ABSTRACT

BHAUMIK, PRITHWISH. Bayesian Estimation and Uncertainty Quantification in Differential Equation Models. (Under the direction of Subhashis Ghoshal.)

In engineering, physics, biomedical sciences, pharmacokinetics and pharmacodynamics (PKPD) and many other fields the regression function is often specified as solution of a system of ordinary differential equations (ODEs) given by

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

here $\boldsymbol{F}$ is a known appropriately smooth vector valued function. Our interest lies in estimating $\boldsymbol{\theta}$ from the noisy data.

A two-step approach to solve this problem consists of the first step fitting the data nonparametrically, and the second step estimating the parameter by minimizing the distance between the nonparametrically estimated derivative and the derivative suggested by the system of ODEs. In Chapter 2 we consider a Bayesian analog of the two step approach by putting a finite random series prior on the regression function using B-spline basis. We establish a Bernstein-von Mises theorem for the posterior distribution of the parameter of interest induced from that on the regression function with the $n^{-1/2}$ contraction rate.

Although this approach is computationally fast, the Bayes estimator is not asymptotically efficient. This can be remedied by directly considering the distance between the function in the nonparametric model and a Runge-Kutta (RK4) approximate solution of the ODE while inducing the posterior distribution on the parameter as done in Chapter 3. We also study the asymptotic properties of a direct Bayesian method obtained from the approximate likelihood obtained by the RK4 method in Chapter 3.

Chapters 4 and 5 contain the extensions of the methods discussed so far for higher order ODE's and partial differential equations (PDE's) respectively.

We have mentioned the scopes of some future works in Chapter 6.

Bayesian Estimation and Uncertainty Quantification in
Differential Equation Models

by
Prithwish Bhaumik

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2015

APPROVED BY:

Soumendra Nath Lahiri	Alyson Wilson
Yichao Wu	Subhashis Ghoshal Chair of Advisory Committee

DEDICATION

To my parents and all well wishers.

BIOGRAPHY

Prithwish Bhaumik was born on 11th of September, 1987 in Kolkata, India. After finishing his schooling at Ballygunge Govt. High School, he went to St. Xavier's College, Kolkata to pursue Bachelors degree in Statistics. He received Masters in Statistics from Indian Statistical Institute, Kolkata. Then he joined the Department of Statistics at North Carolina State University as a Ph.D. student. Prithwish will be joining the University of Texas at Austin as a post-doctoral fellow from July, 2015.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Subhashis Ghoshal for guiding me to lay the foundation of my research career. His constant help has made it possible for me to finish my doctoral work in a proper and timely fashion. I would also like to express my sincere gratitude to Dr. Soumendra Nath Lahiri, Dr. Alyson Wilson and Dr. Yichao Wu for agreeing to be in my graduate advisory committee.

The Statistics department of North Carolina State university has provided me with all the kinds of help a graduate student can look for. The faculty members have been extremely helpful and easily approachable. I would specifically mention Dr. Sujit Ghosh, Dr. Donald Martin and Dr. Arnab Maity in this regard. I have also received a lot of help from the department staff members. I am extremely grateful to Alison McCoy and Terry Byron who have always answered all my queries with a smiling face.

My four years of stay in the USA has been a memorable chapter of my life. I have been fortunate enough to have a wonderful group of friends including Souvik Chandra, Sayantan Banerjee, Ayan Das Gupta, Shuva Gupta, Lopamudra Kundu, Ritwik Chatterjee, Saswata Sahoo, Priyam Das, Debraj Das, Arnab Hazra, Arkaprava Roy, Indrani Sahoo, Moumita Chakraborty and Sohini Raha.

Finally I owe all of my achievements to my parents. Their constant guidance and support has made it possible for me to stay focussed and reach my target.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$:=$ equality by definition.

$(\!(A_{i,j})\!)$: a matrix $\boldsymbol{A}$ with $(i,j)^{th}$ element being $A_{i,j}$.

$\boldsymbol{A}^T$: transpose of the matrix $\boldsymbol{A}$.

$\boldsymbol{A}_{i,}$: the i^{th} row of $\boldsymbol{A}$.

$\boldsymbol{A}_{,j}$: the j^{th} column of $\boldsymbol{A}$.

$\text{rows}_r^s(\boldsymbol{A})$: the sub-matrix of $\boldsymbol{A}$ consisting of r^{th} to s^{th} rows of $\boldsymbol{A}$ with $r < s$.

$\text{cols}_r^s(\boldsymbol{A})$: the sub-matrix of $\boldsymbol{A}$ consisting of r^{th} to s^{th} columns of $\boldsymbol{A}$ with $r < s$.

$\boldsymbol{x}_{r:s}$: the sub-vector consisting of r^{th} to s^{th} elements of a vector $\boldsymbol{x}$.

$\text{vec}(\boldsymbol{A})$: the vector obtained by stacking the columns of the matrix $\boldsymbol{A}$ one over another.

$\boldsymbol{A} \otimes \boldsymbol{B}$: the Kronecker product between $\boldsymbol{A}$ and $\boldsymbol{B}$.

$\boldsymbol{I}_p$: the identity matrix of order p.

$\text{maxeig}(\boldsymbol{A})$: the maximum eigenvalue of the matrix $\boldsymbol{A}$.

$\text{mineig}(\boldsymbol{A})$: the minimum eigenvalue of the matrix $\boldsymbol{A}$.

$\|\boldsymbol{x}\|$: $(\sum_{i=1}^{p} x_i^2)^{1/2}$, the L_2 norm of the vector $\boldsymbol{x}$.

$f'(t)$: $\frac{d}{dt}f(t)$, the derivative of the function $f(\cdot)$.

$f^{(r)}(t)$: $\frac{d^r}{dt^r}f(t)$, the r^{th} order derivative of the function $f(\cdot)$.

$\boldsymbol{f}(\cdot)$: a vector valued function.

$\|\boldsymbol{f}\|_w$: $(\int_0^1 \|\boldsymbol{f}(t)\|^2 w(t)dt)^{1/2}$ for functions $\boldsymbol{f} : [0,1] \to \mathbb{R}^p$ and $w : [0,1] \to [0,\infty)$.

$f(\boldsymbol{x})$: $(f(x_1),\ldots,f(x_p))^T$ for a real-valued function $f : [0,1] \to \mathbb{R}$ and a vector $\boldsymbol{x} \in \mathbb{R}^p$.

$\langle \cdot,\cdot \rangle$: inner product.

$\mathbb{1}_A(\cdot)$: indicator function of the set A.

$a_n = o(b_n)$: $a_n/b_n \to 0$ as $n \to \infty$ for numerical sequences a_n and b_n.

$a_n = O(b_n)$: a_n/b_n is bounded.

$a_n \asymp b_n$: $a_n = O(b_n)$ and $b_n = O(a_n)$.

$a_n \lesssim b_n$: $a_n = O(b_n)$.

$a_n \gg b_n$: $b_n = o(a_n)$.

$a_n \ll b_n$: $a_n = o(b_n)$.

$o_P(1)$: a sequence of random variables which converges in probability to zero.

$O_P(1)$: a sequence of random variables bounded in probability.

$\mathbf{E}(\cdot)$: the mean vector of a random vector.

$\mathbf{Var}(\cdot)$: the dispersion matrix of a random vector.

$\|P-Q\|_{TV}$: $\sup_{B\in\mathscr{R}^p}|P(B)-Q(B)|$, the total variation distance between P and Q.
$C^m(E)$: the collection of functions defined on an open set E with first m continuous partial derivatives with respect to its arguments.
$\dot{f}_{\boldsymbol{\theta}}(x)$: $\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x)$ for the function $\boldsymbol{\theta}\mapsto f_{\boldsymbol{\theta}}(x)$.
$\ddot{f}_{\boldsymbol{\theta}}(x)$: $\frac{\partial^2}{\partial\boldsymbol{\theta}^2}f_{\boldsymbol{\theta}}(x)$ for the function $\boldsymbol{\theta}\mapsto f_{\boldsymbol{\theta}}(x)$.
$\mathbb{P}_n\psi$: $n^{-1}\sum_{i=1}^n\psi(\boldsymbol{X}_i)$ for a sample $\{\boldsymbol{X}_i:\ i=1,\ldots,n\}$ and a measurable function $\psi(\cdot)$.
$\mathbb{G}_n\psi$: $\sqrt{n}\,(\mathbb{P}_n\psi-\mathrm{E}\psi)$.
$a\wedge b$: the minimum of two real numbers a and b.
$D^{\boldsymbol{r}}f(\boldsymbol{t})$: $\frac{\partial^{r_1+\cdots+r_s}}{\partial t_1^{r_1}\cdots\partial t_s^{r_s}}f(\boldsymbol{t})$ for a vector $\boldsymbol{r}=(r_1,\ldots,r_s)^T$ of nonnegative integers and a function $f:\mathbb{R}^s\mapsto\mathbb{R}$.
$|\boldsymbol{r}|$: $\sum_{j=1}^s r_j$ for a vector $\boldsymbol{r}=(r_1,\ldots,r_s)^T$.

Chapter 1

Introduction

Differential equations are encountered in various branches of science such as in genetics (Chen et al., 1999), viral dynamics of infectious diseases [Anderson and May (1992), Nowak and May (2000)]. Diggle (1990) contains a data set on the growth of three closed colonies of paramecium aurelium in a nutritive medium over a period of 19 days. The logarithm of the growth for the three colonies are plotted in Figure 1.1. Denoting by $\mu(t)$ as the logarithm of growth at time t, Ghosh and Goyal (2010) suggested the possible ordinary differential equation (ODE) model as

$$\begin{aligned} \frac{d\mu(t)}{dt} &= \theta_1 - \theta_2 \exp\{\mu(t)\}, \\ \mu(0) &= \log 2. \end{aligned}$$

There are numerous applications in the fields of pharmacokinetics and pharmacodynamics (PKPD) as well. There are a lot of instances where no closed form solution exist. Such an example can be found in the feedback system (Gabrielsson and Weiner, 2006, page 332) modeled by the ODEs

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

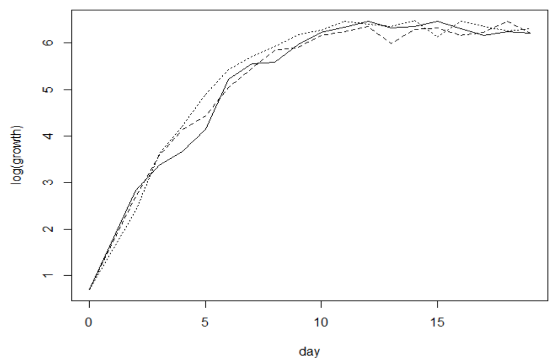

Figure 1.1: *Logarithm of growth over time of the three colonies*

where $R(t)$ and $M(t)$ stand for loss of response and modulator at time t respectively. Here $k_{\text{in}}, k_{\text{out}}$ and k_{tol} are unknown parameters which have to be estimated from the noisy observations. Another popular example is the Lotka-Volterra equations, also known as predator-prey equations. At time $t \in [0,1]$ the prey and predator populations change according to the equations

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

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. A noisy data from this model is shown in Figure 1.2. These models can be put in a regression model

$$\boldsymbol{Y}_i = \boldsymbol{f}_{\boldsymbol{\theta}}(t_i) + \boldsymbol{\varepsilon}_i, \tag{1.1}$$

for $i = 1, \ldots, n$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ and $\boldsymbol{f}_{\boldsymbol{\theta}} : [0,1] \mapsto \mathbb{R}^d$ and $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ satisfies the ODE

$$\frac{d\boldsymbol{f}_{\boldsymbol{\theta}}(t)}{dt} = \boldsymbol{F}(t, \boldsymbol{f}_{\boldsymbol{\theta}}(t), \boldsymbol{\theta}),\ t \in [0,1]; \tag{1.2}$$

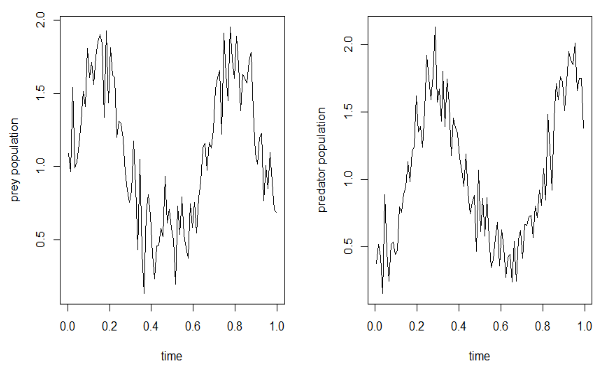

Figure 1.2: *Prey and predator population following the Lotka-Volterra equations*

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

Some physical events can be described using higher order ODE's. Let $G(t)$ and $H(t)$ respectively be the concentrations of glucose and hormone in blood at time t. Denoting by $G_0(t)$ and $H_0(t)$ the corresponding optimal values at time t, we are interested in studying the behavior of $g(t) = G(t) - G_0(t)$ and $h(t) = H(t) - H_0(t)$. These quantities change according to the system of ODE's

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

$J(t)$ being a known function denoting the external rate of increase of blood glucose concentration. Here m_1, m_2, m_3 and m_4 are unknown parameters. Since we cannot measure $h(t)$, the two equations are combined to get the second order ODE

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

where $\alpha = (m_1 + m_3)/2$, $\omega_0^2 = m_1 m_3 + m_2 m_4$ and $S(t) = m_3 J(t) + J'(t)$. More generally, we often come across higher order ODE models of the form

$$\boldsymbol{F}\left(t, \boldsymbol{f_\theta}(t), \boldsymbol{f_\theta}^{(1)}(t), \ldots, \boldsymbol{f_\theta}^{(q)}(t), \boldsymbol{\theta}\right) = \boldsymbol{0}, \tag{1.3}$$

$\boldsymbol{F}$ being a known sufficiently smooth vector valued function.

There are also instances when we encounter partial differential equation (PDE) models. For example Xun et al. (2013) addressed a problem using the PDE model

$$\frac{\partial g(t,z)}{\partial t} - \theta_D \frac{\partial^2 g(t,z)}{\partial z^2} - \theta_S \frac{\partial g(t,z)}{\partial z} - \theta_A g(t,z) = 0$$

where $g(t,z)$ is the signal over time t and range z and θ_A, θ_D and θ_S are the parameters of interest. Formally speaking, the regression function is governed by the α^{th} order PDE

$$\boldsymbol{F}\left(\boldsymbol{t}, \left(D^{\boldsymbol{r}} \boldsymbol{f_\theta} : |\boldsymbol{r}| \leq \alpha\right), \boldsymbol{\theta}\right) = \boldsymbol{0},$$

$\boldsymbol{F}$ being known and $\boldsymbol{t} = (t_1, \ldots, t_s)^T$.

1.1 Literature review

There have been a significant number of works done on parameter estimation in the past few decades. We present them below.

1.1.1 Non-linear least squares approach

If the ODEs can be solved analytically, then the usual non linear least squares (NLS) [Levenberg (1944), Marquardt (1963)] can be used to estimate the unknown parameters. Thus

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \left\| \boldsymbol{Y}_i - \boldsymbol{f_\theta}(x_i) \right\|^2.$$

In many contexts, such closed form solutions are not available as evidenced in some of the previous examples. Hairer et al. (1987) and Mattheij and Molenaar (1996) used the 4-stage Runge-Kutta algorithm as an alternative approach. The statistical properties of the corresponding estimator have been studied by Xue et al. (2010). The strong consistency, $\sqrt{n}$-consistency and asymptotic normality of the estimator were established in their work.

1.1.2 Generalized profiling approach

Ramsay et al. (2007) proposed the generalized profiling procedure where the solution is approximated by a linear combination of basis functions given by $\boldsymbol{f}(\cdot) = \boldsymbol{\beta}^T\boldsymbol{\phi}$, where $\boldsymbol{\beta}$ is the vector of coefficients. The coefficients of the basis functions are estimated by solving a penalized optimization problem using an initial choice of the parameters of interest. A data-dependent fitting criterion $H_n(\boldsymbol{f})$ and a penalty term $J(\boldsymbol{f}, \boldsymbol{\theta})$ are constructed. For a given $\boldsymbol{\theta}$ they define

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\beta}} \left(H_n(\boldsymbol{\beta}^T\boldsymbol{\phi}) - \lambda J(\boldsymbol{\beta}^T\boldsymbol{\phi}, \boldsymbol{\theta})\right),$$

λ being a tuning parameter. Then the parameter estimate is obtained as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(H_n(\boldsymbol{\beta}^T(\boldsymbol{\theta})\boldsymbol{\phi})\right).$$

Qi and Zhao (2010) explored the statistical properties of this estimator including $\sqrt{n}$-consistency and asymptotic normality. Despite having desirable statistical properties, these approaches are computationally cumbersome especially for high-dimensional systems of ODEs as well as when $\boldsymbol{\theta}$ is high-dimensional.

