}(1).$$

Using the laws of large numbers and (2.1), we find that

$$\mathbb{G}_n\big(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\mathbf{h}_n/\sqrt{n})}-\ell_{\boldsymbol{\gamma}_0}) - \mathbf{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}\big) - \mathbb{G}_n\big(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\mathbf{h}_n/\sqrt{n}),n}-\ell_{\boldsymbol{\gamma}_0,n}) - \mathbf{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}\big)$$

is $O_{P_0}(\sqrt{n}r_n^{-4})$ which is $o_{P_0}(1)$ by the condition (3.3) on r_n. Hence,

$$\mathbb{G}_n\big(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\mathbf{h}_n/\sqrt{n}),n}-\ell_{\boldsymbol{\gamma}_0,n}) - \mathbf{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}\big) = o_{P_0}(1).$$

We note that

$$\begin{aligned}&-P_0\log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\\ &\quad= \log\sigma - \log\sigma_* + \frac{1}{2\sigma^2}\left[\sigma_0^2 + \int_0^1\big|f_0(t)-f_{\boldsymbol{\theta},r_n}(t)\big|^2 g(t)\,dt\right]\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2\sigma_*^2}\left[\sigma_0^2 + \int_0^1 \left|f_0(t) - f_{\boldsymbol{\theta}_0,r_n}(t)\right|^2 g(t)\,dt\right] \\
&= \log\sigma - \log\sigma_* + \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma_*^2}\right)\left[\sigma_0^2 + \int_0^1 \left|f_0(t) - f_{\boldsymbol{\theta},r_n}(t)\right|^2 g(t)\,dt\right] \\
&\quad + \frac{1}{2\sigma_*^2}\left[2\int_0^1 \big(f_0(t) - f_{\boldsymbol{\theta}_0,r_n}(t)\big)\big(f_{\boldsymbol{\theta}_0,r_n}(t) - f_{\boldsymbol{\theta},r_n}(t)\big)g(t)\,dt\right. \\
&\quad \left. + \int_0^1 \left|f_{\boldsymbol{\theta}_0,r_n}(t) - f_{\boldsymbol{\theta},r_n}(t)\right|^2 g(t)\,dt\right].
\end{aligned} \tag{7.1}
$$

Using (3.7), the last term inside the third bracket in (7.1) can be expanded as

$$
(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{V}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O\big(r_n^{-4}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\big) + o\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\big),
$$

where $\mathbf{V}_{\boldsymbol{\theta}_0} = \int_0^1 (\dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial\boldsymbol{\theta}}(\dot{f}_{\boldsymbol{\theta}}^T(t)(f_0(t) - f_{\boldsymbol{\theta}_0}(t)))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0})g(t)\,dt$. Also, writing $\sigma_*^2 = \sigma_0^2 + \int_0^1 |f_0(t) - f_{\boldsymbol{\theta}_0}(t)|^2 g(t)\,dt$ and using (3.7), the first term in (7.1) is given by

$$
\begin{aligned}
&-\frac{1}{2}\log\left(\frac{\sigma_*^2}{\sigma^2} - 1 + 1\right) + \frac{1}{2}\left(\frac{\sigma_*^2}{\sigma^2} - 1\right) + O\big(r_n^{-4}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|\big) + o\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big) \\
&= \frac{(\sigma^2 - \sigma_*^2)^2}{4\sigma_*^4} + O\big(r_n^{-4}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|\big) + o\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big).
\end{aligned}
$$

Hence,

$$
P_0 \log \frac{p_{\boldsymbol{\gamma}_0 + \mathbf{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}} + \frac{1}{2n}\mathbf{h}_n^T \mathbf{V}_{\boldsymbol{\gamma}_0}\mathbf{h}_n = o\big(n^{-1}\big). \tag{7.2}
$$

We have already shown that

$$
n\mathbb{P}_n \log \frac{p_{\boldsymbol{\gamma}_0 + \mathbf{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}} - \mathbb{G}_n \mathbf{h}_n^T \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n} - nP_0 \log \frac{p_{\boldsymbol{\gamma}_0 + \mathbf{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}} = o_{P_0}(1). \tag{7.3}
$$

Substituting (7.2) in (7.3), we get the desired result. □

Now our objective is to prove (6.2). We define the measure $Q_{\boldsymbol{\gamma}}(A) = P_0(p_{\boldsymbol{\gamma}}/p_{\boldsymbol{\gamma}_0}\mathbb{1}_A)$ and the corresponding density $q_{\boldsymbol{\gamma}} = p_0 p_{\boldsymbol{\gamma}}/p_{\boldsymbol{\gamma}_0}$ as given in [17]. Also, we define a measure $Q_{\boldsymbol{\gamma},n}$ by $Q_{\boldsymbol{\gamma},n}(A) = P_0(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}\mathbb{1}_A)$ with $q_{\boldsymbol{\gamma},n} = p_0 p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}$. The misspecified Kullback–Leibler neighborhood of $\boldsymbol{\gamma}_0$ is defined as

$$
B(\epsilon, \boldsymbol{\gamma}_0, P_0) = \left\{\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : -P_0 \log\left(\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}}\right) \le \epsilon^2,\ P_0\left(\log\left(\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}}\right)\right)^2 \le \epsilon^2\right\}.
$$

By Theorem 3.1 of [17], condition (6.2) is satisfied if we can ensure that for every $\epsilon > 0$, there exists a sequence of tests $\{\phi_n\}$ such that

$$P_0^{(n)}\phi_n \to 0, \qquad \sup_{\{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\geq\epsilon\}} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\phi_n) \to 0. \tag{7.4}$$

The above condition is ensured by the next lemma.

Lemma 3. *Assume that $\boldsymbol{\gamma}_0$ is a unique point of minimum of $\boldsymbol{\gamma} \mapsto -P_0 \log p_{\boldsymbol{\gamma}}$. Then there exist tests ϕ_n satisfying* (7.4).

Proof. For given $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_0$ consider the tests $\phi_{n,\boldsymbol{\gamma}_1} = \mathbb{1}\{\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n}) < 0\}$. Note that $\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n}) = \mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1}) + O_{P_0}(r_n^{-4}) \overset{P_0^{(n)}}{\to} P_0 \log(p_0/q_{\boldsymbol{\gamma}_1})$ and $P_0 \log(p_0/q_{\boldsymbol{\gamma}_1}) = P_0 \log(p_{\boldsymbol{\gamma}_0}/p_{\boldsymbol{\gamma}_1}) > 0$ for $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_0$ by the definition of $\boldsymbol{\gamma}_0$. Hence, $P_0^{(n)}\phi_{n,\boldsymbol{\gamma}_1} \to 0$ as $n \to \infty$. By Markov's inequality we have that

$$\begin{aligned} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\phi_{n,\boldsymbol{\gamma}_1}) &= Q_{\boldsymbol{\gamma},n}^{(n)}\big(\exp\{sn\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n})\} > 1\big) \\ &\leq Q_{\boldsymbol{\gamma},n}^{(n)} \exp\{sn\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n})\} \\ &= \big(Q_{\boldsymbol{\gamma},n}(p_0/q_{\boldsymbol{\gamma}_1,n})^s\big)^n = \big(\rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma},s) + O\big(r_n^{-4}\big)\big)^n, \end{aligned}$$