1.1.3 Multiple shooting approach

Baake et al. (1992) viewed the ODE as a multi-point boundary value problem by dividing $[0, 1]$ into a grid of multiple shooting nodes $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_m < \tau_{m+1} = 1$ and considering the initial value problems

$$\frac{df_{\boldsymbol{\theta}}(t)}{dt} = F(t, f_{\boldsymbol{\theta}}, \boldsymbol{\theta}), \quad f_{\boldsymbol{\theta}}(\tau_j) = s_j \tag{1.4}$$

for $j = 1, \ldots, m+1$. Here $s_1, \ldots, s_{m+1}$ are also treated as unknown parameters and the new parameter vector $(s_1, \ldots, s_{m+1}, \boldsymbol{\theta})^T$ is estimated using least squares technique subject to the constraints (1.4). Naturally this method will give a solution with discontinuous trajectory which can be made continuous by introducing additional continuity criteria. This method also has high computational cost.

1.1.4 Two-step approach

Varah (1982) used an approach of two-step procedure. In the first step each of the state variables is approximated by a cubic spline using the least squares technique. Let us denote the approximation by $\hat{\boldsymbol{f}}(\cdot)$. In the second step, the parameter is estimated as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{f}}'(x_i) - \boldsymbol{F}(x_i, \hat{\boldsymbol{f}}(x_i), \boldsymbol{\theta}) \right\|^2 .$$

This method does not depend on the initial or boundary conditions of the state variables and is computationally very efficient. An example given in Voit and Almeida (2004) showed the computational superiority of the two-step approach over the usual least squares technique. Brunel (2008) replaced the sum of squares of the second step by a weighted integral of the squared deviation, that is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_0^1 \left\| \hat{\boldsymbol{f}}'(t) - \boldsymbol{F}(x_i, \hat{\boldsymbol{f}}(t), \boldsymbol{\theta}) \right\|^2 w(t)dt$$

and proved $\sqrt{n}$-consistency as well as asymptotic normality of $\hat{\boldsymbol{\theta}}$. The order of the B-spline basis is determined by the smoothness of $\boldsymbol{F}(\cdot,\cdot,\cdot)$ with respect to its first two arguments. Gugushvili and Klaassen (2012) used the same approach but used kernel smoothing instead of spline. They also established $\sqrt{n}$-consistency of the estimator. Another modification has been made in the work of Wu et al. (2012). They used penalized smoothing spline in the first step and numerical derivatives instead of actual derivatives of the nonparametrically estimated functions. In another work Brunel et al. (2014) used nonparametric approximation of the true solution to (1.2) and then used a set of orthogonality conditions to estimate the parameters. The $\sqrt{n}$-consistency as well as asymptotic normality of the estimator were also established in their work.

1.1.5 Bayesian estimation techniques

In ODE models Bayesian estimation was considered in the works of Gelman et al. (1996), Rogers et al. (2007) and Girolami (2008). First they solved the ODEs numerically to approximate the expected response and hence constructed the likelihood. A prior was assigned on $\boldsymbol{\theta}$ and MCMC technique was used to generate samples from the posterior. Computational cost might be an issue in this case as well.

Campbell and Steele (2012) proposed the smooth functional tempering approach which is a population MCMC technique and it utilizes the generalized profiling approach (Ramsay et al., 2007) and the parallel tempering algorithm. Campbell (2007) and Jaeger et al. (2012) also used a Bayesian analog of the generalized profiling by putting prior on the coefficients of the basis functions.

Chkrebtii et al. (2013) divided the time range into discrete grid points. They put a Gaussian process prior on the solution of the ODE and its derivative. The posterior distribution of the solution is used to draw the posterior sample of the parameter of interest. This method is computationally expensive and the likelihood is required to be known for this approach. The theoretical aspects of Bayesian estimation methods have not been yet explored in the literature.

1.1.6 Parameter estimation for higher order ODE model

Now we discuss some works done for higher order ODE models. Bergstrom (1983) considered parameter estimation in higher order stochastic differential equation models. The parameter is estimated by maximum likelihood estimation (MLE) using either the Gaussian distribution or a frequency domain approximation to Gaussian distribution as the working model. The estimator is shown to be asymptotically normally distributed as well as asymptotically efficient. In another paper Bergstrom (1985) described an efficient algorithm to compute the Gaussian likelihood while estimating the unknown parameters involved in a non-stationary higher order ODE. This algorithm avoids the computation of the covariance matrix of the observations using appropriate linear transformations.

The approach has been extended for the case of open higher order ODE's in Bergstrom (1986).

1.1.7 Parameter estimation for PDE model

Müller and Timmer (2002) used the multiple shooting approach for PDE models. A two step approach similar to Varah (1982) using splines was applied by Müller and Timmer (2004) to estimate the parameters. Xun et al. (2013) used a parameter cascading method which is a two-step optimization procedure. In the first step a nonparametric B-spline model is fit using penalized least squares approach. The optimal coefficients of the basis functions are expressed as a function of the parameter vector. The second step involves estimating the parameter by least squares method. They established asymptotic normality of the estimator with $n^{-1/2}$ rate of convergence. In the Bayesian framework, the Bayesian P-splines approach has been used by Xun et al. (2013).

1.2 Contributions of the thesis

In Chapter 2 we consider a Bayesian analog of the approach of Brunel (2008) fitting a nonparametric regression model using B-spline basis. We assign priors on the coefficients of the basis functions. A posterior is induced on $\boldsymbol{\theta}$ from the joint posterior of the coefficients of the basis functions. In this chapter we study the asymptotic properties of the posterior distribution of $\boldsymbol{\theta}$ and establish a Bernstein-von Mises theorem with the $n^{-1/2}$ contraction rate. Normal distribution is used as the working model for error distribution, but the true distribution of errors may be different. Interestingly, the original model is parametric but it is embedded in a nonparametric model, which is further approximated by high dimensional parametric models. Note that the slow rate of nonparametric estimation does not influence the convergence rate of the parameter in the original parametric model.

In Chapter 3 we propose two separate approaches based on the numerical solution obtained from four stage Runge-Kutta method (RK4). We use Gaussian distribution as the working model for error, although the true distribution may be different. The first approach is a fully Bayesian method assigning a direct prior on $\boldsymbol{\theta}$ and then constructing the posterior of $\boldsymbol{\theta}$ using an approximate likelihood function constructed using the numerical solution. We call this method Runge-Kutta sieve Bayesian (RKSB) method. In the second approach we define $\boldsymbol{\theta}$ as the minimizer of a weighted distance between the function in the nonparametric model and the RK4 numerical solution. We call this approach Runge-Kutta two-step Bayesian (RKTB) method. Thus, this approach is similar in spirit to the two-step Bayesian approach. A 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. Both RKSB and RKTB lead to Bernstein-von Mises Theorem with dispersion matrix inverse of Fisher information and hence both the Bayesian methods are asymptotically efficient. This is not the case for the two step-Bayesian approach described in Chapter 2. Bernstein-von Mises Theorem implies that credible intervals have asymptotically correct frequentist coverage. The computation cost of the two-step Bayesian method is the least. RKTB is more expensive and RKSB is even more expensive from computational point of view. RKTB is better than other highly computationally intensive Bayesian approaches based on MCMC and Gaussian process prior.

In Chapters 4 and 5 we extend these ideas for higher order ODE models and PDE models. Here the weight function used for two-step approach has to satisfy some additional criteria.

Chapter 6 contains discussions on possible future works using the numerical solution of PDE. We also briefly discuss the ODE model in the framework of non-additive error and scopes of research in this context.

Chapter 2

Bayesian two-step procedure based on splines

2.1 Introduction

The two step method introduced by Varah (1982) is computationally quite efficient as shown through an example given in Voit and Almeida (2004). Brunel (2008) used a weighted integral of the squared deviation instead of the sum of squares in the second step and proved the asymptotic normality of the estimator with $n^{-1/2}$ convergence rate. The degree of smoothness of $\boldsymbol{F}(\cdot,\cdot,\cdot)$ with respect to its first two arguments determines the order of the B-spline basis

In this chapter we consider a Bayesian analog of the approach of Brunel (2008). We put a prior on the coefficient vector of the spline based regression model. The Gaussian distribution is used as the working model for error although the true distribution may be different. We also let the ODE model to be misspecified, that is, the true regression function may not be a solution of the ODE. The response variable is also allowed to be multidimensional with possibly correlated errors. A posterior is induced on the parameter of interest from the posterior of the coefficient vector. A Bernstein-von Mises theorem is established for the posterior distribution of $\boldsymbol{\theta}$ with the $n^{-1/2}$ contraction rate.

The chapter is organized as follows. Section 2.2 contains the description of the model as well as the priors used for the analysis. The main results are given in Section 2.3.

We extend the results to more generalized setups in Section 2.4. In Section 2.5 we have carried out a simulation study under different settings. Section 2.6 contains the analysis of a real life data. The proofs of the main results are given in Section 2.7. The Section 2.8 contains the proofs of the required lemmas.

2.2 Model assumption and prior specification

We have a system of d ordinary differential equations given by

$$\frac{df_{j\boldsymbol{\theta}}(t)}{dt} = F_j(t, \boldsymbol{f_\theta}(t), \boldsymbol{\theta}),\ t \in [0,1],\ j = 1, \ldots, d, \tag{2.1}$$

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

Consider an $n \times d$ matrix of observations $\boldsymbol{Y}$ with $Y_{i,j}$ denoting the measurement taken on the j^{th} response at the point x_i, $0 \leq x_i \leq 1$, $i = 1, \ldots, n$; $j = 1, \ldots, d$. Denoting $\boldsymbol{\varepsilon} = (\!(\varepsilon_{i,j})\!)$ as the corresponding matrix of errors, the proposed model is given by

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

while the data is generated by the model

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

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

Since the expression of $\boldsymbol{f_\theta}$ is usually not available, the proposed model is embedded in the nonparametric regression model

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

where $\boldsymbol{X}_n = ((N_j(x_i)))_{1\le i\le n, 1\le j\le k_n+m-1}$, $\{N_j(\cdot)\}_{j=1}^{k_n+m-1}$ being the B-spline basis functions of order m with $k_n - 1$ interior knots. Here we denote

$$\boldsymbol{B}_n = \left(\boldsymbol{\beta}_1^{(k_n+m-1)\times 1}, \ldots, \boldsymbol{\beta}_d^{(k_n+m-1)\times 1}\right),$$

the matrix containing the coefficients of the basis functions. Also we consider P_0 to be unknown and use $N(0,\sigma^2)$ as the working distribution for the error where σ may be treated as another unknown parameter. Let us denote by $t_1, \ldots, t_{k_n-1}$ the set of interior knots with $t_l = l/k_n$ for $l = 1, \ldots, k_n - 1$. Hence the meshwidth is $\xi_n = 1/k_n$. Denoting by Q_n, the empirical distribution function of x_i, $i = 1, \ldots, n$, we assume

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

for some distribution function $Q(\cdot)$ with positive continuous density. Let the prior distribution on the coefficients be given by

$$\boldsymbol{\beta}_j \overset{iid}{\sim} N_{k_n+m-1}(\mathbf{0}, nc^{-1}k_n^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}) \tag{2.5}$$

for some constant $c > 0$. Simple calculation yields the posterior distribution for $\boldsymbol{\beta}_j$ as

$$\boldsymbol{\beta}_j|\boldsymbol{Y} \sim N_{k_n+m-1}\left(c_n^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j}, c_n^{-1}\sigma^2(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right) \tag{2.6}$$

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

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

$(0, 1)$. We define

$$R_{\boldsymbol{f}}(\boldsymbol{\eta}) = \left\{ \int_0^1 \left\| \boldsymbol{f}'(t) - \boldsymbol{F}(t, \boldsymbol{f}(t), \boldsymbol{\eta}) \right\|^2 w(t) dt \right\}^{1/2},$$
$$\boldsymbol{\psi}(\boldsymbol{f}) = \arg\min_{\boldsymbol{\eta} \in \boldsymbol{\Theta}} R_{\boldsymbol{f}}(\boldsymbol{\eta}). \tag{2.7}$$

It is easy to check that $\boldsymbol{\psi}(\boldsymbol{f}_{\boldsymbol{\eta}}) = \boldsymbol{\eta}$ for all $\boldsymbol{\eta} \in \Theta$. Thus the map $\boldsymbol{\psi}$ extends the definition of the parameter $\boldsymbol{\theta}$ beyond the model. Let us define $\boldsymbol{\theta}_0 = \boldsymbol{\psi}(\boldsymbol{f}_0)$. We assume that $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}$. From now on, we shall write $\boldsymbol{\theta}$ for $\boldsymbol{\psi}(\boldsymbol{f})$ and treat it as the parameter of interest. A posterior is induced on $\boldsymbol{\Theta}$ through the mapping $\boldsymbol{\psi}$ acting on $\boldsymbol{f}(\cdot) = \boldsymbol{B}_n^T \boldsymbol{N}(\cdot)$ and the posterior of $\boldsymbol{B}_n$ given by (2.6).

2.3 Main results

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

$$\text{for all } \epsilon > 0, \quad \inf_{\boldsymbol{\eta}: \|\boldsymbol{\eta} - \boldsymbol{\theta}_0\| \geq \epsilon} R_{\boldsymbol{f}_0}(\boldsymbol{\eta}) > R_{\boldsymbol{f}_0}(\boldsymbol{\theta}_0). \tag{2.8}$$

We denote $D_{l,r,s}\boldsymbol{F}(t, \boldsymbol{f}, \boldsymbol{\theta}) = \partial^{l+r+s}/\partial\boldsymbol{\theta}^s \partial\boldsymbol{f}^r \partial t^l \boldsymbol{F}(t, \boldsymbol{f}(t), \boldsymbol{\theta})$. Since the posterior distributions of $\boldsymbol{\beta}_j$ are mutually independent when $\boldsymbol{\varepsilon}_{,j}$ are mutually independent for $j = 1, \ldots, d$, we can assume $d = 1$ in Theorem 2.1 for the sake of simplicity in notation and write $f(\cdot)$, $f_0(\cdot)$, $F(\cdot, \cdot, \cdot)$, $\boldsymbol{\beta}$ instead of $\boldsymbol{f}(\cdot)$, $\boldsymbol{f}_0(\cdot)$, $\boldsymbol{F}(\cdot, \cdot, \cdot)$ and $\boldsymbol{B}_n$ respectively. Extension to d-dimensional case is straightforward as shown in Remark 2.5 after the statement of Theorem 2.1. We deal with the situation of correlated errors in Section 2.4.

Theorem 2.1. *Let the matrix*

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = \int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0) w(t) dt$$
$$- \int_0^1 \left(D_{0,0,1}\boldsymbol{S}(t, f_0(t), \boldsymbol{\theta}_0)\right) w(t) dt$$

be nonsingular, where

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

Let m be an integer greater than or equal to 5 and $n^{1/2m} \ll k_n \ll n^{1/8}$. If $D_{0,2,1}F(t,y,\boldsymbol{\theta})$ and $D_{0,0,2}F(t,y,\boldsymbol{\theta})$ are continuous in their arguments, then under the assumption (2.8), *there exists $E_n \subseteq C^m((0,1)) \times \boldsymbol{\Theta}$ with $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$, such that uniformly for $(f,\boldsymbol{\theta}) \in E_n$,*

$$\left\|\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - (\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\sqrt{n}(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0))\right\| \to 0 \tag{2.9}$$

as $n \to \infty$, where

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

Remark 2.2. Condition (2.8) implies that $\boldsymbol{\theta}_0$ is the unique point of minimum of $R_{\boldsymbol{f}_0}(\cdot)$ and $\boldsymbol{\theta}_0$ should be a well-separated point of minimum.

Remark 2.3. The posterior distribution of $\boldsymbol{\Gamma}(\boldsymbol{f})-\boldsymbol{\Gamma}(\boldsymbol{f}_0)$ contracts at $\mathbf{0}$ at the rate $n^{-1/2}$ as indicated by Lemma 2.14. Hence, the posterior distribution of $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ contracts at $\mathbf{0}$ at the rate $n^{-1/2}$ with high probability under the truth. We refer to Theorem 2.6 for a more refined version of this result.

Remark 2.4. We note that at least fifth order smoothness of the true mean function is sufficient to ensure the contraction rate $n^{-1/2}$. We do not gain anything more by assuming a higher order of smoothness. For $m = 5$, the required condition becomes $n^{1/10} \ll k_n \ll n^{1/8}$. Also, the knots are chosen deterministically and there is no need to assign a prior on the number of terms of the random series used. Hence, the issue of Bayesian adaptation, that is, improving convergence rate with higher smoothness without knowing the smoothness, does not arise in the present context.

Remark 2.5. When the response is a d-dimensional vector, (2.9) holds with the scalars being replaced by the corresponding d-dimensional vectors. Let $\boldsymbol{A}(t)$ stands for the $p \times d$

matrix

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

Then we have

$$
(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{\Gamma}(\boldsymbol{f}) = \sum_{j=1}^{d}\int_0^1 \boldsymbol{A}_{,j}(t)\boldsymbol{N}^T(t)\boldsymbol{\beta}_j dt = \sum_{j=1}^{d}\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j, \tag{2.10}
$$

where $\boldsymbol{G}_{n,j}^T = \int_0^1 \boldsymbol{A}_{,j}(t)\boldsymbol{N}^T(t)dt$ which is a $p\times(k_n+m-1)$ matrix for $j=1,\ldots,d$.

Thus in order to approximate the posterior distribution of $\boldsymbol{\theta}$, it suffices to study the asymptotic posterior distribution of the linear combination of $\boldsymbol{\beta}_j$ given by (2.10). The next theorem describes the approximate posterior distribution of $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$.

Theorem 2.6. *Define*

$$
\begin{aligned}
\boldsymbol{\mu}_n &= \sqrt{n}\sum_{j=1}^{d}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j} - \sqrt{n}\,(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{\Gamma}(\boldsymbol{f}_0),\\
\boldsymbol{\Sigma}_n &= n\sum_{j=1}^{d}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{G}_{n,j}
\end{aligned}
$$

and $\boldsymbol{B}_j = ((\langle A_{k,j}(\cdot), A_{k',j}(\cdot)\rangle))_{k,k'=1,\ldots,p}$ *for* $j=1,\ldots,d$*. If* $\boldsymbol{B}_j$ *is non-singular for all* $j=1,\ldots,d$*, then under the conditions of Theorem 2.1,*

$$
\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right) - N\left(\boldsymbol{\mu}_n,\sigma^2\boldsymbol{\Sigma}_n\right)\right\|_{TV} = o_{P_0}(1). \tag{2.11}
$$

Inspecting the proof, we can conclude that (2.11) is uniform over σ^2 belonging to a compact subset of $(0,\infty)$. Also note that the scale of the approximating normal distribution involves the working variance σ^2 assuming that it is given, even though the convergence is studied under the true distribution P_0 with variance σ_0^2, not necessarily equal to the given σ^2. Thus, the distribution matches with the frequentist distribution of the estimator in Brunel (2008) only if σ is correctly specified as σ_0. The next result assures that putting a prior on σ rectifies the problem.

Theorem 2.7. *We assign independent $N(\mathbf{0}, nc^{-1}k_n^{-1}\sigma^2(\boldsymbol{X}_n^T X_n)^{-1})$ prior on $\boldsymbol{\beta}_j$ for $j = 1,\ldots,d$ and a constant $c > 0$ and inverse gamma prior on σ^2 with shape and scale parameters a and b respectively. If the fourth order moment of the true error distribution is finite, then under the conditions of Theorems 2.1 and 2.6*

$$\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right)-N\left(\boldsymbol{\mu}_n,\sigma_0^2\boldsymbol{\Sigma}_n\right)\right\|_{TV}=o_{P_0}(1). \tag{2.12}$$

2.4 Extensions

The results obtained so far can be extended for the case where $\boldsymbol{\varepsilon}_{i,j}$ and $\boldsymbol{\varepsilon}_{i,j'}$ are associated for $i = 1,\ldots,n$ and $j \neq j'$; $j, j' = 1,\ldots,d$. Let under the working model, $\boldsymbol{\varepsilon}_{i,}$ have the dispersion matrix $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$ for $i = 1,\ldots,n$, $\boldsymbol{\Omega}$ being a known positive definite matrix. Denoting $\boldsymbol{\Omega}^{-1/2} = (\!(\omega^{jk})\!)_{j,k=1}^d$, we have the following extension of Theorem 2.6.

Theorem 2.8. *Define*

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

Then under the conditions of Theorems 2.1 and 2.6,

$$\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right)-N\left(\boldsymbol{\mu}_n^*,\sigma^2\boldsymbol{\Sigma}_n^*\right)\right\|_{TV}=o_{P_0}(1). \tag{2.13}$$

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

Theorems 2.1 and 2.7,

$$\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right)-N\left(\boldsymbol{\mu}_n^*,\sigma_0^2\boldsymbol{\Sigma}_n^*\right)\right\|_{TV}=o_{P_0}(1), \tag{2.14}$$

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

Remark 2.9. In many applications, the regression function is modeled as $\boldsymbol{h}_{\boldsymbol{\theta}}(t)=\boldsymbol{g}(\boldsymbol{f}_{\boldsymbol{\theta}}(t))$ instead of $\boldsymbol{f}_{\boldsymbol{\theta}}(t)$, where $\boldsymbol{g}$ is a known invertible function and $\boldsymbol{h}_{\boldsymbol{\theta}}(t)\in\mathbb{R}^d$. It should be noted that

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

which is a known function of $t,\boldsymbol{h}_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Now we can carry out our analysis replacing $\boldsymbol{F}$ and $\boldsymbol{f}_{\boldsymbol{\theta}}$ in (1.2) by $\boldsymbol{H}$ and $\boldsymbol{h}_{\boldsymbol{\theta}}$ respectively.

Remark 2.10. Often we do not have the data on all the state variables. For the sake of simplicity let $d=2$. Let the true regression function be $(f_{1\boldsymbol{\theta}_0}(\cdot),f_{2\boldsymbol{\theta}_0}(\cdot))^T$ and suppose that only the first component Y_1 of the response variable $\boldsymbol{Y}$ is observable. Let the system of ODE be given by

$$\frac{d}{dt}f_{1\boldsymbol{\theta}}(t) = F_1(t,f_{1\boldsymbol{\theta}}(t),f_{2\boldsymbol{\theta}}(t),\boldsymbol{\theta}) \tag{2.15}$$

$$\frac{d}{dt}f_{2\boldsymbol{\theta}}(t) = F_2(t,f_{1\boldsymbol{\theta}}(t),f_{2\boldsymbol{\theta}}(t),\boldsymbol{\theta}). \tag{2.16}$$

Model $f_1(\cdot)$ by a spline series $\boldsymbol{\beta}^T\boldsymbol{N}(\cdot)$, where $\boldsymbol{\beta}$ is a free parameter. We substitute this expression for $f_1(\cdot)$ and apply the four stage Runge-Kutta method on (2.16) with n grid points to obtain the corresponding nonparametric regression model for Y_2 on t given by $f_2(t)=\phi_n(t,f_1(t),\boldsymbol{\theta})$. Now we define

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}}\int_0^1\left(f_1'(t)-F_1(t,f_1(t),f_2(t),\boldsymbol{\eta})\right)^2w(t)dt. \tag{2.17}$$

Then the posterior distribution of $\boldsymbol{\theta}$ induced from that of $f_1(\cdot)$ will satisfy the Bernstein-von Mises theorem with $n^{-1/2}$ rate of contraction towards $\boldsymbol{\theta}_0$. The proof of this assertion is given later.

2.5 Simulation Study

We consider the Lotka-Volterra equations to study the posterior distribution of $\boldsymbol{\theta}$. For a sample of size n, the x_i's are chosen as $x_i = (2i-1)/2n$ for $i = 1, \ldots, n$. Samples of sizes 100 and 500 are considered. The weight function is chosen as $w(t) = t(1-t)$, $t \in [0, 1]$. We simulate 1000 replications for each case. Under each replication a sample of size 1000 is directly drawn from the posterior distribution of $\boldsymbol{\theta}$ and then 95% equal tailed credible interval is obtained. Each replication took around one minute. We calculate the coverage and the average length of the corresponding credible interval over these 1000 replications. The estimated standard errors of the interval length and coverage are given inside the parentheses in the tables. We also consider 1000 replications to construct the 95% equal tailed confidence interval based on asymptotic normality as obtained from the estimation method introduced by Varah (1982) and modified and studied by Brunel (2008). We abbreviate this method by "VB" in tables. The estimated standard errors of the interval length and coverage are given inside the parentheses in the tables. Thus we have $p = 4, d = 2$ and the ODE's are given by

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

for $t \in [0, 1]$ with initial condition $f_{1\boldsymbol{\theta}}(0) = 1, f_{2\boldsymbol{\theta}}(0) = 0.5$. The above system is not analytically solvable. The true mean vector is given by $(f_{1\boldsymbol{\theta}_0}(t), f_{2\boldsymbol{\theta}_0}(t))^T$, where $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})^T$. We take $\theta_{10} = \theta_{20} = \theta_{30} = \theta_{40} = 10$ to be the true value of the parameter.

The true distribution of error is taken either $N(0, (0.2)^2)$ or a scaled t-distribution with 6 degrees of freedom, where scaling is done in order to make the standard deviation 0.2. We put an inverse gamma prior on σ^2 with shape and scale parameters being 99 and 1 respectively and independent Gaussian priors on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with mean vector $\mathbf{0}$

and dispersion matrix $nc^{-1}k_n^{-1}\sigma^2(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}$ with $c = 3.5$. We choose $m = 5$. As far as choosing k_n is concerned, we take $k_n = 17, 20$ for $n = 100$ and 500 respectively. The simulation results are summarized in the Table 2.1. It is clear that both methods give more accurate results as we go on increasing sample size.

Table 2.1: *Coverages and average lengths of the Bayesian credible interval and confidence interval obtained from VB method*

n		$N(0,(0.2)^2)$				scaled t_6			
		Bayes		VB		Bayes		VB	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
100	θ_1	92.8	4.24	88.6	3.38	91.8	4.22	87.6	3.34
		(0.02)	(0.15)	(0.03)	(0.46)	(0.03)	(0.15)	(0.03)	(0.47)
	θ_2	97.4	4.22	89.0	3.15	97.8	4.19	88.1	3.13
		(0.02)	(0.15)	(0.03)	(0.44)	(0.01)	(0.15)	(0.03)	(0.45)
	θ_3	93.4	4.49	88.9	3.60	93.4	4.53	89.7	3.61
		(0.02)	(0.17)	(0.03)	(0.51)	(0.02)	(0.21)	(0.03)	(0.56)
	θ_4	97.5	4.27	87.4	3.19	98.2	4.29	89.4	3.21
		(0.02)	(0.14)	(0.03)	(0.46)	(0.01)	(0.18)	(0.03)	(0.50)
500	θ_1	95.5	1.71	94.6	1.55	95.2	1.72	93.8	1.55
		(0.01)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)	(0.01)	(0.10)
	θ_2	97.0	1.63	93.8	1.45	96.5	1.64	93.2	1.45
		(0.01)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)	(0.01)	(0.10)
	θ_3	95.3	1.84	93.8	1.66	94.7	1.83	93.8	1.66
		(0.01)	(0.01)	(0.01)	(0.10)	(0.01)	(0.01)	(0.01)	(0.12)
	θ_4	97.4	1.66	93.5	1.48	97.8	1.66	94.3	1.48
		(0.01)	(0.01)	(0.01)	(0.09)	(0.01)	(0.01)	(0.01)	(0.10)

2.6 Real life data

Diggle (1990) considered the growth of three closed colonies of paramecium aurelium in a nutritive medium over a period of 19 days with growth on day zero being 2. The

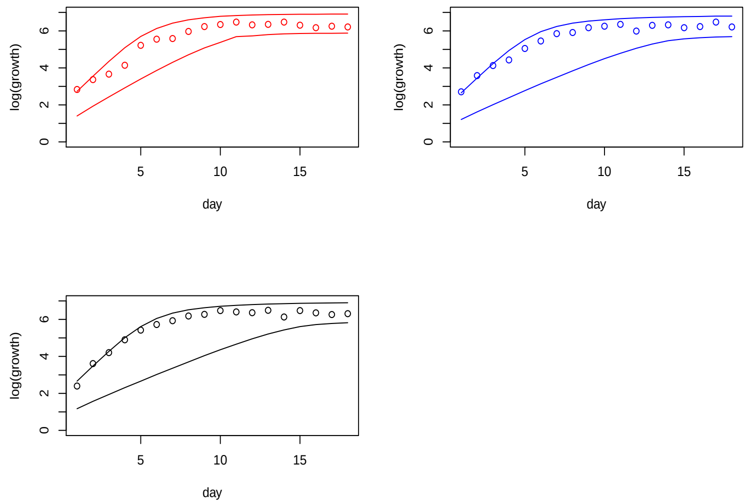

Figure 2.1: *Observed values and the posterior predictive intervals for the three colonies*

logarithm of growth, denoted by $\mu(t)$, changes over time according to the ODE

$$\frac{d\mu(t)}{dt} = \theta_1 - \theta_2 \exp\{\mu(t)\},\ \ \mu(0) = \log(2),$$

as suggested in Ghosh and Goyal (2010). The three colonies are analyzed separately. In this example there are $n = 18$ observations for each colony. We put an inverse gamma prior on σ^2 with shape and scale parameters 10 and 1 respectively and use $w(t) = t^7(1-t)^7$. Conditional on σ^2 we put Gaussian prior on $\boldsymbol{\beta}$ with mean vector $\mathbf{0}$ and dispersion matrix $nc^{-1}k_n^{-1}\sigma^2(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}$ with $c = 1000$. We take $m = 5$ and $k_n = 2$. The ODE has the solution

$$\mu(t) = \log\theta_1 + \log 2 + t\theta_1 - \log[2\theta_2(\exp\{t\theta_1\} - 1) + \theta_1].$$

Samples of size 1000 are drawn from the posterior distributions of θ_1 and θ_2. The 95% posterior predictive interval of the functional value at the 18 data points is superimposed on the data in Figure 2.1.

2.7 Proofs of theorems

Proof of Theorem 2.1. The structure of the proof follows that of Proposition 3.1 of Brunel (2008) and Proposition 3.3 of Gugushvili and Klaassen (2012), but differs substantially in detail since we address posterior variation and also allow misspecification. First note that

$$\begin{aligned}
&\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0) \\
&= \int_0^1 \big(-(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T D_{0,1,0}F(t, f_0(t), \boldsymbol{\theta}_0)w(t) \qquad (2.18)\\
&\quad -\frac{d}{dt}[(D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T w(t)] + (D_{0,1,0}\boldsymbol{S}(t, f_0(t), \boldsymbol{\theta}_0))\, w(t)\Big) (f(t) - f_0(t))dt.
\end{aligned}$$

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

$$\int_0^1 (D_{0,0,1}F(t, f(t), \boldsymbol{\theta}))^T (f'(t) - F(t, f(t), \boldsymbol{\theta}))w(t)dt = \mathbf{0}, \qquad (2.19)$$

$$\int_0^1 (D_{0,0,1}F(t, f_0(t), \boldsymbol{\theta}_0))^T (f_0'(t) - F(t, f_0(t), \boldsymbol{\theta}_0))w(t)dt = \mathbf{0}. \qquad (2.20)$$

Taking difference, we get

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

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

$$\begin{aligned}
&\boldsymbol{M}(f,\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&= \int_0^1 \left(D_{0,0,1}F(t,f(t),\boldsymbol{\theta}_0) - D_{0,0,1}F(t,f_0(t),\boldsymbol{\theta}_0)\right)^T \left(f_0'(t) - F(t,f_0(t),\boldsymbol{\theta}_0)\right)w(t)dt\\
&\quad + \int_0^1 \left(D_{0,0,1}F(t,f(t),\boldsymbol{\theta}_0)\right)^T \left(f'(t) - f_0'(t) + F(t,f_0(t),\boldsymbol{\theta}_0) - F(t,f(t),\boldsymbol{\theta}_0)\right)w(t)dt,
\end{aligned}$$

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

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

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

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

where $\epsilon_n \to 0$. By Lemmas 2.12 and 2.13, there exists a sequence $\{\epsilon_n\}$ so that $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$. Then, $\boldsymbol{M}(f,\boldsymbol{\theta})$ is invertible and the eigenvalues of $[\boldsymbol{M}(f,\boldsymbol{\theta})]^{-1}$ are bounded away from zero and infinity for sufficiently large n and

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

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

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

for sufficiently large n, where

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

In view of Lemmas 2.12 and 2.14, on a set in the sample space with high true probability, the posterior distribution of $(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\sqrt{n}(\boldsymbol{T}_{1n}+\boldsymbol{T}_{2n}+\boldsymbol{T}_{3n})$ assigns most of its mass inside a large compact set. Thus, we can assert that inside the set E_n, the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ is given by that of

$$(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\sqrt{n}(\boldsymbol{T}_{1n}+\boldsymbol{T}_{2n}+\boldsymbol{T}_{3n}). \tag{2.21}$$

We shall extract $\sqrt{n}\,(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0))$ from (2.21) and show that the remainder term goes to zero. First write

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

which follows by integration by parts and the fact that $w(0)=w(1)=0$. Note that the first integral of the above equation appears in (2.18). The norm of the second integral can be bounded above by a constant multiple of $\sup_{t\in[0,1]}|f(t)-f_0(t)|^2+\sup_{t\in[0,1]}|f'(t)-f_0'(t)|^2$ using the continuity of $D_{0,1,1}F(t,y,\boldsymbol{\theta})$. Now we consider $\boldsymbol{T}_{3n}$ in (2.21). Then,

$$\begin{aligned}\boldsymbol{T}_{3n} &= \int_0^1 \left(D_{0,0,1}F(t,f_0(t),\boldsymbol{\theta}_0)\right)^T \left(F(t,f_0(t),\boldsymbol{\theta}_0) - F(t,f(t),\boldsymbol{\theta}_0)\right)w(t)dt\\ &\quad + \int_0^1 \left(D_{0,0,1}F(t,f(t),\boldsymbol{\theta}_0) - D_{0,0,1}F(t,f_0(t),\boldsymbol{\theta}_0)\right)^T\\ &\qquad \times\left(F(t,f_0(t),\boldsymbol{\theta}_0) - F(t,f(t),\boldsymbol{\theta}_0)\right)w(t)dt.\end{aligned} \tag{2.22}$$