for $\rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma},s) = \int p_0^s q_{\boldsymbol{\gamma}_1}^{-s} q_{\boldsymbol{\gamma}}\, d\mu$. By [16] the function $s \mapsto \rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_1,s)$ converges to $P_0(q_{\boldsymbol{\gamma}_1} > 0) = P_0(p_{\boldsymbol{\gamma}_1} > 0)$ as $s \uparrow 1$ and has left derivative $P_0 \log(\frac{q_{\boldsymbol{\gamma}_1}}{p_0})\mathbb{1}\{q_{\boldsymbol{\gamma}_1} > 0\} = P_0 \log(\frac{p_{\boldsymbol{\gamma}_1}}{p_{\boldsymbol{\gamma}_0}})\mathbb{1}\{p_{\boldsymbol{\gamma}_1} > 0\}$ at $s = 1$. Then either $P_0(p_{\boldsymbol{\gamma}_1} > 0) < 1$ or $P_0(p_{\boldsymbol{\gamma}_1} > 0) = 1$ and $P_0 \log(\frac{p_{\boldsymbol{\gamma}_1}}{p_{\boldsymbol{\gamma}_0}})\mathbb{1}\{p_{\boldsymbol{\gamma}_1} > 0\} = P_0 \log(\frac{p_{\boldsymbol{\gamma}_1}}{p_{\boldsymbol{\gamma}_0}}) < 0$ or both. In either case it follows that there exists $s_{\boldsymbol{\gamma}_1} < 1$ arbitrarily close to 1 such that $\rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_1,s_{\boldsymbol{\gamma}_1}) < 1$. It is easy to show that the map $\boldsymbol{\gamma} \mapsto \rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma},s_{\boldsymbol{\gamma}_1})$ is continuous at $\boldsymbol{\gamma}_1$ by the dominated convergence theorem. Therefore, for every $\boldsymbol{\gamma}_1$, there exists an open neighborhood $G_{\boldsymbol{\gamma}_1}$ such that

$$u_{\boldsymbol{\gamma}_1} = \sup_{\boldsymbol{\gamma}\in G_{\boldsymbol{\gamma}_1}} \rho(\boldsymbol{\gamma}_1,\boldsymbol{\gamma},s_{\boldsymbol{\gamma}_1}) < 1.$$

The set $\{\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \geq \epsilon\}$ is compact and hence can be covered with finitely many sets of the type $G_{\boldsymbol{\gamma}_i}$ for $i = 1,\dots,k$. Let us define $\phi_n = \max_i\{\phi_{n,\boldsymbol{\gamma}_i} : i = 1,\dots,k\}$. This test satisfies $P_0^{(n)}\phi_n \leq \sum_{i=1}^k P_0^{(n)}\phi_{n,\boldsymbol{\gamma}_i} \to 0$, and

$$Q_{\boldsymbol{\gamma},n}^{(n)}(1-\phi_n) \leq \max_{i=1,\dots,k} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\phi_{n,\boldsymbol{\gamma}_i}) \leq \max_{i=1,\dots,k}\big(u_{\boldsymbol{\gamma}_i} + O\big(r_n^{-4}\big)\big)^n \to 0$$

uniformly in $\boldsymbol{\gamma} \in \bigcup_{i=1}^k G_{\boldsymbol{\gamma}_i}$. Therefore, the tests ϕ_n meet (7.4). □

The proof of Theorem 3.1 of [17] also uses the results of the next two lemmas.

Lemma 4. *Suppose that $P_0\dot{\ell}_{\boldsymbol{\gamma}_0}\dot{\ell}_{\boldsymbol{\gamma}_0}^T$ is invertible. Then for every sequence $\{M_n\}$ such that $M_n \to \infty$, there exists a sequence of tests $\{\omega_n\}$ such that for some constant $D > 0$, $\epsilon > 0$ and large enough n,*

$$P_0^{(n)}\omega_n \to 0, \qquad Q_{\boldsymbol{\gamma},n}^{(n)}(1-\omega_n) \le e^{-nD(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 \wedge \epsilon^2)},$$

for all $\boldsymbol{\gamma} \in \Theta \times U$ such that $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \ge M_n/\sqrt{n}$.

Proof. Let $\{M_n\}$ be given. We construct two sequences of tests. The first sequence is used to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \boldsymbol{\gamma} \in (\Theta \times U)_1\}$ with $(\Theta \times U)_1 = \{\boldsymbol{\gamma} \in \Theta \times U : M_n/\sqrt{n} \le \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \epsilon\}$ and the second to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \boldsymbol{\gamma} \in (\Theta \times U)_2\}$ with $(\Theta \times U)_2 = \{\boldsymbol{\gamma} \in \Theta \times U : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| > \epsilon\}$. These two sequences are combined to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \ge M_n/\sqrt{n}\}$.

To construct the first sequence, a constant $L > 0$ is chosen to truncate the score-function, that is, $\dot{\ell}_{\gamma_0}^L = 0$ if $\|\dot{\ell}_{\boldsymbol{\gamma}_0}\| > L$ and $\dot{\ell}_{\boldsymbol{\gamma}_0}^L = \dot{\ell}_{\boldsymbol{\gamma}_0}$ otherwise. Similarly we define $\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L$. We define

$$\omega_{1,n} = \mathbb{1}\big\{\big\|(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L\big\| > \sqrt{M_n/n}\big\}.$$

Since the function $\dot{\ell}_{\boldsymbol{\gamma}_0}$ is square-integrable, we observe that the matrices $P_0\dot{\ell}_{\boldsymbol{\gamma}_0,n}\dot{\ell}_{\boldsymbol{\gamma}_0,n}^T$, $P_0\dot{\ell}_{\boldsymbol{\gamma}_0,n}(\dot{\ell^L}_{\boldsymbol{\gamma}_0,n})^T$ and $P_0\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L(\dot{\ell^L})_{\boldsymbol{\gamma}_0,n}^T$ can be made sufficiently close to each other for sufficiently large choices of L and n. We fix such an L. Now,

$$\begin{aligned}
P_0^{(n)}\omega_{1,n} &= P_0^{(n)}\big(\big\|\sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L\big\|^2 > M_n\big) \\
&\le P_0^{(n)}\big(\big\|\sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0}^L\big\|^2 > M_n/4\big) \\
&\quad + P_0^{(n)}\big(\big\|\sqrt{n}(\mathbb{P}_n - P_0)\big(\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L - \dot{\ell}_{\boldsymbol{\gamma}_0}^L\big)\big\|^2 > M_n/4\big).
\end{aligned}$$

The right-hand side of the above inequality converges to zero since both sequences inside the brackets are stochastically bounded. The rest of the proof follows from the proof of Theorem 3.3 of [17] and Lemma 2. As far as $Q_{\boldsymbol{\gamma},n}^{(n)}(1-\omega_{1,n})$ for $\boldsymbol{\gamma} \in (\Theta \times U)_1$ is concerned, for all $\boldsymbol{\gamma}$

$$\begin{aligned}
&Q_{\boldsymbol{\gamma},n}^{(n)}\big(\big\|(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L\big\| \le \sqrt{M_n/n}\big) \\
&\quad = Q_{\boldsymbol{\gamma},n}^{(n)}\Big(\sup_{\mathbf{v}\in S}\mathbf{v}^T(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L \le \sqrt{M_n/n}\Big) \\
&\quad \le \inf_{\mathbf{v}\in S} Q_{\boldsymbol{\gamma},n}^{(n)}\big(\mathbf{v}^T(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L \le \sqrt{M_n/n}\big),
\end{aligned}$$

where S is the unit sphere in $\mathbb{R}^{p+1}$. Choosing $\mathbf{v} = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)/\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|$, the right-hand side of the previous display can be bounded by

$$\begin{aligned}
&Q_{\boldsymbol{\gamma},n}^{(n)}\big((\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T(\mathbb{P}_n - P_0)\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L \le \sqrt{M_n/n}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|\big) \\
&\quad = Q_{\boldsymbol{\gamma},n}^{(n)}\big((\boldsymbol{\gamma}_0 - \boldsymbol{\gamma})^T(\mathbb{P}_n - \tilde{Q}_{\boldsymbol{\gamma},n})\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L \ge (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T(\tilde{Q}_{\boldsymbol{\gamma},n} - \tilde{Q}_{\boldsymbol{\gamma}_0,n})\dot{\ell}_{\boldsymbol{\gamma}_0,n}^L \\
&\qquad - \sqrt{M_n/n}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|\big),
\end{aligned}$$

where $\tilde{Q}_{\boldsymbol{\gamma},n} = \|Q_{\boldsymbol{\gamma},n}\|^{-1} Q_{\boldsymbol{\gamma},n}$ and also note that $P_0 = Q_{\boldsymbol{\gamma}_0,n} = \tilde{Q}_{\boldsymbol{\gamma}_0,n}$. It should be noted that

$$\begin{aligned}
&(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T (\tilde{Q}_{\boldsymbol{\gamma},n} - \tilde{Q}_{\boldsymbol{\gamma}_0,n}) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} \\
&\quad = \big(P_0(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\big)^{-1} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \big(P_0\big((p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n} - 1)\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big) \\
&\qquad + \big(1 - P_0(p_{\gamma,n}/p_{\boldsymbol{\gamma}_0,n})\big) P_0 \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big).
\end{aligned}$$

By Lemma 3.4 of [17],

$$\big(P_0(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n} - 1)\big) = O\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big)$$

as $\boldsymbol{\gamma} \to \boldsymbol{\gamma}_0$. Using the differentiability of $\boldsymbol{\gamma} \mapsto \log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})$ and Lemma 3.4 of [17], we see that

$$\begin{aligned}
&P_0\left\| \left(\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}} - 1 - (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n} \right) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} \right\| \\
&\quad \le P_0\left\| \left(\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}} - 1 - \log \frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}} \right) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} \right\| \\
&\qquad + P_0\left\| \left(\log \frac{p_{\gamma,n}}{p_{\gamma_0,n}} - (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n} \right) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} \right\|,
\end{aligned} \tag{7.5}$$

which is $o(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|)$. Also note that for all $\boldsymbol{\gamma} \in (\Theta \times U)_1$,

$$-\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \sqrt{M_n/n} \ge -\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2 (M_n)^{-1/2}.$$

Then we observe that for every $\delta > 0$, there exist $\epsilon > 0$, $L > 0$ and $N \ge 1$ such that for all $n \ge N$ and all $\boldsymbol{\gamma} \in (\Theta \times U)_1$,

$$\begin{aligned}
&(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T (\tilde{Q}_{\boldsymbol{\gamma},n} - \tilde{Q}_{\boldsymbol{\gamma}_0,n}) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} - \sqrt{M_n/n} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \\
&\quad \ge (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T P_0\big(\dot{\boldsymbol{\ell}}_{\gamma_0,n} \dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n}\big)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) - \delta \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2.
\end{aligned}$$

Denoting $\Delta(\boldsymbol{\gamma}) = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T P_0(\dot{\boldsymbol{\ell}}_{\gamma_0,n} \dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n})(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$ and using the positive definiteness of $P_0(\dot{\boldsymbol{\ell}}_{\gamma_0,n} \dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n})$ for sufficiently large n, there exists a positive constant c such that $-\delta\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2 \ge -\delta/c\Delta(\boldsymbol{\gamma})$. Also, there exists a constant $r(\delta)$ which depends only on $P_0(\dot{\boldsymbol{\ell}}_{\gamma_0,n} \dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n})$ and has the property that $r(\delta) \to 1$ if $\delta \to 0$. We can choose such an $r(\delta)$ to satisfy

$$Q^{(n)}_{\boldsymbol{\gamma},n}(1 - \omega_{1,n}) \le Q^{(n)}_{\boldsymbol{\gamma},n}\big((\boldsymbol{\gamma}_0 - \boldsymbol{\gamma})^T (\mathbb{P}_n - \tilde{Q}_{\boldsymbol{\gamma},n}) \dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n} \ge r(\delta)\Delta(\boldsymbol{\gamma})\big),$$

for sufficiently small ϵ, sufficiently large L and n, making the type-II error bounded above by the unnormalized tail probability $Q^{(n)}_{\boldsymbol{\gamma},n}(\bar{W}_n \ge r(\delta)\Delta(\boldsymbol{\gamma}))$ where $W_i = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T (\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}(X_i, Y_i) -$

$\tilde{Q}_{\boldsymbol{\gamma},n}\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n})$, $(1 \le i \le n)$. We note that $\tilde{Q}_{\boldsymbol{\gamma},n} W_i = 0$ and W_i are independent. Also,

$$|W_i| \le \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|\big(\big\|\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}(X_i, \mathbf{Y}_i)\big\| + \big\|\tilde{Q}_{\boldsymbol{\gamma},n}\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big\|\big) \le 2L\sqrt{p+1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|.$$

Then we have

$$\begin{aligned}
&\mathrm{Var}_{\tilde{Q}_{\boldsymbol{\gamma},n}} W_i \\
&= (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T\big[\tilde{Q}_{\boldsymbol{\gamma},n}\big(\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big(\dot{\ell}^L_{\boldsymbol{\gamma}_0,n}\big)^T\big) - \tilde{Q}_{\boldsymbol{\gamma},n}\dot{\ell}^L_{\boldsymbol{\gamma}_0,n}\tilde{Q}_{\boldsymbol{\gamma},n}\big(\dot{\ell}^L_{\boldsymbol{\gamma}_0,n}\big)^T\big](\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \\
&\le (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T\tilde{Q}_{\boldsymbol{\gamma},n}\big(\dot{\ell}^L_{\boldsymbol{\gamma}_0,n}\big(\dot{\ell}^L_{\boldsymbol{\gamma}_0,n}\big)^T\big)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \\
&= \big(P_0(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\big)^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T P_0\big((p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n} - 1)\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big(\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big)^T\big)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \\
&\quad + \big(P_0(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\big)^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T P_0\big(\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big(\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}\big)^T\big)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0).
\end{aligned}$$

The first term on the right-hand side above is $o(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2)$ by similar argument as in (7.5). Then it follows that $\mathrm{Var}_{\tilde{Q}_{\boldsymbol{\gamma},n}} W_i \le s(\delta)\Delta(\boldsymbol{\gamma})$ for small enough ϵ and large enough L, where $s(\delta) \to 1$ as $\delta \to 0$ for $i = 1, \ldots, n$. We apply Bernstein's inequality to obtain

$$\begin{aligned}
Q^{(n)}_{\boldsymbol{\gamma},n}(1 - \omega_{1,n}) &= \|Q_{\boldsymbol{\gamma},n}\|^n \tilde{Q}^{(n)}_{\boldsymbol{\gamma},n}\big(W_1 + \cdots + W_n \ge nr(\delta)\Delta(\boldsymbol{\gamma})\big) \\
&\le \|Q_{\boldsymbol{\gamma},n}\|^n \exp\bigg(-\frac{1}{2}\frac{r^2(\delta)n\Delta(\boldsymbol{\gamma})}{s(\delta) + 1.5L\sqrt{p+1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|r(\delta)}\bigg).
\end{aligned}$$

We can make the factor $t(\delta) = r^2(\delta)(s(\delta) + 1.5L\sqrt{p+1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|r(\delta))^{-1}$ arbitrarily close to 1 for sufficiently small δ and ϵ. By Lemma 3.4 of [17], we have

$$\begin{aligned}
\|Q_{\boldsymbol{\gamma},n}\| &= 1 + P_0 \log\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}} + \frac{1}{2}P_0\bigg(\log\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}}\bigg)^2 + o\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big) \\
&\le 1 + P_0 \log\frac{p_{\boldsymbol{\gamma},n}}{p_{\boldsymbol{\gamma}_0,n}} + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T P_0\big(\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}\dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n}\big)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + o\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big) \\
&\le 1 - \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T\mathbf{V}_{\boldsymbol{\gamma}_0}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + \frac{1}{2}u(\delta)\Delta(\boldsymbol{\gamma}),
\end{aligned}$$

for some constant $u(\delta)$ such that $u(\delta) \to 1$ as $\delta \to 0$ for large n. Using the inequality $1 + x \le e^x$ for all $x \in \mathbb{R}$, we have, for sufficiently small $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|$,

$$Q^{(n)}_{\boldsymbol{\gamma},n}(1 - \omega_{1,n}) \le \exp\bigg(-\frac{n}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T\mathbf{V}_{\boldsymbol{\gamma}_0}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + \frac{n}{2}\big(u(\delta) - t(\delta)\big)\Delta(\boldsymbol{\gamma})\bigg).$$

Clearly, $u(\delta) - t(\delta) \to 0$ as $\delta \to 0$ and $\Delta(\boldsymbol{\gamma})$ is bounded above by a multiple of $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2$. Utilizing the positive definiteness of $\mathbf{V}_{\boldsymbol{\gamma}_0}$, we conclude that there exists a constant $C > 0$ such that for sufficiently large L and n and sufficiently small $\epsilon > 0$,

$$Q^{(n)}_{\boldsymbol{\gamma},n}(1 - \omega_{1,n}) \le \exp\big(-Cn\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2\big).$$