The first integral on the right hand side of (2.22) can be written as

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

say. Now $\boldsymbol{T}_{31n}$ appears in (2.18). By the continuity of $D_{0,2,0}F(t, y, \boldsymbol{\theta})$, $\|\boldsymbol{T}_{32n}\|$ can be bounded above up to a constant by a multiple of $\sup_{t\in[0,1]} |f(t) - f_0(t)|^2$. We apply the Cauchy-Schwarz inequality and the continuity of $D_{0,1,1}F(t, y, \boldsymbol{\theta})$ to bound the second integral on the right hand side of (2.22) by a constant multiple of $\sup\{|f(t) - f_0(t)|^2 : t \in [0, 1]\}$. As far as the first term inside the bracket of (2.21) is concerned, we have

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

The first integral appears in (2.18). The norm of the second integral of the above display can be bounded by a constant multiple of $\sup\{|f(t) - f_0(t)|^2 : t \in [0, 1]\}$ utilizing the continuity of $D_{0,2,1}F(t, y, \boldsymbol{\theta})$ with respect to its arguments. Combining these, we find that the norm of the vector $(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1} \sqrt{n}(\boldsymbol{T}_{1n} + \boldsymbol{T}_{2n} + \boldsymbol{T}_{3n}) - (\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1} \sqrt{n} (\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0))$ is bounded above by a constant multiple of

$$
\sqrt{n} \sup_{t\in[0,1]} |f(t) - f_0(t)|^2 + \sqrt{n} \sup_{t\in[0,1]} |f'(t) - f_0'(t)|^2.
$$

Now applying Lemma 2.12, we get the desired result. $\square$

Proof of Theorem 2.6. By Theorem 2.1 and (2.10), it suffices to show that

$$\left\| \Pi\left(\sqrt{n}\sum_{j=1}^{d}\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j - \sqrt{n}\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{\Gamma}(\boldsymbol{f}_0) \in \cdot|\boldsymbol{Y}\right) - N(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Sigma}_n)\right\|_{TV} = o_{P_0}(1). \tag{2.23}$$

Note that the posterior distribution of $\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j$ is a normal distribution with mean vector given by $(1+\sigma^2 ck_n/n)^{-1}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j}$ and dispersion matrix

$$\sigma^2(1+\sigma^2 ck_n/n)^{-1}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{G}_{n,j}.$$

We calculate the Kullback-Leibler divergence between two Gaussian distributions to prove the assertion. Alternatively, we can also follow the approach given in Theorem 1 and Corollary 1 of Bontemps (2011). The Kulback-Leibler divergence between the distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_2)$ is given by

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

Taking

$$\boldsymbol{\mu}_1 = (1+\sigma^2 ck_n/n)^{-1}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j},\ \boldsymbol{\Omega}_1 = \sigma^2(1+\sigma^2 ck_n/n)^{-1}\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{G}_{n,j}$$

and $\boldsymbol{\mu}_2 = \boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j}$, $\boldsymbol{\Omega}_2 = \sigma^2\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{G}_{n,j}$, we get

$$\mathrm{tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) = p + o(1).$$

Also,

$$\log(\det(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2)) = p\log(1+\sigma^2 ck_n/n) \asymp k_n/n = o(1).$$

From the proof of Lemma 2.14, it follows that

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

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

Proof of Theorem 2.7. The marginal posterior of σ^2 is also inverse gamma with parameters $(dn+2a)/2$ and $b + \sum_{j=1}^d \boldsymbol{Y}_{,j}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n}(1+(ck_n/n))^{-1})\boldsymbol{Y}_{,j}/2$, where $\boldsymbol{P}_{\boldsymbol{X}_n} = \boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T$. Straightforward calculations show that

$$\begin{aligned}
\mathrm{E}(\sigma^2|\boldsymbol{Y}) &= \frac{\frac{1}{2}\sum_{j=1}^d\left\{\boldsymbol{Y}_{,j}^T\boldsymbol{Y}_{,j} - \boldsymbol{Y}_{,j}^T\boldsymbol{P}_{\boldsymbol{X}_n}\boldsymbol{Y}_{,j}(1+ck_nn^{-1})^{-1}\right\}+b}{\frac{1}{2}dn+a-1},\\
\mathrm{Var}(\sigma^2|\boldsymbol{Y}) &= \frac{\left(\mathrm{E}(\sigma^2|\boldsymbol{Y})\right)^2}{\frac{1}{2}dn+a-2},
\end{aligned}$$

which give rise to $|\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2| = O_{P_0}(n^{-1/2})$ and $\mathrm{Var}(\sigma^2|\boldsymbol{Y}) = O_{P_0}(n^{-1})$. In particular, the marginal posterior distribution of σ^2 is consistent at the true value of error variance. Let $\mathscr{N}$ be an arbitrary neighborhood of σ_0. Then, $\Pi(\mathscr{N}^c|\boldsymbol{Y}) = o_{P_0}(1)$. We observe that

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

The total variation distance between the two normal distributions appearing in the second term is bounded by a constant multiple of $|\sigma - \sigma_0|$, and hence the term can be made arbitrarily small by choosing $\mathscr{N}$ appropriately. The first term converges in probability to zero by Theorem 2.6. The third term converges in probability to zero by the posterior consistency. Hence, we get the desired result. $\square$

Proof of Theorem 2.8. According to the fitted model, $\boldsymbol{Y}_{i,}^{1\times d} \sim N_d((\boldsymbol{X}_n)_{i,}\boldsymbol{B}_n, \boldsymbol{\Sigma}^{d\times d})$ for $i = 1, \ldots, n$. The logarithm of the posterior probability density function (p.d.f.) is negative half times

$$\sum_{i=1}^{n} ((\boldsymbol{X}_n)_{i,}\boldsymbol{B}_n - \boldsymbol{Y}_{i,})\,\boldsymbol{\Sigma}^{-1}\left(B_n^T(\boldsymbol{X}_n^T)_{,i} - \boldsymbol{Y}_{i,}^T\right) + \sum_{j=1}^{d} \boldsymbol{\beta}_j^T \frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{nc^{-1}k_n^{-1}}\boldsymbol{\beta}_j, \tag{2.24}$$

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

$$\mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n\boldsymbol{I}_d}{n}\right)\boldsymbol{B}_n^T\boldsymbol{X}_n^T\boldsymbol{X}_n\boldsymbol{B}_n\right). \tag{2.25}$$

The term in (2.24) which is linear in $\boldsymbol{\beta}_j$, $j = 1, \ldots, d$, is given by

$$\sum_{i=1}^{n}(\boldsymbol{X}_n)_{i,}(\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_d)\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{i,}^T = \mathrm{tr}\left(\boldsymbol{X}_n\boldsymbol{B}_n\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}^T\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}^T\boldsymbol{X}_n\boldsymbol{B}_n\right).$$

A completing square argument shows that the posterior density is proportional to

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

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

$$\begin{aligned}\mathrm{vec}(\boldsymbol{B}_n)|\boldsymbol{Y} \quad \sim \quad & N\left(\mathrm{vec}\left((\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n\boldsymbol{I}_d}{n}\right)^{-1}\right),\right.\\ & \left.\left(\boldsymbol{\Sigma}^{-1} + \frac{ck_n\boldsymbol{I}_d}{n}\right)^{-1} \otimes (\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right).\end{aligned}$$

Fixing a $j \in \{1, \ldots, d\}$, we observe that the posterior mean of $\boldsymbol{\beta}_j$ is a weighted sum of $(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j'}$ for $j' = 1, \ldots, d$. The weight attached with $(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j}$ is of the order of 1, whereas for $j' \neq j$, the contribution from $(\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^T \boldsymbol{Y}_{,j'}$ is of the order of k_n/n which goes to zero as n goes to infinity. Thus, the results of Lemmas 2.11 to 2.14 can be shown to hold under this setup. We are interested in the limiting distribution of $(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1} \boldsymbol{\Gamma}(\boldsymbol{f}) = \sum_{j=1}^{d} \boldsymbol{G}_{n,j}^T \boldsymbol{\beta}_j = (\boldsymbol{G}_{n,1}^T \ldots \boldsymbol{G}_{n,d}^T)\text{vec}(\boldsymbol{B}_n)$. We note that the posterior distribution of $\left((\boldsymbol{\Sigma}^{-1} + ck_n \boldsymbol{I}_d/n)^{1/2} \otimes \boldsymbol{I}_{k_n+m-1}\right) \text{vec}(\boldsymbol{B}_n)$ is a $(k_n+m-1)d$-dimensional normal distribution with mean vector and dispersion matrix being given by

$$\text{vec}\left((\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^T \boldsymbol{Y} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Sigma}^{-1} + ck_n \boldsymbol{I}_d/n\right)^{-1/2}\right)$$

and $\boldsymbol{I}_d \otimes (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1}$ respectively, since by the properties of Kronecker product, for the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$ of appropriate orders $(\boldsymbol{B}^T \otimes \boldsymbol{A})\text{vec}(\boldsymbol{D}) = \text{vec}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{B})$.
Let us consider the mean vector of the posterior distribution of the vector

$$\left(\left(\boldsymbol{\Sigma}^{-1} + ck_n \boldsymbol{I}_d/n\right)^{1/2} \otimes \boldsymbol{I}_{k_n+m-1}\right) \text{vec}(\boldsymbol{B}_n).$$

We observe that

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

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

$$\boldsymbol{Z}_k | \boldsymbol{Y} \sim N_{k_n+m-1}\left(\left(\boldsymbol{X}_n^T \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_n^T \sum_{j=1}^{d} \boldsymbol{Y}_{,j} c_{jk}, \left(\boldsymbol{X}_n^T \boldsymbol{X}_n\right)^{-1}\right).$$

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

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

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

$$\left\|(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_{,j}(c_{jk}-\sigma^{jk})\right\| \lesssim \frac{k_n}{n}O_{P_0}(\sqrt{n}) = o_{P_0}(1). \tag{2.26}$$

Hence, we conclude that the total variation distance between the distributions

$$N\left((\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\sum_{j=1}^d \boldsymbol{Y}_{,j}c_{jk}, (\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right)$$

and $N\left((\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\sum_{j=1}^d \boldsymbol{Y}_{,j}\sigma^{jk}, (\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right)$ converges to zero in P_0-probability. We can write $\left(\boldsymbol{G}_{n,1}^T\ldots\boldsymbol{G}_{n,d}^T\right)\mathrm{vec}(\boldsymbol{B}_n)$ in terms of $\boldsymbol{Z}_k$ as

$$\sum_{k=1}^d \mathrm{cols}_{(k-1)(k_n+m-1)+1}^{k(k_n+m-1)}\left(\left(\boldsymbol{G}_{n,1}^T\ldots\boldsymbol{G}_{n,d}^T\right)\left(\left(\boldsymbol{\Sigma}^{-1}+\frac{ck_n\boldsymbol{I}_d}{n}\right)^{1/2}\otimes\boldsymbol{I}_{k_n+m-1}\right)^{-1}\right)\boldsymbol{Z}_k.$$

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

$$\left\|\left(\sqrt{n}\left(\boldsymbol{G}_{n,1}^T\ldots\boldsymbol{G}_{n,d}^T\right)\mathrm{vec}(\boldsymbol{B}_n)-\sqrt{n}\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}(\boldsymbol{f}_0)\right)-N(\boldsymbol{\mu}_n^{**},\boldsymbol{\Sigma}_n^{**})\right\|_{TV} = o_{P_0}(1),$$

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

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

and $\boldsymbol{\Sigma}_n^{**}$ is given by

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

Following the steps of the proof of Lemma 2.14, it can be shown that the eigenvalues of the matrix $\boldsymbol{\Sigma}_n^*$ mentioned in the statement of Theorem 2.8 are bounded away from zero and infinity. We can show that the Kullback-Leibler divergence of $N(\boldsymbol{\mu}_n^{**},\boldsymbol{\Sigma}_n^{**})$ from $N(\boldsymbol{\mu}_n^*,\sigma^2\boldsymbol{\Sigma}_n^*)$ converges in probability to zero by going through some routine matrix manipulations. Hence,

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

The above expression is equivalent to (2.23) of the proof of Theorem 2.6. Following steps similar to those of Theorem 2.6, we get (2.13). We obtain (2.14) by following the proof of Theorem 2.7. □

Proof of Remark 2.10. Using the definition (2.17) we get

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

Subtracting the latter from the former we get

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

Since $f_{2\boldsymbol{\theta}_0}(t)$ is a known function of t, it can be absorbed in the first argument of F_1 which then becomes a function of three arguments. Then the second part of the left side above can be analyzed as in Theorem 2.1. To deal with the first part of left side it is sufficient to study the difference $f_2(\cdot)-f_{2\boldsymbol{\theta}_0}(\cdot)$. Note that $f_2(t)$ can be written as

$$\begin{aligned}&\phi_n(t,f_{1\boldsymbol{\theta}_0}(t),\boldsymbol{\theta}_0)+(f_1(t)-f_{1\boldsymbol{\theta}_0}(t))D_{0,1,0}\phi_n(t,f_{1\boldsymbol{\theta}_0}(t),\boldsymbol{\theta}_0)\\&+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^TD_{0,0,1}\phi_n(t,f_{1\boldsymbol{\theta}_0}(t),\boldsymbol{\theta}_0)+O\left((f_1(t)-f_{1\boldsymbol{\theta}_0}(t))^2\right)+O\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\right).\end{aligned}$$

Also, the difference $\sup_{t\in[0,1]}|\phi_n(t,f_{1\boldsymbol{\theta}_0}(t),\boldsymbol{\theta}_0)-f_{2\boldsymbol{\theta}_0}(t)|$ is of the order n^{-1}. Now using Lemmas 2.11 to 2.13, we can conclude that

$$\|\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\boldsymbol{J}_{\boldsymbol{\theta}_0}^{-1}\sqrt{n}(\boldsymbol{\Gamma}(f_1)-\boldsymbol{\Gamma}(f_{1\boldsymbol{\theta}_0}))\|\to 0$$

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

2.8 Proofs of the lemmas

We need to go through a couple of lemmas in order to prove the main results. We denote by $\mathrm{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ the expectation and variance operators respectively with respect to P_0-probability. The following lemma helps to estimate the bias of the Bayes estimator.

Lemma 2.11. *For $m \geq 2$ and k_n satisfying $n^{1/2m} \ll k_n \ll n$, for $r = 0, 1$,*

$$\sup_{t\in[0,1]} |\mathrm{E}_0(\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y})) - f_0^{(r)}(t)| = o(k_n^{r+1/2}/\sqrt{n}).$$

Proof. We note that $f^{(r)}(t) = (\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}$ for $r = 0, 1$ with $\boldsymbol{N}^{(r)}(\cdot)$ standing for the r^{th} order derivative of $\boldsymbol{N}(\cdot)$. By (2.6),

$$\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y}) = \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} (\boldsymbol{N}^{(r)}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}. \tag{2.27}$$

Theorem A.1 gives

$$(\boldsymbol{N}^{(r)}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{N}^{(r)}(t) \asymp \frac{k_n^{2r+1}}{n}. \tag{2.28}$$

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

$$\sup_{t\in[0,1]} |f_0^{(r)}(t) - (\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}^*| = O(k_n^{-(m-r)}). \tag{2.29}$$

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

$$\begin{aligned}
&\sqrt{n}k_n^{-r-1/2} \sup_{t\in[0,1]} |\mathrm{E}_0(\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y})) - f_0^{(r)}(t)| \\
&\leq \sqrt{n}k_n^{-r-1/2} \sup_{t\in[0,1]} \left| \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} (\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}^* - (\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}^* \right| \\
&\quad + \sqrt{n}k_n^{-r-1/2} \left(1 + \frac{ck_n\sigma^2}{n}\right)^{-1} \sup_{t\in[0,1]} |(\boldsymbol{N}^{(r)}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T(f_0(\boldsymbol{x}) - \boldsymbol{X}_n\boldsymbol{\beta}^*)| \\
&\quad + \sqrt{n}k_n^{-r-1/2} \sup_{t\in[0,1]} |f_0^{(r)}(t) - (\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}^*|.
\end{aligned}$$

Using the fact that $\sup_{t\in[0,1]} |(\boldsymbol{N}^{(r)}(t))^T\boldsymbol{\beta}^*| = O(1)$, first term on the right hand side of the previous inequality is of the order of $k_n^{-r+1/2}/\sqrt{n}$. Using the Cauchy-Schwarz inequality, (2.28) and (2.29), we can bound the second term up to a constant multiple by

$\sqrt{n}k_n^{-m}$. The third term has the order of $\sqrt{n}k_n^{-m-1/2}$ as a result of (2.29). By the assumed conditions on m and k_n, the assertion holds. $\square$

The following lemma controls posterior variability.

Lemma 2.12. *If $m \geq 5$ and $n^{1/2m} \ll k_n \ll n^{1/8}$, then for $r = 0, 1$ and for all $\epsilon > 0$,*

$$\Pi\left(\sqrt{n}\sup_{t\in[0,1]}|f^{(r)}(t) - f_0^{(r)}(t)|^2 > \epsilon|\boldsymbol{Y}\right) = o_{P_0}(1).$$

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

$$2\frac{\sqrt{n}}{\epsilon}\left\{\sup_{t\in[0,1]}\left|\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y}) - f_0^{(r)}(t)\right|^2 + \mathrm{E}\left[\sup_{t\in[0,1]}\left|f^{(r)}(t) - \mathrm{E}(f^{(r)}(t)|\boldsymbol{Y})\right|^2|\boldsymbol{Y}\right]\right\}. \quad (2.30)$$

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

$$2\sup_{t\in[0,1]}\left|\mathrm{E}_0(\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y})) - f_0^{(r)}(t)\right|^2 \\ +2\mathrm{E}_0\left[\sup_{t\in[0,1]}\left|\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y}) - \mathrm{E}_0(\mathrm{E}[f^{(r)}(t)|\boldsymbol{Y}])\right|^2\right]. \quad (2.31)$$

Using (2.27), $\sup_{t\in[0,1]}\left|\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y}) - \mathrm{E}_0(\mathrm{E}[f^{(r)}(t)|\boldsymbol{Y}])\right|$ can be bounded up to a constant multiple by

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

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 inside the E_0- expectation in the second term of (2.31) by a constant multiple of

$$\max_{1\le k\le n}\left|(\boldsymbol{N}^{(r)}(s_k))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{\varepsilon}\right|^2+\sup_{t\in[0,1]}\frac{1}{n^2}\left|(\boldsymbol{N}^{(r+1)}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{\varepsilon}\right|^2. \quad (2.32)$$

By the spectral decomposition, we can write $\boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T=\boldsymbol{P}^T\boldsymbol{D}\boldsymbol{P}$, where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{D}$ 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

$$\begin{aligned}&\mathrm{E}_0\left(\max_{1\le k\le n}\left|(\boldsymbol{N}^{(r)}(s_k))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{\varepsilon}\right|^2\right)\\ &\le \max_{1\le k\le n}\left\{(\boldsymbol{N}^{(r)}(s_k))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{N}^{(r)}(s_k)\right\}\mathrm{E}_0\left(\boldsymbol{\varepsilon}^T\boldsymbol{P}^T\boldsymbol{D}\boldsymbol{P}\boldsymbol{\varepsilon}\right).\end{aligned}$$

By Theorem A.1 and the fact that $\mathbf{Var}_0(\boldsymbol{P}\boldsymbol{\varepsilon})=\mathbf{Var}_0(\boldsymbol{\varepsilon})$, we can conclude that the expectation of the first term of (2.32) is $O(k_n^{2r+2}/n)$. Again applying the Cauchy-Schwarz inequality, the second term of (2.32) is bounded by

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

whose expectation is of the order $n(k_n^{2r+3}/n^3)=k_n^{2r+3}/n^2$, using Theorem A.1. Thus, the expectation of the bound given by (2.32) is of the order k_n^{2r+2}/n. Combining it with (2.31) and Lemma 2.11, we get

$$\mathrm{E}_0\left[\sup_{t\in[0,1]}\left|\mathrm{E}(f^{(r)}(t)|\boldsymbol{Y})-f_0^{(r)}(t)\right|^2\right]=O\left(\frac{k_n^{2r+2}}{n}\right). \quad (2.33)$$

Define

$$\boldsymbol{\varepsilon}^*:=(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{1/2}\boldsymbol{\beta}-\left(1+\frac{\sigma^2ck_n}{n}\right)^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1/2}\boldsymbol{X}_n^T\boldsymbol{Y}.$$

Note that

$$\boldsymbol{\varepsilon}^*|\boldsymbol{Y}\sim N(\boldsymbol{0},\left(\sigma^{-2}+ck_n/n\right)^{-1}\boldsymbol{I}_{k_n+m-1}).$$

Expressing $\sup_{t\in[0,1]}|f^{(r)}(t)-\mathrm{E}[f^{(r)}(t)|\boldsymbol{Y}]|$ as $\sup_{t\in[0,1]}\left|(\boldsymbol{N}^{(r)}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1/2}\boldsymbol{\varepsilon}^*\right|$ and using the Cauchy-Schwarz inequality and Theorem A.1, the second term inside the bracket in (2.30) is seen to be $O(k_n^{2r+2}/n)$. Combining it with (2.30) and (2.33) and using $2r+2 \leq 4$, we have the desired assertion. □

Lemmas 2.11 and 2.12 can be used to prove the posterior consistency of $\boldsymbol{\theta}$ as shown in the next lemma.

Lemma 2.13. *If $m \geq 5$ and $n^{1/2m} \ll k_n \ll n^{1/8}$, then for all $\epsilon > 0$, $\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon|\boldsymbol{Y}) = o_{P_0}(1)$.*

Proof. By the triangle inequality, using the definition in (2.7),

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

for appropriately chosen constants c_1 and c_2. We denote the set $T_n = \{f : \sup_{t\in[0,1]} |f(t) - f_0(t)| \leq \tau_n, \sup_{t\in[0,1]} |f'(t) - f_0'(t)| \leq \tau_n\}$ for some $\tau_n \to 0$. By Lemma 2.12, there exists such a sequence $\{\tau_n\}$ so that $\Pi(T_n^c|\boldsymbol{Y}) = o_{P_0}(1)$. Hence for $f \in T_n$,

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

Therefore, for any $\delta > 0$, $\Pi(\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} |R_f(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta|\boldsymbol{Y}) = o_{P_0}(1)$. By assumption (2.8), for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon$ there exists a $\delta > 0$ such that

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

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

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

□

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

Lemma 2.14. *Under the conditions of Theorem 2.6, on a set in the sample space with high true probability, the posterior distribution of* $\sqrt{n}\,(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\,(\boldsymbol{\Gamma}(\boldsymbol{f})-\boldsymbol{\Gamma}(\boldsymbol{f}_0))$ *assigns most of its mass inside a large compact set.*

Proof. First note that

$$(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\,\boldsymbol{\Gamma}(\boldsymbol{f})=\sum_{j=1}^{d}\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j$$

and $(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\,\boldsymbol{\Gamma}(\boldsymbol{f}_0)=\sum_{j=1}^{d}\int_0^1\boldsymbol{A}_{,j}(t)f_{j0}(t)dt$, where $\boldsymbol{A}_{,j}(t)$ denotes the j^{th} column of the matrix $\boldsymbol{A}(t)$ as defined in Remark 2.5 for $j=1,\ldots,d$. In order to prove the assertion, we will show that $\mathbf{Var}(\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j|\boldsymbol{Y})$ and $\mathbf{Var}_0(\mathbf{E}(\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j|\boldsymbol{Y}))$ have all eigenvalues of the order n^{-1} and

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

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

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

Hence,

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

Also note that

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

If $A_{k,j}(\cdot)\in C^{m^*}((0,1))$ for some $1\leq m^*<m$, then by equation (2) of De Boor (1978, page 167), we have $\sup\{|A_{k,j}(t)-\tilde{A}_{k,j}(t)|:t\in[0,1]\}=O(k_n^{-1})$, where $\tilde{A}_{k,j}(\cdot)=\boldsymbol{\alpha}_{k,j}^T\boldsymbol{N}(\cdot)$

and $\boldsymbol{\alpha}_{k,j}^T = (A_{k,j}(t_1^*), \ldots, A_{k,j}(t_{k_n+m-1}^*))$ with appropriately chosen $t_1^*, \ldots, t_{k_n+m-1}^*$. We can express $\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{G}_{n,j}$ as

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

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

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

by Theorem A.1 implying that the eigenvalues of the matrix $\int_0^1 \boldsymbol{N}(t)(\boldsymbol{N}(t))^T dt$ are of order k_n^{-1}. The eigenvalues of $\left(\boldsymbol{X}_n^T\boldsymbol{X}_n/n\right)$ are of the order k_n^{-1} by Theorem A.1. Then we have

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

Similarly, it can be shown that $\text{mineig}\left(\tilde{\boldsymbol{G}}_{n,j}^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\tilde{\boldsymbol{G}}_{n,j}\right) \gtrsim n^{-1}$. Let us denote by

$\mathbf{1}_{k_n+m-1}$ the k_n+m-1-component vector with all elements 1. Then for $k=1,\ldots,p$,

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

the last step following by the application of the Cauchy-Schwarz inequality and the facts that

$$\sup\{|A_{k,j}(t)-\tilde{A}_{k,j}(t)| : t\in[0,1]\} = O(k_n^{-1})$$

and $\int_0^1 \|\boldsymbol{N}(t)\|^2\,dt \le 1$. Thus, the eigenvalues of $(\boldsymbol{G}_{n,j}-\tilde{\boldsymbol{G}}_{n,j})^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}(\boldsymbol{G}_{n,j}-\tilde{\boldsymbol{G}}_{n,j})$ are of the order $(nk_n)^{-1}$ or less. Hence, the eigenvalues of $\boldsymbol{G}_{n,j}^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{G}_{n,j}$ are also of the order n^{-1}.

Similarly to the proof of Lemma 2.11, we can write for the $\boldsymbol{\beta}_j^*$ given in (2.29),

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

where $[\boldsymbol{G}_{n,j}^T\boldsymbol{\beta}_j^*]_k = \int_0^1 A_{k,j}(t)f_j^*(t)dt$ and $f_j^*(t)=\boldsymbol{N}^T(t)\boldsymbol{\beta}_j^*$ for $k=1,\ldots,p$. Proceeding in the same way as in the proof of Lemma 2.11, we can show that each term on the right hand side of the above equation converges to zero. Hence, the proof. □

Chapter 3

Bayesian methods based on Runge-Kutta approximation

3.1 Introduction

Although the two-step approach is computationally fast, the Bayes estimator is not asymptotically efficient. Another approach is to use numerical methods to solve the system. A four stage Runge-Kutta (RK4) method is one such method. we propose two separate approaches. The Gaussian distribution is used as the working model for error. 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 $\boldsymbol{f}_{\boldsymbol{\theta},r}(\cdot)$ 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 $\arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} \int_0^1 \left\|\boldsymbol{\beta}^T\boldsymbol{N}(t) - \boldsymbol{f}_{\boldsymbol{\eta},r}(t)\right\|^2 w(t)dt$ 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 $\boldsymbol{N}(\cdot)$ is the B-spline basis vector. We call this approach Runge-Kutta two-step Bayesian (RKTB) method. This approach is similar in spirit to the idea considered in Chapter 2. But instead of using the distance between the derivatives, we consider the distance between function in the nonparametric model and RK4 approximation of the model to extend the definition of the parameter. Ghosh and Goyal (2010) considered Euler's approximation to construct the approximate likelihood and then drew posterior samples. In the same paper they fitted the data using splines and estimated $\boldsymbol{\theta}$ by minimizing the sum of squares of the difference between the spline fitting and the Euler approximation at the grid points. But they did not explore the theoretical aspects of those methods. Both RKSB and RKTB lead to Bernstein-von Mises Theorem with dispersion matrix inverse of Fisher information and hence both the Bayesian methods are asymptotically efficient. This was not the case for the two step-Bayesian approach covered in Chapter 2.

The rest of the chapter is organized as follows. Sections 3.2 contains some preliminaries of Runge-Kutta method. The model assumptions and prior specifications are given in Section 3.3. The main results are given in Section 3.4. In Section 3.5 we have carried out a simulation study. Proofs of the main results are given in Section 3.6. Section 3.7 contains the proofs of the technical lemmas.

3.2 Preliminaries of Runge-Kutta method

Given r equispaced grid points $a_1 = 0, a_2, \ldots, a_r$ with common difference h and an initial condition $\boldsymbol{f}_{\boldsymbol{\theta}}(0) = \boldsymbol{y}_0$, Euler's method (Henrici, 1962, page 9) computes the approximate solution as $\boldsymbol{f}_{\boldsymbol{\theta},r}(a_{k+1}) = \boldsymbol{f}_{\boldsymbol{\theta},r}(a_k) + h\boldsymbol{F}(a_k, \boldsymbol{f}_{\boldsymbol{\theta},r}(a_k), \boldsymbol{\theta})$ for $k = 1, 2, \ldots, r-1$. The RK4 method (Henrici, 1962, page 68) is an improvement over Euler's method. Let us denote

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

Then we obtain $\boldsymbol{f}_{\boldsymbol{\theta},r}(a_{k+1})$ from $\boldsymbol{f}_{\boldsymbol{\theta},r}(a_k)$ as

$$\boldsymbol{f}_{\boldsymbol{\theta},r}(a_{k+1}) = \boldsymbol{f}_{\boldsymbol{\theta},r}(a_{k+1}) + h/6(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4).$$

3.3 Model assumptions and prior specifications

Now we formally describe the model. The proposed model is given by

$$Y_{i,j} = f_{j\boldsymbol{\theta}}(X_i) + \varepsilon_{i,j},\ i = 1, \ldots, n,\ j = 1, \ldots, d, \tag{3.1}$$

where $\boldsymbol{\theta} \subseteq \boldsymbol{\Theta}$, which is a compact subset of $\mathbb{R}^p$. The function $\boldsymbol{f_\theta}(\cdot) := (f_{1\boldsymbol{\theta}}(\cdot), \ldots, f_{d\boldsymbol{\theta}}(\cdot))^T$ satisfies the system of ODE given by

$$\frac{df_{j\boldsymbol{\theta}}(t)}{dt} = F_j(t, \boldsymbol{f_\theta}(t), \boldsymbol{\theta}),\ t \in [0, 1],\ j = 1, \ldots, d. \tag{3.2}$$

Let for a fixed $\boldsymbol{\theta}$, $\boldsymbol{F} \in C^{m-1}((0, 1), \mathbb{R}^d)$ for some integer $m \geq 1$. Then, by successive differentiation we have $\boldsymbol{f_\theta} \in C^m((0, 1))$. 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 $\boldsymbol{f}_0 = (f_{10}, \ldots, f_{d0})^T$ does not necessarily lie in $\{\boldsymbol{f_\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. We assume that $\boldsymbol{f}_0 \in C^m([0, 1])$. Let $\varepsilon_{i,j}$ are identically and independently distributed with mean zero and finite moment generating function for $i = 1, \ldots, n\,; j = 1, \ldots, d$. 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 $X_i \overset{iid}{\sim} G$ with density g. The approximate solution to (3.2) is given by $\boldsymbol{f}_{\theta,r_n}$, where r_n is the number of grid points, which is chosen so that $r_n \gg \sqrt{n}$. By the proof of Theorem 3.3 of Henrici (1962, page 124), we have

$$\sup_{t\in[0,1]} \|\boldsymbol{f_\theta}(t) - \boldsymbol{f}_{\boldsymbol{\theta},r_n}(t)\| = O(r_n^{-1}),\ \sup_{t\in[0,1]} \left\| \frac{\partial}{\partial\boldsymbol{\theta}} \boldsymbol{f_\theta}(t) - \frac{\partial}{\partial\boldsymbol{\theta}} \boldsymbol{f}_{\boldsymbol{\theta},r_n}(t) \right\| = O(r_n^{-1}). \tag{3.3}$$

For the sake of simplicity we assume the response to be one dimensional. The extension to the multidimensional case is straightforward. Let us denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{X} = (X_1, \ldots, X_n)^T$. 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 chapter.

3.3.1 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

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

We also denote

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

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. If $\boldsymbol{\gamma}_0$ is the unique maximizer of the right hand side above, we get

$$\int_0^1 \dot{f}_{\boldsymbol{\theta}_0}^T(t)\left(f_0(t) - f_{\boldsymbol{\theta}_0}(t)\right)g(t)dt = \mathbf{0},\ \sigma_*^2 = \sigma_0^2 + \int_0^1 |f_0(t) - f_{\boldsymbol{\theta}_0}(t)|^2 g(t)dt. \tag{3.6}$$

We assume that the sub-matrix of the Hessian matrix of $-P_0 \log p_{\boldsymbol{\gamma}}$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ given by

$$\int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial\boldsymbol{\theta}}\left(\dot{f}_{\boldsymbol{\theta}}^T(t)\left(f_0(t) - f_{\boldsymbol{\theta}_0}(t)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)g(t)dt \tag{3.7}$$

is positive definite. The prior measure on $\boldsymbol{\Theta}$ is assumed to have a Lebesgue-density continuous and positive on a neighborhood of $\boldsymbol{\theta}_0$. The prior distribution of $\boldsymbol{\theta}$ is assumed to be independent of that of σ^2. The joint prior measure is denoted by Π with corresponding density π. We obtain the posterior of $\boldsymbol{\gamma}$ using the approximate likelihood given by (3.4).

3.3.2 Runge-Kutta Two-step Bayesian Method (RKTB)

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

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

where $\boldsymbol{X}_n = ((N_j(X_i)))_{1\leq i\leq n, 1\leq j\leq k_n+m-1}$, $\{N_j(\cdot)\}_{j=1}^{k_n+m-1}$ being the B-spline basis functions of order m with $k_n - 1$ interior knots. We assume for a given σ^2

$$\boldsymbol{\beta} \sim N_{k_n+m-1}(\boldsymbol{0}, \sigma^2 n^2 k_n^{-1} \boldsymbol{I}_{k_n+m-1}). \tag{3.9}$$

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

$$N_{k_n+m-1}\left(\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}, \sigma^2\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\right). \tag{3.10}$$

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

$$R_{\boldsymbol{f},n}(\boldsymbol{\eta}) = \left\{\int_0^1 |f(t) - f_{\eta,r_n}(t)|^2 g(t)dt\right\}^{1/2}, \; R_{\boldsymbol{f}_0}(\boldsymbol{\eta}) = \left\{\int_0^1 |f_0(t) - f_{\eta}(t)|^2 g(t)dt\right\}^{1/2},$$

where $f(t) = \boldsymbol{\beta}^T\boldsymbol{N}(t)$. Now we define $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} R_{f,n}(\boldsymbol{\eta})$ and induce posterior distribution on $\boldsymbol{\Theta}$ through the posterior of $\boldsymbol{\beta}$ given by (3.10). Also let us define $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} R_{f_0}(\boldsymbol{\eta})$. 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 assume that

$$\text{for all}\, \epsilon > 0, \quad \inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} R_{f_0}(\boldsymbol{\eta}) > R_{f_0}(\boldsymbol{\theta}_0). \tag{3.11}$$

3.4 Main results

The main results of our work are given by Theorems 3.1 and 3.3.