By the assumption of the theorem, there exists a consistent sequence of tests for P_0 versus $Q_{\boldsymbol{\gamma},n}$ for $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|>\epsilon$. Now by Lemma 3.3 of [17], there exists a sequence of tests $\{\omega_{2,n}\}$such that

$$P_0^{(n)}(\omega_{2,n}) \le \exp(-nC_1), \qquad \sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\ge\epsilon} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\omega_{2,n}) \le \exp(-nC_2).$$

We define a sequence $\{\psi_n\}$ as $\psi_n=\omega_{1,n}\vee\omega_{2,n}$ for all $n\ge 1$, in which case $P_0^{(n)}\psi_n \le P_0^{(n)}\omega_{1,n}+P_0^{(n)}\omega_{2,n}\to 0$ and

$$\begin{aligned}\sup_{\boldsymbol{\gamma}\in\Theta\times U} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\psi_n) &= \sup_{\boldsymbol{\gamma}\in(\Theta\times U)_1} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\psi_n)\vee \sup_{\boldsymbol{\gamma}\in(\Theta\times U)_2} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\psi_n)\\ &\le \sup_{\boldsymbol{\gamma}\in(\Theta\times U)_1} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\omega_{1,n})\vee \sup_{\boldsymbol{\gamma}\in(\Theta\times U)_2} Q_{\boldsymbol{\gamma},n}^{(n)}(1-\omega_{2,n}).\end{aligned}$$

Combining the previous bounds, we get the desired result for a suitable choice of $D>0$. □

Lemma 5. *There exists a constant $K>0$ such that the prior mass of the Kullback–Leibler neighborhoods $B(\epsilon_n,\boldsymbol{\gamma}_0,P_0)$ satisfies $\Pi(B(\epsilon_n,\boldsymbol{\gamma}_0,P_0))\ge K\epsilon_n^p$, where $\epsilon_n\gg n^{-1/2}$.*

Proof. From the proof of Lemma 2, we get

$$-P_0\log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}) = O\big(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2\big)+O\big(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|r_n^{-4}\big)\le c_1\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2+c_2\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\epsilon_n$$

for sufficiently large n and positive constants c_1 and c_2. Again, $P_0(\log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}))^2\le c_3\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2$ for some constant $c_3>0$. Let $c=\min((2c_1)^{-1/2},(2c_2)^{-1},c_3^{-1/2})$. Then $\{\boldsymbol{\gamma}\in\Theta\times U:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\le c\epsilon_n\}\subset B(\epsilon_n,\boldsymbol{\gamma}_0,P_0)$. Since the Lebesgue-density π of the prior is continuous and strictly positive in $\boldsymbol{\gamma}_0$, we see that there exists a $\delta'>0$ such that for all $0<\delta\le\delta'$, $\Pi(\boldsymbol{\gamma}\in\Theta\times U:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\le\delta)\ge\frac{1}{2}V\pi(\boldsymbol{\gamma}_0)\delta^{p+1}>0$, V being the Lebesgue-volume of the $(p+1)$-dimensional ball of unit radius. Hence, for sufficiently large n, $c\epsilon_n\le\delta'$ and we obtain the desired result. □

The next lemma is used to estimate the bias of the Bayes estimator in RKTB.

Lemma 6. *For $m\ge 2$ and $n^{1/(2m)}\ll k_n\ll n^{1/2}$,*

$$\sup_{t\in[0,1]}\big|\mathrm{E}\big(f(t)|\mathbf{X},\mathbf{Y},\sigma^2\big)-f_0(t)\big|^2 = O_{P_0}\big(k_n^2/n\big)+O_{P_0}\big(k_n^{1-2m}\big).$$

Proof. By (3.11),

$$\mathrm{E}\big(f(t)|\mathbf{X},\mathbf{Y},\sigma^2\big)=\big(\mathbf{N}(t)\big)^T\big(\mathbf{X}_n^T\mathbf{X}_n+k_nn^{-2}\mathbf{I}_{k_n+m-1}\big)^{-1}\mathbf{X}_n^T\mathbf{Y}. \tag{7.6}$$

By Lemma 12 in the Appendix, we have uniformly over $t\in[0,1]$,

$$\big(\mathbf{N}(t)\big)^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\mathbf{N}(t)\asymp\frac{k_n}{n}\big(1+o_{P_0}(1)\big). \tag{7.7}$$

Since $f_0 \in C^m$, there exists a $\boldsymbol{\beta}^*$ ([7], Theorem XII.4, page 178) such that

$$\sup_{t\in[0,1]} \left| f_0(t) - \left(\mathbf{N}(t)\right)^T \boldsymbol{\beta}^* \right| = O\left(k_n^{-m}\right). \tag{7.8}$$

We can bound $\sup_{t\in[0,1]} |\mathrm{E}(f(t)|\mathbf{X},\mathbf{Y},\sigma^2) - f_0(t)|^2$ up to a constant multiple by

$$\begin{aligned} &\sup_{t\in[0,1]} \left|\mathrm{E}\left(f(t)|\mathbf{X},\mathbf{Y},\sigma^2\right) - \left(N(t)\right)^T \left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\mathbf{Y}\right|^2 \\ &\quad + \sup_{t\in[0,1]} \left|\left(\mathbf{N}(t)\right)^T \left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\left(\mathbf{Y} - f_0(\mathbf{x})\right)\right|^2 \\ &\quad + \sup_{t\in[0,1]} \left|\left(\mathbf{N}(t)\right)^T \left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\left(f_0(\mathbf{x}) - \mathbf{X}_n\boldsymbol{\beta}^*\right)\right|^2 \\ &\quad + \sup_{t\in[0,1]} \left| f_0(t) - \left(\mathbf{N}(t)\right)^T \boldsymbol{\beta}^* \right|^2. \end{aligned} \tag{7.9}$$

Using the Binomial Inverse theorem, the Cauchy–Schwarz inequality and (7.7), the first term of (7.9) can be shown to be $O_{P_0}(k_n^6/n^8)$. The second term can be bounded up to a constant multiple by

$$\begin{aligned} &\max_{1\le k\le n} \left|\left(\mathbf{N}(s_k)\right)^T \left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\boldsymbol{\varepsilon}\right|^2 \\ &\quad + \sup_{t,t':|t-t'|\le n^{-1}} \left|\left(\mathbf{N}(t) - \mathbf{N}(t')\right)^T \left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\boldsymbol{\varepsilon}\right|^2, \end{aligned} \tag{7.10}$$

where $s_k = k/n$ for $k = 1,\ldots,n$. Applying the mean value theorem to the second term of the above sum, we can bound the expression by a constant multiple of

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

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

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

Note that $\mathrm{Var}_0(\mathbf{P}\boldsymbol{\varepsilon}) = \mathrm{E}_0(\mathrm{Var}(\mathbf{P}\boldsymbol{\varepsilon}|\mathbf{X})) + \mathrm{Var}_0(\mathrm{E}(\mathbf{P}\boldsymbol{\varepsilon}|\mathbf{X})) = \sigma_0^2\mathbf{I}_{k_n+m-1}$. Hence, we get $\mathrm{E}_0(\boldsymbol{\varepsilon}^T\mathbf{P}^T\mathbf{D}\mathbf{P}\boldsymbol{\varepsilon}) = \sigma_0^2(k_n + m - 1)$. In view of Lemma 12, we can conclude that the first term of (7.10) is $O_{P_0}(k_n^2/n)$. Again applying the Cauchy–Schwarz inequality, the second term of

(7.10) is bounded by

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

which is $O_{P_0}(n(k_n^3/n^3)) = O_{P_0}(k_n^3/n^2)$, using Lemma 12. Thus, the second term of (7.9) is $O_{P_0}(k_n^2/n)$. Using the Cauchy–Schwarz inequality, (7.7) and (7.8), the third term of (7.9) is $O_{P_0}(k_n^{1-2m})$. The fourth term of (7.9) is of the order of k_n^{-2m} as a result of (7.8). □

The following lemma controls posterior variability in RKTB.