Theorem 3.1. *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*

$$\left\|\Pi_n\left(\sqrt{n}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\in\cdot|\boldsymbol{X},\boldsymbol{Y}\right)-\boldsymbol{N}(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0},\sigma_*^2\boldsymbol{V}_{\boldsymbol{\gamma}_0}^{-1})\right\|_{TV}\stackrel{P_0}{\to}0,$$

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

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

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

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 $\sigma_*^2\boldsymbol{V}_{\boldsymbol{\gamma}_0}^{-1}$ respectively. We also get the following important corollary.

Corollary 3.2. *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.*

In RKTB we assume that the matrix

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

is nonsingular. Let us denote $\boldsymbol{C}(t)=(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\left(\dot{f}_{\boldsymbol{\theta}_0}(t)\right)^T$ and $\boldsymbol{H}_n^T=\int_0^1\boldsymbol{C}(t)\boldsymbol{N}^T(t)g(t)dt$. Also, we denote the posterior probability measure of RKTB by Π_n^*. Now we have the following result.

Theorem 3.3. *Let*

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

If $\boldsymbol{B}$ is non-singular, then for $m \geq 3$ and $n^{1/(2m)} \ll k_n \ll n^{1/4}$,

$$\left\|\Pi_n^*\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \in \cdot|Y\right) - N\left(\boldsymbol{\mu}_n, \sigma_0^2\boldsymbol{\Sigma}_n\right)\right\|_{TV} = o_{P_0}(1). \tag{3.12}$$

Remark 3.4. It will be proved later in Lemma 3.16 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.

We also get the following important corollary.

Corollary 3.5. *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 3.6. RKSB is the Bayesian analog of estimating $\boldsymbol{\theta}$ as

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

Similarly, RKTB is the Bayesian analog of

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

where $\hat{f}(\cdot)$ stands for the nonparametric estimate of f_0 based on B-splines. Arguments similar to ours should be able to establish analogous convergence results for these estimators.

3.5 Simulation Study

We consider the Lotka-Volterra equations to study the posterior distribution of $\boldsymbol{\theta}$. We consider the situation where the true regression function belongs to the solution set. 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. Under each replication a sample of size 1000 is drawn from the posterior distribution of $\boldsymbol{\theta}$ using RKSB, RKTB and Bayesian two-step methods and then 95% equal tailed credible intervals are obtained. Bayesian two-step method is abbreviated as "TS" in the table. We calculate the coverage and the average length of the corresponding credible interval over these 900 replications. The estimated standard errors of the interval length and coverage are given inside the parentheses in the table. Thus we have $p = 4, d = 2$ and the ODE's are given by

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

for $t \in [0, 1]$ with initial condition $f_{1\boldsymbol{\theta}}(0) = 1, f_{2\boldsymbol{\theta}}(0) = 0.5$. The true parameter vector is chosen as $\boldsymbol{\theta}_0 = (10, 10, 10, 10)^T$. The above system is not analytically solvable. 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. We take $m = 3$ and $m = 5$ for RKTB and Bayesian two-step method respectively. Looking at the order of k_n suggested by Theorem 3.3, k_n is chosen as 13 and 18 for $n = 100$ and $n = 500$ respectively in RKTB. In Bayesian two-step method the choices are 17 and 20 for $n = 100$ and $n = 500$ respectively. The simulation results are summarized in the Table 3.1. Not surprisingly the first two methods lead to shorter intervals compared to the third one because of asymptotic efficiency obtained from Corollaries 3.2 and 3.5 respectively.

3.6 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 3.1. From Lemma 3.7 we know that there exists a compact subset U of $(0, \infty)$ such that $\Pi_n(\sigma^2 \in U | \boldsymbol{X}, \boldsymbol{Y}) \xrightarrow{P_0} 1$. Let $\Pi_{U,n}(\cdot | \boldsymbol{X}, \boldsymbol{Y})$ be the posterior distribution conditioned on $\sigma^2 \in U$. By Theorem A.2 if we can ensure that there exist stochastically bounded random variables $\boldsymbol{\Delta}_{n,\gamma_0}$ and a positive definite matrix $\boldsymbol{V}_{\gamma_0}$ such that for every

Table 3.1: *Coverages and average lengths of the Bayesian credible intervals for the three methods*

n		RKSB		RKTB		TS	
		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	6.96
		(0.00)	(0.29)	(0.00)	(0.65)	(0.00)	(4.99)
	θ_2	100.0	2.57	100.0	2.48	100.0	6.70
		(0.00)	(0.33)	(0.00)	(0.74)	(0.00)	(4.94)
	θ_3	99.9	2.50	100.0	2.44	100.0	7.14
		(0.00)	(0.34)	(0.00)	(1.44)	(0.00)	(4.92)
	θ_4	100.0	2.27	100.0	2.20	100.0	6.62
		(0.00)	(0.32)	(0.00)	(1.19)	(0.00)	(4.78)
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)	(0.04)
	θ_3	100.0	0.82	99.3	0.61	99.0	1.16
		(0.00)	(0.07)	(0.00)	(0.02)	(0.00)	(0.06)
	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)

compact set $K \subset \mathbb{R}^{p+1}$,

$$\sup_{h \in K} \left| \log \frac{p^{(n)}_{\boldsymbol{\gamma}_0 + \boldsymbol{h}/\sqrt{n}, n}}{p^{(n)}_{\boldsymbol{\gamma}_0, n}} (\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{h}^T \boldsymbol{V}_{\boldsymbol{\gamma}_0} \boldsymbol{\Delta}_{n, \boldsymbol{\gamma}_0} + \frac{1}{2} \boldsymbol{h}^T \boldsymbol{V}_{\boldsymbol{\gamma}_0} \boldsymbol{h} \right| \to 0, \tag{3.13}$$

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

$$P_0^{(n)} \Pi_{U,n} \left(\sqrt{n} \, \| \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \| > M_n | \boldsymbol{X}, \boldsymbol{Y} \right) \to 0, \tag{3.14}$$

then

$$\left\| \Pi_{U,n} \left(\sqrt{n} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \in \cdot | \boldsymbol{X}, \boldsymbol{Y} \right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n, \boldsymbol{\gamma}_0}, \boldsymbol{V}_{\boldsymbol{\gamma}_0}^{-1}) \right\|_{TV} \xrightarrow{P_0} 0.$$

We show that the conditions (3.13) and (3.14) hold in Lemmas 3.7 − 3.11. Lemma 3.8 gives that $\boldsymbol{V}_{\gamma_0} = \begin{pmatrix} \sigma_*^{-2}\boldsymbol{V}_{\theta_0} & \mathbf{0} \\ \mathbf{0} & \sigma_*^{-4}/2 \end{pmatrix}$ with

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

and $\boldsymbol{\Delta}_{n,\gamma_0} = \boldsymbol{V}_{\gamma_0}^{-1}\mathbb{G}_n\dot{\ell}_{\gamma_0,n}$. Since $\|\Pi_n - \Pi_{U,n}\|_{TV} = o_{P_0}(1)$, we get

$$\left\|\Pi_n \left(\sqrt{n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \in \cdot | \boldsymbol{X}, Y \right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n,\gamma_0}, \boldsymbol{V}_{\gamma_0}^{-1})\right\|_{TV} \overset{P_0}{\to} 0.$$

Hence, we get the desired result. □

Proof of Corollary 3.2. The log-likelihood of the correctly specified model with Gaussian error is given by

$$\ell_{\gamma_0}(X,Y) \;\; = \;\; -\log\sigma_0 - \frac{1}{2\sigma_0^2}|Y - f_{\boldsymbol{\theta}_0}(X)|^2 + \log g(X).$$

Thus

$$\frac{\partial}{\partial \boldsymbol{\theta}_0}\ell_{\gamma_0}(X,Y) = \sigma_0^{-2}\left(\dot{f}_{\boldsymbol{\theta}_0}(X)\right)^T (Y - f_{\boldsymbol{\theta}_0}(X))$$

and

$$\frac{\partial}{\partial \sigma_0^2}\ell_{\gamma_0}(X,Y) = -\frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_0^4}|Y - f_{\boldsymbol{\theta}_0}(X)|^2.$$

Hence, the Fisher information is given by

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

Thus $\boldsymbol{V}_{\gamma_0}^{-1} = (\boldsymbol{I}(\boldsymbol{\gamma}_0))^{-1}$ if the regression function is correctly specified and the true error distribution is $N(0,\sigma_0^2)$. □

Proof of Theorem 3.3. We have for $f(\cdot) = \boldsymbol{\beta}^T\boldsymbol{N}(\cdot)$

$$\int_0^1 \boldsymbol{C}(t)\boldsymbol{\beta}^T\boldsymbol{N}(t)g(t)dt = \boldsymbol{H}_n^T\boldsymbol{\beta}, \tag{3.15}$$

where $\boldsymbol{H}_n^T = \int_0^1 \boldsymbol{C}(t)\boldsymbol{N}^T(t)g(t)dt$ which is a matrix of order $p \times (k_n + m - 1)$. Consequently, the asymptotic variance of the conditional posterior distribution of $\boldsymbol{H}_n^T\boldsymbol{\beta}$ is $\sigma^2 \boldsymbol{H}_n^T \left(\boldsymbol{X}_n^T\boldsymbol{X}_n + \frac{k_n}{n^2}\boldsymbol{I}_{k_n+m-1}\right)^{-1} \boldsymbol{H}_n$. By Lemma 3.15 and the posterior consistency of the σ^2 given by Lemma 3.17, it suffices to show that for any neighborhood $\mathscr{N}$ of σ_0^2,

$$\sup_{\sigma^2 \in \mathscr{N}} \Bigg\| \Pi_n^* \left(\sqrt{n}\boldsymbol{H}_n^T\boldsymbol{\beta} - \sqrt{n}\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1} \int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}(t)\right)^T f_0(t)g(t) \in \cdot | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2\right) \\ - N(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Sigma}_n)\Bigg\|_{TV} = o_{P_0}(1). \tag{3.16}$$

Note that $\Pi(\mathscr{N}^c|\boldsymbol{X}, \boldsymbol{Y}) = o_{P_0}(1)$. It is straightforward to verify that the Kullback-Leibler divergence between $N\left(\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}, \sigma^2\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\right)$ and the distribution given by (3.10) converges in P_0-probability to zero uniformly over $\sigma^2 \in \mathscr{N}$ and hence, so is the total variation distance. By linear transformation (3.15), we get (3.16). Note that

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

Using the fact that $\boldsymbol{\Sigma}_n$ is stochastically bounded given by Lemma 3.16, 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 $\mathscr{N}$ accordingly. The first term converges in probability to zero by (3.16). The third term converges in probability to zero by the posterior consistency. □

Proof of Corollary 3.5. The log-likelihood of the correctly specified model is given by

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

Thus $\dot{\ell}_{\boldsymbol{\theta}_0}(X,Y) = -\sigma_0^{-2}\left(\dot{f}_{\boldsymbol{\theta}_0}(X)\right)^T(Y - f_{\boldsymbol{\theta}_0}(X))$ and the Fisher information is given by $\boldsymbol{I}(\boldsymbol{\theta}_0) = \sigma_0^{-2}\int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}(X)\right)^T \dot{f}_{\boldsymbol{\theta}_0}(t)g(t)dt$. In the proof of Lemma 3.16 we obtained that

$$\sigma_0^2\boldsymbol{\Sigma}_n \xrightarrow{P_0} \sigma_0^2\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}\int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}(t)\right)^T \dot{f}_{\boldsymbol{\theta}_0}(t)g(t)dt\ \left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}.$$

This limit is equal to $\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}$ under the correct specification of the regression function as well as the likelihood. □

3.7 Proofs of the 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 3.7. *There exists a compact set U independent of $\boldsymbol{\theta}$ and n such that* $\Pi_n(\sigma^2 \in U|\boldsymbol{X},\boldsymbol{Y}) \xrightarrow{P_0} 1$.

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. Hence, it is easy to show that the mean of the conditional posterior of σ^2 converges to $\sigma_{\boldsymbol{\theta}}^2 := \sigma_0^2 + \int_0^1 \left(f_0(t) - f_{\boldsymbol{\theta}}(t)\right)^2 g(t)dt$ in P_0-probability. Then it follows that for any $\epsilon > 0$, $\Pi_n\left(\sigma^2 \in [\sigma_{\boldsymbol{\theta}}^2 - \epsilon, \sigma_{\boldsymbol{\theta}}^2 + \epsilon]|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}\right)$ converges in P_0-probability to 1. Since $\boldsymbol{\Theta}$ 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 $\Pi_n\left(\sigma^2 \notin U|\boldsymbol{X},\boldsymbol{Y}\right)$ is bounded above by

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

It suffices to prove that

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

Clearly $\mathrm{E}(\sigma^2|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n (Y_i - f_{\boldsymbol{\theta}}(X_i))^2 + o(1)\,a.s.$ 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 A.6 and Example 19.7 of Van der Vaart (1998), it follows that

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

Also, it can be easily shown that the quantity $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\mathrm{E}_0\left(\mathrm{E}(\sigma^2|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})\right) - \sigma^2_{\boldsymbol{\theta}}\right| \to 0$ as $n \to \infty$ which gives the first assertion. To see the second assertion, observe that $\mathrm{Var}(\sigma^2|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}) = n^{-1}O(1)\,a.s.$ by the previous assertion and the fact that the conditional posterior of σ^2 given $\boldsymbol{\theta}$ is inverse gamma. $\square$

In view of the previous lemma, we choose the parameter space for $\boldsymbol{\gamma}$ to be $\boldsymbol{\Theta}\times U$ from now onwards. We show that the condition (3.13) holds by the following lemma.

Lemma 3.8. *For the model induced by Runge-Kutta method as described in Section 3.3.1, we have*

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

in (outer) $P_0^{(n)}$-probability for every compact set $\boldsymbol{K}\subset\mathbb{R}^{p+1}$, where $\boldsymbol{\Delta}_{n,\gamma_0} = \boldsymbol{V}_{\gamma_0}^{-1}\mathbb{G}_n\dot{\ell}_{\boldsymbol{\gamma}_0,n}$ and $\boldsymbol{V}_{\gamma_0} = \begin{pmatrix}\sigma_*^{-2}\boldsymbol{V}_{\theta_0} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma_*^{-4}/2\end{pmatrix}$ for some positive definite matrix $\boldsymbol{V}_{\boldsymbol{\theta}_0}$.

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

$$\left|\log(p_{\boldsymbol{\gamma}_1}(X_1,Y_1)/p_{\boldsymbol{\gamma}_2}(X_1,Y_1))\right| \le m(X_1,Y_1)\left\|\gamma-\gamma_0\right\|,$$

where $m(X_1,Y_1)$ is

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

which is square integrable. Therefore, by Lemma A.7, for any sequence $\{\boldsymbol{h}_n\}$ bounded in P_0-probability, $\mathbb{G}_n\left(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\boldsymbol{h}_n/\sqrt{n})} - \ell_{\boldsymbol{\gamma}_0}) - \boldsymbol{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}\right) \stackrel{P_0}{\to} 0$. Using the laws of large numbers and (3.3), we find that

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

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

$$\mathbb{G}_n\left(\sqrt{n}(\ell_{\boldsymbol{\gamma}_0+(\boldsymbol{h}_n/\sqrt{n}),n}-\ell_{\boldsymbol{\gamma}_0,n})-\boldsymbol{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}\right)\xrightarrow{P_0}0.$$

We note that

$$\begin{aligned}
&-P_0\log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\\
&= \log\sigma-\log\sigma_*+\frac{1}{2\sigma^2}\left[\sigma_0^2+\int_0^1|f_0(t)-f_{\boldsymbol{\theta},r_n}(t)|^2g(t)dt\right]\\
&\quad -\frac{1}{2\sigma_*^2}\left[\sigma_0^2+\int_0^1|f_0(t)-f_{\boldsymbol{\theta}_0,r_n}(t)|^2g(t)dt\right]\\
&= \log\sigma-\log\sigma_*+\left(\frac{1}{2\sigma^2}-\frac{1}{2\sigma_*^2}\right)\left[\sigma_0^2+\int_0^1|f_0(t)-f_{\boldsymbol{\theta},r_n}(t)|^2g(t)dt\right]\\
&\quad +\frac{1}{2\sigma_*^2}\left[2\int_0^1\left(f_0(t)-f_{\boldsymbol{\theta}_0,r_n}(t)\right)\left(f_{\boldsymbol{\theta}_0,r_n}(t)-f_{\boldsymbol{\theta},r_n}(t)\right)g(t)dt\right.\\
&\quad \left.+\int_0^1|f_{\boldsymbol{\theta}_0,r_n}(t)-f_{\boldsymbol{\theta},r_n}(t)|^2g(t)dt\right].
\end{aligned}\tag{3.17}$$

Using (3.6), the last term inside the third bracket in (3.17) can be expanded as

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\boldsymbol{V}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+O\left(r_n^{-1}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\right)+o\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\right),$$

where $\boldsymbol{V}_{\boldsymbol{\theta}_0}=\int_0^1\left(\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}^T(t)\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)-\frac{\partial}{\partial\boldsymbol{\theta}}\left(\dot{\boldsymbol{f}}_{\boldsymbol{\theta}}^T(t)\left(f_0(t)-f_{\boldsymbol{\theta}_0}(t)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)g(t)dt$. Also, writing $\sigma_*^2=\sigma_0^2+\int_0^1|f_0(t)-f_{\boldsymbol{\theta}_0}(t)|^2dG(t)$ and using (3.6), the first term in (3.17) is given by

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

Hence,

$$
P_0 \log \frac{p_{\boldsymbol{\gamma}_0+\boldsymbol{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}}+\frac{1}{2n}\boldsymbol{h}_n^T\boldsymbol{V}_{\boldsymbol{\gamma}_0}\boldsymbol{h}_n=o(n^{-1}). \tag{3.18}
$$

We have already shown that

$$
n\mathbb{P}_n \log \frac{p_{\boldsymbol{\gamma}_0+\boldsymbol{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}}-\mathbb{G}_n\boldsymbol{h}_n^T\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}-nP_0\log\frac{p_{\boldsymbol{\gamma}_0+\boldsymbol{h}_n/\sqrt{n},n}}{p_{\boldsymbol{\gamma}_0,n}}=o_{P_0}(1). \tag{3.19}
$$

Substituting (3.18) in (3.19), we get the desired result. □

Now our objective is to prove (3.14). We define the measure $Q_{\boldsymbol{\gamma}}(A)=P_0\left(p_{\boldsymbol{\gamma}}/p_{\boldsymbol{\gamma}_0}\mathbb{1}_A\right)$ and the corresponding density $q_{\boldsymbol{\gamma}}$ as given in Kleijn and van der Vaart (2012). Also, we define a measure $Q_{\boldsymbol{\gamma},n}$ by $Q_{\boldsymbol{\gamma},n}(A)=P_0\left(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}\mathbb{1}_A\right)$ with $q_{\boldsymbol{\gamma},n}$ being the corresponding density. The misspecified Kullback-Leibler neighborhood of $\boldsymbol{\gamma}_0$ is defined as

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

By Theorem A.3, condition (3.14) is satisfied if we can ensure that for every $\epsilon>0$, there exists a sequence of tests $\{\phi_n\}$ such that

$$
P_0^n\phi_n\to 0, \quad \sup_{\{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\geq\epsilon\}} Q_{\boldsymbol{\gamma},n}^n(1-\phi_n)\to 0. \tag{3.20}
$$

The above condition is ensured by the next lemma.

Lemma 3.9. *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* (3.20).

Proof. For given $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_0$ consider the tests $\phi_{n,\boldsymbol{\gamma}_1} = \mathbb{1}\{\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n}) < 0\}$. As a result we have $\mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1,n}) = \mathbb{P}_n \log(p_0/q_{\boldsymbol{\gamma}_1}) + O_{P_0}(r_n^{-1}) \stackrel{P_0^n}{\to} P_0 \log(p_0/q_{\boldsymbol{\gamma}_1})$ 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

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

for $\rho(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}, s) = \int p_0^s q_{\boldsymbol{\gamma}_1}^{-s} q_{\boldsymbol{\gamma}} d\mu$. Following the proof of Theorem A.4, we can assert 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 $i = 1, \ldots, k$. Let us define $\phi_n = \max_i\{\phi_{n,\boldsymbol{\gamma}_i} : i = 1, \ldots, k\}$. This test satisfies $P_0^n \phi_n \leq \sum_{i=1}^k P_0^n \phi_{n,\boldsymbol{\gamma}_i} \to 0$, and

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

uniformly in $\boldsymbol{\gamma} \in \cup_{i=1}^k G_{\boldsymbol{\gamma}_i}$. Therefore, the tests ϕ_n meets (3.20). □

The proof of Theorem A.3 also uses the results of the next two lemmas.

Lemma 3.10. *Suppose that $P_0 \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0} \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}^T$ 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,*

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

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

Proof. Let $\{M_n\}$ be given. We construct two sequences of tests. The first sequence is used to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \boldsymbol{\gamma} \in (\boldsymbol{\Theta} \times U)_1\}$ with

$$(\boldsymbol{\Theta} \times U)_1 = \{\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : M_n/\sqrt{n} \leq \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \leq \epsilon\}$$

and the second to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \boldsymbol{\gamma} \in (\boldsymbol{\Theta} \times U)_2\}$ with

$$(\boldsymbol{\Theta} \times U)_2 = \{\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| > \epsilon\}.$$

These two sequences are combined to test P_0 versus $\{Q_{\boldsymbol{\gamma},n} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \geq M_n/\sqrt{n}\}$.
To construction the first sequence, a constant $L > 0$ is chosen to truncate the score-function, that is, $\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0} = 0$ if $\|\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}\| > L$ and $\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0} = \dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}$ otherwise. We define

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

Since the function $\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0}$ is square-integrable, we observe that the matrices $P_0\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}\dot{\boldsymbol{\ell}}^T_{\boldsymbol{\gamma}_0,n}$, $P_0\dot{\boldsymbol{\ell}}_{\boldsymbol{\gamma}_0,n}(\dot{\boldsymbol{\ell}}^L)^T_{\boldsymbol{\gamma}_0,n}$ and $P_0\dot{\boldsymbol{\ell}}^L_{\boldsymbol{\gamma}_0,n}(\dot{\boldsymbol{\ell}}^L)^T_{\boldsymbol{\gamma}_0,n}$ are also square integrable for a sufficiently large choices of L and n. We fix such an L. Now,

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

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 A.5 and Lemma 3.8. □

Lemma 3.11. *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 $\Pi(B(\epsilon_n, \boldsymbol{\gamma}_0, P_0)) \geq K\epsilon_n^p$, where $\epsilon_n \gg n^{-1/2}$.*

Proof. From the proof of Lemma 3.8 we get

$$-P_0 \log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n}) = O(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2) + O(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| r_n^{-1}) \leq c_1 \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 + c_2 \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \epsilon_n$$

for sufficiently large n, c_1, c_2 being a suitably chosen constants. Again,

$$P_0 \left(\log(p_{\boldsymbol{\gamma},n}/p_{\boldsymbol{\gamma}_0,n})\right)^2 \leq c_3 \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2$$

for some constant $c_3 > 0$. Let us define the quantity

$$c = \min\left((2c_1)^{-1/2}, (2c_2)^{-1}, c_3^{-1/2}\right).$$

Then $\{\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leq c\epsilon_n\} \subset B(\epsilon_n, \boldsymbol{\gamma}_0, P_0)$. 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 $0 < \delta \leq \delta'$,

$$\Pi\left(\boldsymbol{\gamma} \in \boldsymbol{\Theta} \times U : \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leq \delta\right) \geq \frac{1}{2} V \pi(\boldsymbol{\gamma}_0) \delta^p > 0,$$

V being the Lebesgue-volume of the $(p+1)$-dimensional ball of unit radius. Hence, for sufficiently large n, $c\epsilon_n \leq \delta'$ and we obtain the desired result. □

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

Lemma 3.12. *For $m \geq 3$ and $n^{1/(2m)} \ll k_n \ll n^{1/4}$,*

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

.

Proof. By (3.10),

$$\mathrm{E}(f(t)|\boldsymbol{X}, \boldsymbol{Y}, \sigma^2) = (\boldsymbol{N}(t))^T (\boldsymbol{X}_n^T \boldsymbol{X}_n + k_n n^{-2} \boldsymbol{I}_{k_n+m-1})^{-1} \boldsymbol{X}_n^T \boldsymbol{Y}. \tag{3.21}$$

By Theorem A.1 we have uniformly over $t \in [0,1]$,

$$(\boldsymbol{N}(t))^T (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \boldsymbol{N}(t) \asymp \frac{k_n}{n}(1 + o_{P_0}(1)). \tag{3.22}$$

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

$$\sup_{t\in[0,1]} |f_0(t) - (\boldsymbol{N}(t))^T\boldsymbol{\beta}^*| = O(k_n^{-m}). \tag{3.23}$$

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

$$\begin{aligned}
&\sup_{t\in[0,1]} \left|\mathrm{E}(f(t)|\boldsymbol{X},\boldsymbol{Y},\sigma^2) - (N(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}\right|^2 \\
&\quad + \sup_{t\in[0,1]} \left|(\boldsymbol{N}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T(\boldsymbol{Y} - f_0(\boldsymbol{x}))\right|^2 \\
&\quad + \sup_{t\in[0,1]} |(\boldsymbol{N}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T(f_0(\boldsymbol{x}) - \boldsymbol{X}_n\boldsymbol{\beta}^*)|^2 \\
&\quad + \sup_{t\in[0,1]} |f_0(t) - (\boldsymbol{N}(t))^T\boldsymbol{\beta}^*|^2.
\end{aligned} \tag{3.24}$$

Using Theorem A.9, we get

$$\begin{aligned}
&\mathrm{E}(f(t)|\boldsymbol{X},\boldsymbol{Y},\sigma^2) - (N(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y} \\
&= (N(t))^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + \frac{k_n}{n^2}\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y} - (N(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y} \\
&= \frac{k_n}{n^2}(N(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\left(\boldsymbol{I}_{k_n+m-1} + \frac{k_n}{n^2}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right)^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}.
\end{aligned}$$

Now applying the Cauchy-Schwarz inequality and (3.22), the first term of (3.24) can be shown to be $O_{P_0}(k_n^5/n^6)$. The second term can be bounded up to a constant multiple by

$$\begin{aligned}
&\max_{1\le k\le n} \left|(\boldsymbol{N}(s_k))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{\varepsilon}\right|^2 \\
&\quad + \sup_{t,t':|t-t'|\le n^{-1}} \left|(\boldsymbol{N}(t) - \boldsymbol{N}(t'))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{\varepsilon}\right|^2,
\end{aligned} \tag{3.25}$$

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

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

By the spectral decomposition, we can write $\boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T = \boldsymbol{P}^T\boldsymbol{D}\boldsymbol{P}$, where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{D}$ 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

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

Note that

$$\mathbf{Var}_0(\boldsymbol{P}\boldsymbol{\varepsilon}) = \mathbf{E}_0(\mathrm{Var}(\boldsymbol{P}\boldsymbol{\varepsilon}|\boldsymbol{X})) + \mathbf{Var}_0(\mathrm{E}(\boldsymbol{P}\boldsymbol{\varepsilon}|\boldsymbol{X})) = \sigma_0^2\boldsymbol{I}_{k_n+m-1}.$$

As a result, we get $\mathbf{E}_0(\boldsymbol{\varepsilon}^T\boldsymbol{P}^T\boldsymbol{D}\boldsymbol{P}\boldsymbol{\varepsilon}) = \sigma_0^2(k_n+m-1)$. Now using Theorem A.1, we can conclude that the first term of (3.25) is $O_{P_0}(k_n^2/n)$. Again applying the Cauchy-Schwarz inequality, the second term of (3.25) is bounded by

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

which is $O_{P_0}\left(n(k_n^3/n^3)\right) = O_{P_0}\left(k_n^3/n^2\right)$, using Theorem A.1. Thus, the second term of (3.24) is $O_{P_0}\left(k_n^2/n\right)$. Using the Cauchy-Schwarz inequality, (3.22) and (3.23), the third term of (3.24) is $O_{P_0}\left(k_n^{1-2m}\right)$. The fourth term of (3.24) is of the order of k_n^{-2m} as a result of (3.23). □

The following lemma controls posterior variability in RKTB.

Lemma 3.13. *If $m \ge 3$ and $n^{1/(2m)} \ll k_n \ll n^{1/4}$, then for all $\epsilon > 0$,*

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

Proof. By Markov's inequality and the fact that $|a+b|^2 \le 2(|a|^2+|b|^2)$ for two real numbers a and b, we can bound $\Pi_n^*\left(\sup_{t\in[0,1]}|f(t)-f_0(t)|>\epsilon|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right)$ by

$$2\epsilon^{-2}\left\{\sup_{t\in[0,1]}\left|\mathrm{E}(f(t)|\boldsymbol{X},\boldsymbol{Y},\sigma^2)-f_0(t)\right|^2\right.$$
$$\left.+\mathrm{E}\left[\sup_{t\in[0,1]}\left|f(t)-\mathrm{E}(f(t)|\boldsymbol{X},\boldsymbol{Y},\sigma^2)\right|^2|\boldsymbol{X},\boldsymbol{Y},\sigma^2\right]\right\}. \tag{3.26}$$

By Lemma 3.12, the first term inside the bracket above is $O_{P_0}(k_n^2/n)+O_{P_0}(k_n^{1-2m})$. Let us define $\boldsymbol{\varepsilon}^*=(\boldsymbol{X}_n^T\boldsymbol{X}_n+k_nn^{-2}\boldsymbol{I}_{k_n+m-1})^{1/2}\boldsymbol{\beta}-(\boldsymbol{X}_n^T\boldsymbol{X}_n+k_nn^{-2}\boldsymbol{I}_{k_n+m-1})^{-1/2}\boldsymbol{X}_n^T\boldsymbol{Y}$. It follows that $\boldsymbol{\varepsilon}^*|\boldsymbol{X},\boldsymbol{Y},\sigma^2\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_{k_n+m-1})$. Noting that $\sup_{t\in[0,1]}|f(t)-\mathrm{E}[f(t)|\boldsymbol{X},\boldsymbol{Y},\sigma^2]|$ can be written as

$$\sup_{t\in[0,1]}\left|(\boldsymbol{N}(t))^T(\boldsymbol{X}_n^T\boldsymbol{X}_n+k_nn^{-2}\boldsymbol{I}_{k_n+m-1})^{-1/2}\boldsymbol{\varepsilon}^*\right|$$

and using the Cauchy-Schwarz inequality and Theorem A.1, the second term inside the bracket in (3.26) 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 3.12 and 3.13.

Lemma 3.14. *If $m\geq 3$ and $n^{1/(2m)}\ll k_n\ll n^{1/4}$, then for all $\epsilon>0$, $\Pi_n^*(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\epsilon|\boldsymbol{X},\boldsymbol{Y},\sigma^2)=o_{P_0}(1)$.*

Proof. By the triangle inequality,

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

for appropriately chosen constants c_1' and c_2'. We denote the set $T_n=\{f:\sup_{t\in[0,1]}|f(t)-f_0(t)|\leq\tau_n\}$ for some $\tau_n\to 0$. By Lemma 3.13, there exists such a sequence $\{\tau_n\}$ so that $\Pi(T_n^c|\boldsymbol{X},\boldsymbol{Y},\sigma^2)=o_{P_0}(1)$. Hence for $f\in T_n$,

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

Therefore, for any $\delta > 0$, $\Pi_n^* \left(\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} |R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2\right) = o_{P_0}(1)$. By assumption (3.11), for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon$ there exists a $\delta > 0$ such that

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

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

$$\Pi_n^*(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2) \leq \Pi_n^* \left(\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} |R_{f,n}(\boldsymbol{\eta}) - R_{f_0}(\boldsymbol{\eta})| > \delta/2 | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2 \right) = o_{P_0}(1).$$

□

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 3.3 to obtain the limiting posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$.