Lemma 7. *If $m \geq 2$ and $n^{1/(2m)} \ll k_n \ll n^{1/2}$, then for all $\epsilon > 0$,*

$$\Pi_n^*\Big(\sup_{t\in[0,1]}\big|f(t) - f_0(t)\big| > \epsilon|\mathbf{X},\mathbf{Y},\sigma^2\Big) = o_{P_0}(1).$$

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 $\Pi_n^*(\sup_{t\in[0,1]}|f(t)-f_0(t)| > \epsilon|\mathbf{X},\mathbf{Y},\sigma^2)$ by

$$\begin{aligned}2\epsilon^{-2}\Big\{&\sup_{t\in[0,1]}\big|\mathrm{E}\big(f(t)|\mathbf{X},\mathbf{Y},\sigma^2\big) - f_0(t)\big|^2\\ &+\mathrm{E}\Big[\sup_{t\in[0,1]}\big|f(t)-\mathrm{E}\big(f(t)|\mathbf{X},\mathbf{Y},\sigma^2\big)\big|^2|\mathbf{X},\mathbf{Y},\sigma^2\Big]\Big\}.\end{aligned} \tag{7.11}$$

By Lemma 6, the first term inside the bracket above is $O_{P_0}(k_n^2/n) + O_{P_0}(k_n^{1-2m})$. For $\boldsymbol{\varepsilon}^* := (\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{1/2}\boldsymbol{\beta} - (\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{-1/2}\mathbf{X}_n^T\mathbf{Y}$, we have $\boldsymbol{\varepsilon}^*|\mathbf{X},\mathbf{Y},\sigma^2 \sim N(\mathbf{0},\sigma^2\mathbf{I}_{k_n+m-1})$. Writing

$$\sup_{t\in[0,1]}\big|f(t)-\mathrm{E}\big[f(t)|\mathbf{X},\mathbf{Y},\sigma^2\big]\big| = \sup_{t\in[0,1]}\big|\big(\mathbf{N}(t)\big)^T\big(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1}\big)^{-1/2}\boldsymbol{\varepsilon}^*\big|$$

and using the Cauchy–Schwarz inequality and Lemma 12, the second term inside the bracket in (7.11) is seen to be $O_{P_0}(k_n^2/n)$. By the assumed conditions on m and k_n, the lemma follows. □

The next lemma proves the posterior consistency of $\boldsymbol{\theta}$ using the results of Lemmas 6 and 7.

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

Proof. By the triangle inequality,

$$\begin{aligned}\big|R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})\big| &\leq \big\|f(\cdot)-f_0(\cdot)\big\|_g + \big\|f_{\boldsymbol{\eta},r_n}(\cdot) - f_{\boldsymbol{\eta}}(\cdot)\big\|_g\\ &\leq c_1'\sup_{t\in[0,1]}\big|f(t)-f_0(t)\big| + c_2' r_n^{-4},\end{aligned}$$

for appropriately chosen constants c_1' and c_2'. For a sequence $\tau_n \to 0$, define

$$T_n = \Big\{ f : \sup_{t\in[0,1]} \big| f(t) - f_0(t) \big| \le \tau_n \Big\}.$$

By Lemma 7, we can choose τ_n to satisfy $\Pi(T_n^c | \mathbf{X}, \mathbf{Y}, \sigma^2) = o_{P_0}(1)$. Hence for $f \in T_n$,

$$\sup_{\boldsymbol{\eta}\in\Theta} \big| R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta}) \big| \le c_1' \tau_n + c_2' r_n^{-4} = o(1).$$

Therefore, for any $\delta > 0$, $\Pi_n^*(\sup_{\boldsymbol{\eta}\in\Theta} |R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta | \mathbf{X}, \mathbf{Y}, \sigma^2) = o_{P_0}(1)$. By assumption (3.12), for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \ge \epsilon$ there exists a $\delta > 0$ such that

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

since $R_{f,n}(\boldsymbol{\theta}) \le R_{f,n}(\boldsymbol{\theta}_0)$. Consequently,

$$\begin{aligned} &\Pi_n^*\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon | \mathbf{X}, \mathbf{Y}, \sigma^2 \big) \\ &\quad \le \Pi_n^*\Big(\sup_{\boldsymbol{\eta}\in\Theta} \big| R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta}) \big| > \delta/2 | \mathbf{X}, \mathbf{Y}, \sigma^2 \Big) \\ &\quad = o_{P_0}(1). \end{aligned}$$

□

In the following lemma, we approximate $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ by a linear functional of f which is later used in Theorem 4.2 to obtain the limiting posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$.

Lemma 9. *Let m be an integer greater than or equal to 2 and $n^{1/(2m)} \ll k_n \ll n^{1/2}$. Then there exists $E_n \subseteq C^m((0,1)) \times \boldsymbol{\Theta}$ with $\Pi(E_n^c | \mathbf{X}, \mathbf{Y}, \sigma^2) = o_{P_0}(1)$, such that uniformly for $(f, \boldsymbol{\theta}) \in E_n$,*

$$\big\| \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{J}_{\boldsymbol{\theta}_0}^{-1} \sqrt{n}\big(\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0) \big) \big\| \lesssim \sqrt{n} r_n^{-4}, \tag{7.12}$$

where $\boldsymbol{\Gamma}(z) = \int_0^1 (\dot{f}_{\boldsymbol{\theta}_0}(t))^T z(t) g(t)\, dt$.

Proof. By definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$,

$$\int_0^1 \big(\dot{f}_{\theta, r_n}(t) \big)^T \big(f(t) - f_{\theta, r_n}(t) \big) g(t)\, dt = \mathbf{0}, \tag{7.13}$$

$$\int_0^1 \big(\dot{f}_{\theta_0}(t) \big)^T \big(f_0(t) - f_{\theta_0}(t) \big) g(t)\, dt = \mathbf{0}. \tag{7.14}$$

We can rewrite (7.13) as

$$\begin{aligned}&\int_0^1 \bigl(\dot f_{\theta_0}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt + \int_0^1 \bigl(\dot f_{\theta}(t) - \dot f_{\theta_0}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt \\ &\quad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t) - \dot f_{\theta}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt \\ &\quad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t)\bigr)^T \bigl(f_{\boldsymbol\theta}(t) - f_{\boldsymbol\theta,r_n}(t)\bigr) g(t)\,dt = \mathbf{0}.\end{aligned}$$

Subtracting (7.14) from the above equation, we get

$$\begin{aligned}&\int_0^1 \bigl(\dot f_{\theta_0}(t)\bigr)^T \bigl(f(t) - f_0(t)\bigr) g(t)\,dt - \int_0^1 \bigl(\dot f_{\theta_0}(t)\bigr)^T \bigl(f_{\boldsymbol\theta}(t) - f_{\boldsymbol\theta_0}(t)\bigr) g(t)\,dt \\ &\quad + \int_0^1 \bigl(\dot f_{\theta}(t) - \dot f_{\theta_0}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt \\ &\quad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t) - \dot f_{\theta}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt \\ &\quad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t)\bigr)^T \bigl(f_{\boldsymbol\theta}(t) - f_{\boldsymbol\theta,r_n}(t)\bigr) g(t)\,dt = \mathbf{0}.\end{aligned}$$

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

$$\begin{aligned}&\mathbf{M}(f,\boldsymbol\theta)(\boldsymbol\theta - \boldsymbol\theta_0) \\ &\quad = \int_0^1 \bigl(\dot f_{\theta_0}(t)\bigr)^T \bigl(f(t) - f_0(t)\bigr) g(t)\,dt \\ &\qquad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t) - \dot f_{\theta}(t)\bigr)^T \bigl(f(t) - f_{\boldsymbol\theta}(t)\bigr) g(t)\,dt \\ &\qquad + \int_0^1 \bigl(\dot f_{\theta,r_n}(t)\bigr)^T \bigl(f_{\boldsymbol\theta}(t) - f_{\boldsymbol\theta,r_n}(t)\bigr) g(t)\,dt,\end{aligned}$$

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

$$\begin{aligned}&\int_0^1 \int_0^1 \bigl(\dot f_{\theta_0}(t)\bigr)^T \dot f_{\theta_0+\lambda(\boldsymbol\theta-\boldsymbol\theta_0)}(t)\,d\lambda g(t)\,dt \\ &\quad - \int_0^1 \int_0^1 \ddot f_{\boldsymbol\theta_0+\lambda(\boldsymbol\theta-\boldsymbol\theta_0)}(t)\bigl(f_0(t) - f_{\theta_0}(t)\bigr)\,d\lambda g(t)\,dt \\ &\quad - \int_0^1 \int_0^1 \ddot f_{\boldsymbol\theta_0+\lambda(\boldsymbol\theta-\boldsymbol\theta_0)}(t)\bigl(f_{\boldsymbol\theta_0}(t) - f_{\boldsymbol\theta}(t)\bigr)\,d\lambda g(t)\,dt\end{aligned}$$

$$- \int_0^1 \int_0^1 \ddot{f}_{\boldsymbol{\theta}_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)\big(f(t)-f_0(t)\big)\,d\lambda g(t)\,dt$$
$$- \int_0^1 \int_0^1 \big(\dot{f}_\theta(t)-\dot{f}_{\theta_0}(t)\big)^T \dot{f}_{\theta_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)\,d\lambda g(t)\,dt.$$

For a sequence $\epsilon \to 0$, define

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

By Lemmas 7 and 8, we can choose ϵ_n so that $\Pi_n^*(E_n^c|\mathbf{X},\mathbf{Y},\sigma^2) = o_{P_0}(1)$. 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 $(f,\boldsymbol{\theta}) \in E_n$. Using (2.1), on E_n

$$\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$
$$= \big(\mathbf{J}_{\boldsymbol{\theta}_0}^{-1} + o(1)\big)\bigg(\sqrt{n}\int_0^1 \big(\dot{f}_{\theta_0}(t)\big)^T\big(f(t)-f_0(t)\big)g(t)\,dt + O\big(\sqrt{n}r_n^{-4}\big)\bigg).$$

Note that $\sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0)) = \sqrt{n}\mathbf{H}_n^T\boldsymbol{\beta} - \sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}(f_0)$. It was shown in the proof of Theorem 4.2 that for a given σ^2, the total variation distance between the posterior distribution of $\sqrt{n}\mathbf{H}_n^T\boldsymbol{\beta} - \sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}(f_0)$ and $N(\boldsymbol{\mu}_n,\sigma^2\boldsymbol{\Sigma}_n)$ converges in P_0-probability to 0. By Lemma 10, both $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are stochastically bounded. Thus the posterior distribution of $\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\sqrt{n}(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0))$ assigns most of its mass inside a large compact set with high true probability. □

The next lemma describes the asymptotic behavior of the mean and variance of the limiting normal distribution given by Theorem 4.2.

Lemma 10. *The mean and variance of the limiting normal approximation given by Theorem 4.2 are stochastically bounded.*

Proof. First, we study the asymptotic behavior of the matrix $\mathrm{Var}(\boldsymbol{\mu}_n|\mathbf{X}) = \boldsymbol{\Sigma}_n = n\mathbf{H}_n^T \times (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{H}_n$. If $C_k(\cdot) \in C^{m^*}(0,1)$ for some $1 \le m^* < m$, then by equation (2) of [7], page 167, we have for all $k = 1,\dots,p$,

$$\sup\big\{\big|C_k(t)-\tilde{C}_k(t)\big| : t\in[0,1]\big\} = O\big(k_n^{-1}\big),$$

where $\tilde{C}_k(\cdot) = \boldsymbol{\alpha}_k^T\mathbf{N}(\cdot)$ and $\boldsymbol{\alpha}_k^T = (C_k(t_1^*),\dots,C_k(t_{k_n+m-1}^*))$ with appropriately chosen $t_1^*,\dots,t_{k_n+m-1}^*$. We can write $\mathbf{H}_n^T(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{H}_n$ as

$$(\mathbf{H}_n-\tilde{\mathbf{H}}_n)^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}(\mathbf{H}_n-\tilde{\mathbf{H}}_n) + \tilde{\mathbf{H}}_n^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\tilde{\mathbf{H}}_n$$
$$+ (\mathbf{H}_n-\tilde{\mathbf{H}}_n)^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\tilde{\mathbf{H}}_n + \tilde{\mathbf{H}}_n^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}(\mathbf{H}_n-\tilde{\mathbf{H}}_n),$$

where the kth row of $\tilde{\mathbf{H}}_n^T$ is given by $\int_0^1 \tilde{C}_k(t)(\mathbf{N}(t))^T g(t)\,dt$ for $k = 1, \ldots, p$. Let us denote $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$. Then

$$\begin{aligned}&\tilde{\mathbf{H}}_n^T \bigl(\mathbf{X}_n^T \mathbf{X}_n\bigr)^{-1} \tilde{\mathbf{H}}_n \\ &\quad = n^{-1} \mathbf{A}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\biggl(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n}\biggr)^{-1}\biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{A}.\end{aligned}$$

We show that

$$\begin{aligned}&\mathbf{A}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\biggl(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n}\biggr)^{-1}\biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{A} \\ &\quad - \mathbf{A}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{A}\end{aligned}$$

converges in P_0-probability to the null matrix of order p. For a $\mathbf{l} \in \mathbb{R}^p$, let $\mathbf{c} = (\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) \times g(t)\,dt)\mathbf{A}\mathbf{l}$. Then we can write

$$\mathbf{l}^T \mathbf{A}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\biggl(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n}\biggr)^{-1}\biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{A}\mathbf{l}$$

as $\mathbf{c}^T(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n})^{-1}\mathbf{c}$. Let us denote by Q_n the empirical distribution function of $X_1, \ldots, X_n$. Note that

$$\begin{aligned}&\biggl|\mathbf{c}^T \biggl(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n}\biggr)\mathbf{c} - \mathbf{c}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{c}\biggr| \\ &\quad \le \sup_{t\in[0,1]} \bigl|Q_n(t) - G(t)\bigr|\mathbf{c}^T\mathbf{c} \sup_{t\in[0,1]} \bigl\|\mathbf{N}(t)\bigr\|^2 \\ &\quad = O_{P_0}\bigl(n^{-1/2}\bigr)\mathbf{c}^T\mathbf{c} \\ &\quad = O_{P_0}\bigl(n^{-1/2}k_n\bigr)\mathbf{c}^T \biggl(\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt\biggr)\mathbf{c},\end{aligned}$$

the third step following from Donsker's theorem and the fact that

$$\sup_{t\in[0,1]} \bigl\|\mathbf{N}(t)\bigr\|^2 \le 1.$$

In the last step we used the fact that the eigenvalues of the matrix $\int_0^1 \mathbf{N}(t)\mathbf{N}^T(t) g(t)\,dt$ are $O(k_n^{-1})$ as proved in Lemma 6.1 of [29]. In that same lemma, it was also proved that the eigenvalues of the matrix $(\mathbf{X}_n^T \mathbf{X}_n/n)$ are $O_{P_0}(k_n^{-1})$. Both these results are applied in the fourth step of the next calculation. Using the fact that $\|\mathbf{R}^{-1} - \mathbf{S}^{-1}\| \le \|\mathbf{S}^{-1}\|\|\mathbf{R} - \mathbf{S}\|\|\mathbf{S}^{-1}\|$ for two nonsingular

matrices $\mathbf{R}$ and $\mathbf{S}$ of the same order, we get

$$\begin{aligned}
&\left|\mathbf{c}^T\left(\frac{\mathbf{X}_n^T\mathbf{X}_n}{n}\right)^{-1}\mathbf{c}-\mathbf{c}^T\left(\int_0^1\mathbf{N}(t)\mathbf{N}^T(t)g(t)\,dt\right)^{-1}\mathbf{c}\right|\\
&\quad=O_{P_0}\bigl(n^{-1/2}k_n\bigr)\mathbf{c}^T\left(\frac{\mathbf{X}_n^T\mathbf{X}_n}{n}\right)^{-1}\mathbf{c}\\
&\quad=O_{P_0}\bigl(n^{-1/2}k_n\bigr)\mathbf{l}^T\mathbf{A}^T\left(\int_0^1\mathbf{N}(t)\mathbf{N}^T(t)g(t)\,dt\right)\left(\frac{\mathbf{X}_n^T\mathbf{X}_n}{n}\right)^{-1}\\
&\qquad\times\left(\int_0^1\mathbf{N}(t)\mathbf{N}^T(t)g(t)\,dt\right)\mathbf{A}\mathbf{l}\\
&\quad=O_{P_0}\bigl(n^{-1/2}k_n\bigr)k_n^{-1}\mathbf{l}^T\mathbf{A}^T\mathbf{A}\mathbf{l}=o_{P_0}(1).
\end{aligned}$$

Now note that the (i,j)th element of the $p\times p$ matrix $\mathbf{A}^T(\int_0^1\mathbf{N}(t)\mathbf{N}^T(t)g(t)\,dt)\mathbf{A}$ is given by $\int_0^1\tilde{C}_i(t)\tilde{C}_j(t)g(t)\,dt$, which converges to $\int_0^1 C_i(t)C_j(t)g(t)\,dt$, the (i,j)th element of the matrix $\int_0^1\mathbf{C}(t)\mathbf{C}^T(t)g(t)\,dt$ which is $\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\int_0^1(\dot{f}_{\boldsymbol{\theta}_0}(t))^T\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt(\mathbf{J}_{\boldsymbol{\theta}_0}^{-1})^T$. Let us denote by $\mathbf{1}_{k_n+m-1}$ the (k_n+m-1)-component vector with all elements 1. Then for $k=1,\ldots,p$, the kth diagonal entry of the matrix $(\mathbf{H}_n-\tilde{\mathbf{H}}_n)^T(\mathbf{X}_n^T\mathbf{X}_n)^{-1}(\mathbf{H}_n-\tilde{\mathbf{H}}_n)$ is given by

$$\begin{aligned}
&\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\bigl(\mathbf{N}(t)\bigr)^Tg(t)\,dt\bigl(\mathbf{X}_n^T\mathbf{X}_n\bigr)^{-1}\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\bigl(\mathbf{N}(t)\bigr)g(t)\,dt\\
&\quad=\frac{1}{n}\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\bigl(\mathbf{N}(t)\bigr)^Tg(t)\,dt\bigl(\mathbf{X}_n^T\mathbf{X}_n/n\bigr)^{-1}\\
&\qquad\times\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\mathbf{N}(t)g(t)\,dt\\
&\quad\asymp\frac{k_n}{n}\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\bigl(\mathbf{N}(t)\bigr)^Tg(t)\,dt\\
&\qquad\times\int_0^1\bigl(C_k(t)-\tilde{C}_k(t)\bigr)\mathbf{N}(t)g(t)\,dt\\
&\quad\lesssim\frac{1}{nk_n},
\end{aligned}$$

the last step following by the application of the Cauchy–Schwarz inequality and the facts that $\sup\{|C_k(t)-\tilde{C}_k(t)|:t\in[0,1]\}=O(k_n^{-1})$ and $\int_0^1\|\mathbf{N}(t)\|^2\,dt\le 1$. Thus, the eigenvalues of $(\mathbf{H}_n-\tilde{\mathbf{H}}_n)^T(\mathbf{X}_n^T\mathbf{X}_n)^{-1}(\mathbf{H}_n-\tilde{\mathbf{H}}_n)$ are of the order $(nk_n)^{-1}$ or less. Hence,

$$n\mathbf{H}_n^T\bigl(\mathbf{X}_n^T\mathbf{X}_n\bigr)^{-1}\mathbf{H}_n\overset{P_0}{\to}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\int_0^1\bigl(\dot{f}_{\boldsymbol{\theta}_0}(t)\bigr)^T\dot{f}_{\boldsymbol{\theta}_0}(t)g(t)\,dt\bigl(\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\bigr)^T.$$

Thus, the eigenvalues of $\mathbf{\Sigma}_n$ are stochastically bounded. Now note that

$$\begin{aligned}\mathrm{E}(\boldsymbol{\mu}_n|\mathbf{X}) &= \sqrt{n}\mathbf{H}_n^T\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T f_0(\mathbf{X}) - \sqrt{n}\mathbf{J}_{\boldsymbol{\theta}_0}^{-1}\mathbf{\Gamma}(f_0)\\ &= \sqrt{n}\mathbf{H}_n^T\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T\left(f_0(\mathbf{X}) - \mathbf{X}_n\boldsymbol{\beta}^*\right)\\ &\quad + \sqrt{n}\int_0^1 \mathbf{C}(t)\big(\mathbf{N}^T(t)\boldsymbol{\beta}^* - f_0(t)\big)g(t)\,dt.\end{aligned}$$

Using the Cauchy–Schwarz inequality and (7.8), we get

$$\begin{aligned}\left\|\mathrm{E}(\boldsymbol{\mu}_n|\mathbf{X})\right\| &\lesssim \sqrt{n}\,\mathrm{maxeig}\left(\mathbf{H}_n^T\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{H}_n\right)^{1/2}\sqrt{n}k_n^{-m} + \sqrt{n}k_n^{-m}\\ &= O_{P_0}\left(\sqrt{n}k_n^{-m}\right) = o_{P_0}(1).\end{aligned}$$

Thus, $Z_n := \|\mathrm{E}(\boldsymbol{\mu}_n|\mathbf{X})\|^2 + \mathrm{maxeig}(\mathrm{Var}(\boldsymbol{\mu}_n|\mathbf{X}))$ is stochastically bounded. Given $M > 0$, there exists $L > 0$ such that $\sup_n P_0(Z_n > L) < M^{-2}$. Hence for all n, $P_0(\|\boldsymbol{\mu}_n\| > M)$ is bounded above by $M^{-2}\mathrm{E}_0[\mathrm{E}(\|\boldsymbol{\mu}_n\|^2|\mathbf{X})\mathbb{1}\{Z_n \le L\}] + P_0(Z_n > L)$ which is less than or equal to $(L+1)/M^2$. Hence, $\boldsymbol{\mu}_n$ is stochastically bounded. □

In the next lemma, we establish the posterior consistency of σ^2.

Lemma 11. *For all $\epsilon > 0$, we have $\Pi_n^*(|\sigma^2 - \sigma_0^2| > \epsilon|\mathbf{X}, \mathbf{Y}) = o_{P_0}(1)$.*

Proof. The joint density of $\mathbf{Y}$, $\boldsymbol{\beta}$ and σ^2 is proportional to

$$\begin{aligned}&\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}_n\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}_n\boldsymbol{\beta})\right\}\\ &\quad\times\sigma^{-k_n-m+1}\exp\left\{-\frac{1}{2n^2k_n^{-1}\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\}\exp\left(-\frac{b}{\sigma^2}\right)\left(\sigma^2\right)^{-a-1},\end{aligned}$$

which implies that the posterior distribution of σ^2 is inverse gamma with shape parameter $n/2+a$ and scale parameter

$$\frac{1}{2}\left\{\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}_n\left(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1}\right)^{-1}\mathbf{X}_n^T\mathbf{Y}\right\} + b.$$

Hence, the posterior mean of σ^2 is given by

$$\mathrm{E}\left(\sigma^2|\mathbf{X},\mathbf{Y}\right) = \frac{\frac{1}{2}\{\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{-1}\mathbf{X}_n^T\mathbf{Y}\} + b}{n/2 + a - 1},$$

which behaves like the $n^{-1}(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{-1}\mathbf{X}_n^T\mathbf{Y})$ asymptotically. The later can be written as

$$n^{-1}\left(\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\mathbf{Y} + \mathbf{Y}^T\left(\mathbf{P}_{\mathbf{X}_n} - \mathbf{X}_n\left(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1}\right)^{-1}\mathbf{X}_n^T\right)\mathbf{Y}\right),$$

where $\mathbf{P}_{\mathbf{X}_n} = \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T$. We will show that $n^{-1}\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\mathbf{Y} \xrightarrow{P_0} \sigma_0^2$ and $n^{-1}\mathbf{Y}^T(\mathbf{P}_{\mathbf{X}_n} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{-1}\mathbf{X}_n^T)\mathbf{Y} = o_{P_0}(1)$ and hence $\mathrm{E}(\sigma^2|\mathbf{X},\mathbf{Y}) \xrightarrow{P_0} \sigma_0^2$. Using $\mathbf{Y} = f_0(\mathbf{X}) + \boldsymbol{\varepsilon}$, we note that

$$\begin{aligned}\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\mathbf{Y} &= \boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon} + f_0(\mathbf{X})^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X}) \\ &\quad + 2\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X}).\end{aligned}$$

We show that $\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}/n \xrightarrow{P_0} \sigma_0^2$, $n^{-1}f_0(\mathbf{X})^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X}) = o_{P_0}(1)$ and $n^{-1}\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X}) = o_{P_0}(1)$. Now, $\mathrm{E}_0(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}/n) \to \sigma_0^2$ as $n \to \infty$. Also,

$$\begin{aligned}\mathrm{Var}_0\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}/n\big) &= n^{-2}\big(\mathrm{E}_0\,\mathrm{Var}\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}|\mathbf{X}\big) \\ &\quad + \mathrm{Var}_0\,\mathrm{E}\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}|\mathbf{X}\big)\big).\end{aligned}$$

Now

$$\begin{aligned}\mathrm{Var}\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}|\mathbf{X}\big) &= \big(\mu_4 - \sigma_0^2\big)(n - k_n - m + 1) \\ \mathrm{E}\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}|\mathbf{X}\big) &= \sigma_0^2(n - k_n - m + 1),\end{aligned}$$

μ_4 being the fourth order central moment of ε_i for $i = 1, \ldots, n$. Hence,

$$\mathrm{Var}_0\big(\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}/n\big) \to 0 \qquad \text{as } n \to \infty.$$

Thus, $\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\boldsymbol{\varepsilon}/n \xrightarrow{P_0} \sigma_0^2$. We can write for $\boldsymbol{\beta}^*$ satisfying (7.8)

$$\begin{aligned}f_0(\mathbf{X})^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X}) &= \big(f_0(\mathbf{X}) - \mathbf{X}_n\boldsymbol{\beta}^*\big)^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\big(f_0(\mathbf{X}) - \mathbf{X}_n\boldsymbol{\beta}^*\big) \\ &\lesssim nk_n^{-2m},\end{aligned}$$

since $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\mathbf{X}_n = \mathbf{0}$. Using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}\big|n^{-1}\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})f_0(\mathbf{X})\big| &= \big|n^{-1}\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_n})\big(f_0(\mathbf{X}) - \mathbf{X}_n\boldsymbol{\beta}^*\big)\big| \\ &\le \sqrt{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/n}\,k_n^{-m} = o_{P_0}(1).\end{aligned}$$

By the Binomial Inverse theorem,

$$\begin{aligned}\mathbf{P}_{\mathbf{X}_n} - \mathbf{X}_n\big(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1}\big)^{-1}\mathbf{X}_n^T &= k_n n^{-2}\mathbf{X}_n\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1} \\ &\quad \times \left(\mathbf{I}_{k_n+m-1} + \big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\frac{k_n}{n^2}\right)^{-1} \\ &\quad \times \big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\mathbf{X}_n^T\end{aligned}$$

whose eigenvalues are of the order $k_n n^{-2} n k_n^{-1} k_n^2 n^{-2} = k_n^2 n^{-3}$. Hence, the random variable $\mathbf{Y}^T(\mathbf{P}_{\mathbf{X}_n} - \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n + k_n n^{-2}\mathbf{I}_{k_n+m-1})^{-1}\mathbf{X}_n^T)\mathbf{Y}/n$ converges in P_0-probability to 0 and $\mathrm{E}(\sigma^2|\mathbf{X},\mathbf{Y}) \xrightarrow{P_0} \sigma_0^2$. Also,

$$\mathrm{Var}\big(\sigma^2|\mathbf{X},\mathbf{Y}\big) = \big(\mathrm{E}\big(\sigma^2|\mathbf{X},\mathbf{Y}\big)\big)^2/(n/2+a-2) = o_{P_0}(1).$$

By using the Markov's inequality, we finally get $\Pi_n^*(|\sigma^2 - \sigma_0^2| > \epsilon|\mathbf{X},\mathbf{Y}) = o_{P_0}(1)$ for all $\epsilon > 0$. □

Appendix

The following result was used to prove Lemmas 6, 7and 10.

Lemma 12. *For any* $0 \le r \le m-2$, *there exist constants* $L_{\max} > L_{\min} > 0$ *such that uniformly in* $t \in [0,1]$,

$$\frac{L_{\min}\sigma^2 k_n^{2r+1}}{n}\big(1+o_{P_0}(1)\big) \le \big(\mathbf{N}^{(r)}(t)\big)^T\big(\mathbf{X}_n^T\mathbf{X}_n\big)^{-1}\mathbf{N}^{(r)}(t) \le \frac{L_{\max}\sigma^2 k_n^{2r+1}}{n}\big(1+o_{P_0}(1)\big).$$

The proof is implicit in Lemma 5.4 of [30].

References

[1] Anderson, R.M. and May, R.M. (1992). *Infectious Diseases of Humans*: *Dynamics and Control*. London: Oxford Univ. Press.

[2] Bhaumik, P. and Ghosal, S. (2015). Bayesian two-step estimation in differential equation models. *Electron. J. Stat.* **9** 3124–3154. MR3453972

[3] Brunel, N.J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2** 1242–1267. MR2471285

[4] Brunel, N.J.-B., Clairon, Q. and d'Alché-Buc, F. (2014). Parametric estimation of ordinary differential equations with orthogonality conditions. *J. Amer. Statist. Assoc.* **109** 173–185. MR3180555

[5] Campbell, D. and Steele, R.J. (2012). Smooth functional tempering for nonlinear differential equation models. *Stat. Comput.* **22** 429–443. MR2865027

[6] Chen, T., He, H.L. and Church, G.M. (1999). Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing* **4** 4.

[7] de Boor, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. New York: Springer. MR0507062

[8] Gabrielsson, J. and Weiner, D. (2000). *Pharmacokinetic and Pharmacodynamic Data Analysis*: *Concepts and Applications*. London: Taylor & Francis.

[9] Gelman, A., Bois, F. and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Statist. Assoc.* **91** 1400–1412.

[10] Ghosh, S.K. and Goyal, L. (2010). Statistical inference for non-linear models involving ordinary differential equations. *J. Stat. Theory Pract.* **4** 727–742. MR2758756

[11] Girolami, M. (2008). Bayesian inference for differential equations. *Theoret. Comput. Sci.* **408** 4–16. MR2460604
[12] Gugushvili, S. and Klaassen, C.A.J. (2012). $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: Bypassing numerical integration via smoothing. *Bernoulli* **18** 1061–1098. MR2948913
[13] Hairer, E., Nørsett, S.P. and Wanner, G. (1993). *Solving Ordinary Differential Equations. I: Nonstiff Problems*, 2nd ed. *Springer Series in Computational Mathematics* **8**. Berlin: Springer. MR1227985
[14] Henrici, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations*. New York: Wiley. MR0135729
[15] Jaeger, J. (2012). Functional estimation in systems defined by differential equation using Bayesian smoothing methods. Ph.D. thesis, UCL.
[16] Kleijn, B.J.K. and van der Vaart, A.W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395
[17] Kleijn, B.J.K. and van der Vaart, A.W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381. MR2988412
[18] Mattheij, R. and Molenaar, J. (2002). *Ordinary Differential Equations in Theory and Practice. Classics in Applied Mathematics* **43**. Philadelphia, PA: SIAM. MR1946758
[19] Nowak, M.A. and May, R.M. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford: Oxford Univ. Press. MR2009143
[20] Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.* **38** 435–481. MR2589327
[21] Ramsay, J.O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 741–796. MR2368570
[22] Rodriguez-Fernandez, M., Egea, J.A. and Banga, J.R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* **7** 483.
[23] Rogers, S., Khanin, R. and Girolami, M. (2007). Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics* **8** S2.
[24] Storn, R. and Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11** 341–359. MR1479553
[25] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247
[26] Varah, J.M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.* **3** 28–46. MR0651865
[27] Wu, H., Xue, H. and Kumar, A. (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
[28] Xue, H., Miao, H. and Wu, H. (2010). 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** 2351–2387. MR2676892
[29] Zhou, S., Shen, X. and Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277
[30] Zhou, S. and Wolfe, D.A. (2000). On derivative estimation in spline regression. *Statist. Sinica* **10** 93–108. MR1742102

Received November 2014 and revised April 2016