Lemma 3.15. *Let m be an integer greater than or equal to 3 and $n^{1/(2m)} \ll k_n \ll n^{1/4}$. Then there exists $E_n \subseteq C^m((0,1)) \times \boldsymbol{\Theta}$ with $\Pi(E_n^c | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2) = o_{P_0}(1)$, such that uniformly for $(f, \boldsymbol{\theta}) \in E_n$,*

$$\left\| \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - (\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1} \sqrt{n}(\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0)) \right\| \to 0 \tag{3.27}$$

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

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

$$\int_0^1 \left(\dot{f}_{\theta,r_n}(t)\right)^T (f(t) - f_{\theta,r_n}(t)) g(t) dt = \mathbf{0}, \tag{3.28}$$

$$\int_0^1 \left(\dot{f}_{\theta_0}(t)\right)^T (f_0(t) - f_{\theta_0}(t)) g(t) dt = \mathbf{0}. \tag{3.29}$$

We can rewrite (3.28) as

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

Subtracting (3.29) from the above equation we get

$$
\begin{aligned}
&\int_0^1 \left(\dot{f}_{\theta_0}(t)\right)^T (f(t) - f_0(t))g(t)dt - \int_0^1 \left(\dot{f}_{\theta_0}(t)\right)^T (f_{\boldsymbol{\theta}}(t) - f_{\boldsymbol{\theta}_0}(t))g(t)dt \\
&\quad + \int_0^1 \left(\dot{f}_{\theta}(t) - \dot{f}_{\theta_0}(t)\right)^T (f(t) - f_{\boldsymbol{\theta}}(t))g(t)dt \\
&\quad + \int_0^1 \left(\dot{f}_{\theta,r_n}(t) - \dot{f}_{\theta}(t)\right)^T (f(t) - f_{\boldsymbol{\theta}}(t))g(t)dt \\
&\quad + \int_0^1 \left(\dot{f}_{\theta,r_n}(t)\right)^T (f_{\boldsymbol{\theta}}(t) - f_{\boldsymbol{\theta},r_n}(t))g(t)dt = \mathbf{0}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\boldsymbol{M}(f, \boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\
&= \int_0^1 \left(\dot{f}_{\theta_0}(t)\right)^T (f(t) - f_0(t))g(t)dt + \int_0^1 \left(\dot{f}_{\theta,r_n}(t) - \dot{f}_{\theta}(t)\right)^T (f(t) - f_{\boldsymbol{\theta}}(t))g(t)dt \\
&\quad + \int_0^1 \left(\dot{f}_{\theta,r_n}(t)\right)^T (f_{\boldsymbol{\theta}}(t) - f_{\boldsymbol{\theta},r_n}(t))g(t)dt,
\end{aligned}
$$

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

$$\begin{aligned}
&\int_0^1\int_0^1 \left(\dot{f}_{\theta_0}(t)\right)^T \dot{f}_{\theta_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)d\lambda g(t)dt\\
&\quad -\int_0^1\int_0^1 \ddot{f}_{\boldsymbol{\theta}_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)(f_0(t)-f_{\theta_0}(t))d\lambda g(t)dt\\
&\quad -\int_0^1\int_0^1 \ddot{f}_{\boldsymbol{\theta}_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)(f_{\boldsymbol{\theta}_0}(t)-f_{\boldsymbol{\theta}}(t))d\lambda g(t)dt\\
&\quad -\int_0^1\int_0^1 \ddot{f}_{\boldsymbol{\theta}_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)(f(t)-f_0(t))d\lambda g(t)dt\\
&\quad -\int_0^1\int_0^1 \left(\dot{f}_{\theta}(t)-\dot{f}_{\theta_0}(t)\right)^T \dot{f}_{\theta_0+\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}(t)d\lambda g(t)dt.
\end{aligned}$$

We also define

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

where $\epsilon_n\to 0$. By Lemmas 3.13 and 3.14, there exists a sequence $\{\epsilon_n\}$ such that $\Pi_n^*(E_n^c|\boldsymbol{X},\boldsymbol{Y},\sigma^2)=o_{P_0}(1)$. Then, $\boldsymbol{M}(f,\boldsymbol{\theta})$ is invertible and the eigenvalues of $[\boldsymbol{M}(f,\boldsymbol{\theta})]^{-1}$ are bounded away from 0 and ∞ for sufficiently large n and $\left\|(\boldsymbol{M}(f,\boldsymbol{\theta}))^{-1}-(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\right\|=o(1)$ for $(f,\boldsymbol{\theta})\in E_n$. Using (3.3), on E_n

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

that is,

$$\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\sqrt{n}\left(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0)\right)=o(1)\sqrt{n}\left(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0)\right)+O(\sqrt{n}r_n^{-1}). \tag{3.30}$$

Note that

$$\sqrt{n}\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}\left(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0)\right)=\sqrt{n}\boldsymbol{H}_n^T\boldsymbol{\beta}-\sqrt{n}\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{\Gamma}(f_0).$$

It was shown in the proof of Theorem 3.3 that for a given σ^2, the total variation distance between the posterior distribution of $\sqrt{n}\boldsymbol{H}_n^T\boldsymbol{\beta} - \sqrt{n}\,(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\,\boldsymbol{\Gamma}(f_0)$ and $N(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Sigma}_n)$ converges in P_0-probability to 0. By Lemma 3.16, both $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are stochastically bounded. Thus the posterior distribution of $\sqrt{n}\,(\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0))$ assigns most of its mass inside a large compact set with high true probability. Hence, for increasing sample size with high true probability approaching one, the right hand side of (3.30) is arbitrarily small with high posterior probability. □

The next lemma describes the asymptotic behavior of the mean and variance of the limiting normal distribution given by Theorem 3.3.

Lemma 3.16. *The mean and variance of the limiting normal approximation given by Theorem 3.3 are stochastically bounded.*

Proof. First we study the asymptotic behavior of the matrix

$$\mathbf{Var}(\boldsymbol{\mu}_n|\boldsymbol{X}) = \boldsymbol{\Sigma}_n = n\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{H}_n.$$

If $C_k(\cdot) \in C^{m^*}(0,1)$ for some $1 \leq m^* < m$, then by equation (2) of De Boor (1978, page 167), we have for all $k = 1,\ldots,p$, $\sup\{|C_k(t) - \tilde{C}_k(t)| : t \in [0,1]\} = O(k_n^{-1})$, where $\tilde{C}_k(\cdot) = \boldsymbol{\alpha}_k^T\boldsymbol{N}(\cdot)$ and $\boldsymbol{\alpha}_k^T = (C_k(t_1^*),\ldots,C_k(t_{k_n+m-1}^*))$ with appropriately chosen $t_1^*,\ldots,t_{r_n+m-1}^*$. We can write $\boldsymbol{H}_n^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\boldsymbol{H}_n$ as

$$\begin{aligned}(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n) + \tilde{\boldsymbol{H}}_n^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\tilde{\boldsymbol{H}}_n\\ +(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\tilde{\boldsymbol{H}}_n + \tilde{\boldsymbol{H}}_n^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n),\end{aligned}$$

where the k^{th} row of $\tilde{\boldsymbol{H}}_n^T$ is given by $\int_0^1 \tilde{C}_k(t)(\boldsymbol{N}(t))^T g(t)dt$ for $k = 1,\ldots,p$. Let us denote $\boldsymbol{A} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_p)$. Then $\tilde{\boldsymbol{H}}_n^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\tilde{\boldsymbol{H}}_n$ can be expressed as

$$n^{-1}\boldsymbol{A}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\left(\frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{n}\right)^{-1}\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A}.$$

We show that the difference between the matrices

$$\boldsymbol{A}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\left(\frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{n}\right)^{-1}\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A} \tag{3.31}$$

and $\boldsymbol{A}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A}$ converges in P_0-probability to the null matrix of order p. For a $\boldsymbol{l}\in\mathbb{R}^p$, let $\boldsymbol{c}=\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A}\boldsymbol{l}$. Then we can write the quadratic form in 3.31 as

$$\boldsymbol{l}^T\boldsymbol{A}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\left(\frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{n}\right)^{-1}\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A}\boldsymbol{l}$$

as $\boldsymbol{c}^T\left(\frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{n}\right)^{-1}\boldsymbol{c}$. Following the proof of Theorem A.1, it can be shown that

$$\begin{aligned}&\left|\boldsymbol{c}^T\left(\frac{\boldsymbol{X}_n^T\boldsymbol{X}_n}{n}\right)\boldsymbol{c}-\boldsymbol{c}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{c}\right|\\
&\quad\le\ O_{P_0}(n^{-1/2})\boldsymbol{c}^T\boldsymbol{c}\\
&\quad=\ O_{P_0}(n^{-1/2}k_n)\boldsymbol{c}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{c}.\end{aligned}$$

Therefore,

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

Now note that the $(i,j)^{th}$ element of the $p\times p$ matrix $\boldsymbol{A}^T\left(\int_0^1 \boldsymbol{N}(t)\boldsymbol{N}^T(t)g(t)dt\right)\boldsymbol{A}$ 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 corresponding $(i,j)^{th}$ element of the matrix $\int_0^1 \boldsymbol{C}(t)\boldsymbol{C}^T(t)g(t)dt$ which can be also expressed as $(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\int_0^1\left(\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)\right)^T\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t)g(t)dt\left((\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\right)^T$. 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 k^{th} diagonal entry of the matrix $(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)$ is given by

$$\begin{aligned}
&\int_0^1 (C_k(t) - \tilde{C}_k(t))(\boldsymbol{N}(t))^T g(t)dt\,(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\int_0^1 (C_k(t) - \tilde{C}_k(t))(\boldsymbol{N}(t))g(t)dt \\
&= \frac{1}{n}\int_0^1 (C_k(t) - \tilde{C}_k(t))(\boldsymbol{N}(t))^T g(t)dt\left(\boldsymbol{X}_n^T\boldsymbol{X}_n/n\right)^{-1}\int_0^1 (C_k(t) - \tilde{C}_k(t))\boldsymbol{N}(t)g(t)dt \\
&\asymp \frac{k_n}{n}\int_0^1 (C_k(t) - \tilde{C}_k(t))(\boldsymbol{N}(t))^T g(t)dt\int_0^1 (C_k(t) - \tilde{C}_k(t))\boldsymbol{N}(t)g(t)dt \\
&\lesssim \frac{1}{nk_n},
\end{aligned}$$

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

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

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

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

Using the Cauchy-Schwarz inequality and (3.23), we get

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

Thus, $Z_n := \|\mathbf{E}(\boldsymbol{\mu}_n|\boldsymbol{X})\|^2 + \mathrm{maxeig}(\mathrm{Var}(\boldsymbol{\mu}_n|\boldsymbol{X}))$ 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 $M^{-2}\mathrm{E}_0\left[\mathrm{E}\left(\|\boldsymbol{\mu}_n\|^2|\boldsymbol{X}\right)\mathbb{1}\{Z_n \le L\}\right] + P_0(Z_n > L)$ 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 3.17. *For all $\epsilon > 0$, we have $\Pi_n^*(|\sigma^2 - \sigma_0^2| > \epsilon | \boldsymbol{X}, \boldsymbol{Y}) \xrightarrow{P_0} 0$.*

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

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

which implies the posterior distribution of σ^2 is inverse gamma with shape and scale parameters $n/2 + a$ and $2^{-1}\left\{\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n + k_n n^{-2}\boldsymbol{I}_{k_n+m-1})^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}\right\} + b$ respectively. Hence, the posterior mean of σ^2 is given by

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

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

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

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

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

First we prove that $\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}/n \xrightarrow{P_0} \sigma_0^2$ and then $n^{-1}f_0(\boldsymbol{X})^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})f_0(\boldsymbol{X}) \xrightarrow{P_0} 0$ and $n^{-1}\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})f_0(\boldsymbol{X}) \xrightarrow{P_0} 0$. Now, $\mathrm{E}_0\left(\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}/n\right) \to \sigma_0^2$ as $n \to \infty$. Also,

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

Now $\mathrm{Var}(\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}|\boldsymbol{X}) = (\mu_4 - \sigma_0^2)(n - k_n - m + 1)$ and $\mathrm{E}(\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma_0^2(n - k_n - m + 1)$, μ_4 being the fourth order central moment of ε_i for $i = 1, \ldots, n$. Hence it follows that $\mathrm{Var}_0\left(\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}/n\right) \to 0$ as $n \to \infty$. Thus, $\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{\varepsilon}/n \stackrel{P_0}{\to} \sigma_0^2$. We can write for $\boldsymbol{\beta}^*$ satisfying (3.23)

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

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

$$\left|n^{-1}\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})f_0(\boldsymbol{X})\right| = \left|n^{-1}\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})(f_0(\boldsymbol{X}) - \boldsymbol{X}_n\boldsymbol{\beta}^*)\right| \leq \sqrt{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/n}k_n^{-m}$$

which is $o_{P_0}(1)$. Note that using Theorem A.9, $\boldsymbol{P}_{\boldsymbol{X}_n} - \boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n + k_nn^{-2}\boldsymbol{I}_{k_n+m-1})^{-1}\boldsymbol{X}_n^T$ can be written as

$$k_nn^{-2}\boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\left(\boldsymbol{I}_{k_n+m-1} + (\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}k_nn^{-2}\right)^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T$$

whose eigenvalues are of the order $k_nn^{-2}nk_n^{-1}k_n^2n^{-2} = k_n^2n^{-3}$. Hence, the random variable

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

converges in P_0-probability to 0 and $\mathrm{E}(\sigma^2|\boldsymbol{X}, \boldsymbol{Y}) \stackrel{P_0}{\to} \sigma_0^2$. Also,

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

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

Chapter 4

Bayesian inference for higher order ordinary differential equation models

4.1 Introduction

Suppose we have a q^{th} order ODE given by

$$F\left(t, f_{\boldsymbol{\theta}}(t), f_{\boldsymbol{\theta}}^{(1)}(t), \ldots, f_{\boldsymbol{\theta}}^{(q)}(t), \boldsymbol{\theta}\right) = 0, \tag{4.1}$$

F being a known sufficiently smooth real valued function in its arguments and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. We are interested in inferring on $\boldsymbol{\theta}$ based on the model $Y = f_{\boldsymbol{\theta}}(x) + \boldsymbol{\varepsilon}$. For example, the equation describing the concentration of glucose in blood is

$$\frac{d^2g(t)}{dt^2} + (\sin t + \theta_1)\frac{dg(t)}{dt} + g(t)(\cos t + \theta_1 \sin t + \theta_2) = \theta_1 t + 1,$$

where $g(t)$ denotes the glucose concentration at time t. Here θ_1 and θ_2 are the parameters of interest. Bergstrom (1983) estimated the parameters involved in higher order stochastic differential equations by maximum likelihood estimation (MLE). The Gaussian distribution or a frequency domain approximation to Gaussian distribution is used as the working model. He proved the asymptotic normality as well as asymptotic efficiency of the estimator. An efficient algorithm was given in Bergstrom (1985) to compute the

Gaussian likelihood for estimating the parameters involved in a non-stationary higher order ODE. Appropriate linear transformations are used in this algorithm to avoid the computation of the covariance matrix of the observations. Bergstrom (1986) extended this idea for open higher order ODE's.

In this chapter we develop three approaches of Bayesian inference on $\boldsymbol{\theta}$. In our first approach we use Runge-Kutta method to obtain an approximate solution $f_{\boldsymbol{\theta},r_n}(\cdot)$ using r_n grid points and then construct an approximate likelihood and obtain the posterior distribution of $\boldsymbol{\theta}$, using the prior of $\boldsymbol{\theta}$. In another approach we assign prior on the coefficient vector $\boldsymbol{\beta}$ of the B-spline approximation $f(\cdot,\boldsymbol{\beta})$ of the regression function. We define $\boldsymbol{\theta}$ as $\arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} \|f(\cdot,\boldsymbol{\beta}) - f_{\boldsymbol{\eta},r_n}(\cdot)\|_w$ and induce a posterior distribution of $\boldsymbol{\theta}$ using the posterior distribution of $\boldsymbol{\beta}$. The third approach is a generalization of the two-step approach. We use the B-spline approximation of the regression function and define $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} \left\|F\left(\cdot, f(\cdot,\boldsymbol{\beta}), f^{(1)}(\cdot,\boldsymbol{\beta}),\ldots,f^{(q)}(\cdot,\boldsymbol{\beta}),\boldsymbol{\eta}\right)\right\|_w$, where the weight function has to meet some criteria. For the sake of simplicity we have assumed the regression function to be one dimensional. Extension to the multidimensional case is straightforward.

The rest of the chapter is organized as follows. Sections 4.2 contains some preliminaries of Runge-Kutta method. The model assumptions and prior specifications are given in Section 4.3. Section 4.4 contains the descriptions of the estimation methods used. The main results are given in Section 4.5. In Section 4.6 we have carried out a simulation study. Proofs of the main results are given in Section 4.7.

4.2 Preliminaries of Runge-Kutta method for higher order ODE

Often the differential equation has the form

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

with initial conditions $f_{\boldsymbol{\theta}}^{(\nu)}(0) = c_\nu$ for $\nu = 0,\ldots,q-1$, H being known. Note that t can be treated as a state variable $\chi(t) = t$ with initial conditions $\chi(0) = 0,\ \chi^{(1)}(0) = 1$ and

$\chi^{(j)}(0) = 0$ for $j = 2, \ldots, q-1$. Denoting $\boldsymbol{\psi}_{\boldsymbol{\theta}}(\cdot) = (f_{\boldsymbol{\theta}}(\cdot), \chi(\cdot))$, we can rewrite the ODE as

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

$\boldsymbol{H}$ being known. Given r equispaced grid points $a_1 = 0, a_2, \ldots, a_r$ with common difference h, the approximate solution to (4.1) is given by $\boldsymbol{\psi}_{\theta,r}(\cdot) = (f_{\boldsymbol{\theta},r}(\cdot), \chi_r(\cdot))$, where $r = r_n$ is the number of grid points, which is chosen so that $r_n \gg \sqrt{n}$. Let us denote

$$\boldsymbol{z}_k = (\boldsymbol{\psi}_{\boldsymbol{\theta},r_n}(a_k), \boldsymbol{\psi}_{\boldsymbol{\theta},r_n}^{(1)}(a_k), \ldots, \boldsymbol{\psi}_{\boldsymbol{\theta},r_n}^{(q-1)}(a_k)).$$

The updated solution is given by $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + r_n^{-1}\boldsymbol{\Phi}(a_k, \boldsymbol{z}_k, r_n)$, where

$$\boldsymbol{\Phi}(t, \boldsymbol{z}, r_n) = \left(\boldsymbol{\Phi}^1(t, \boldsymbol{z}, r_n), \ldots, \boldsymbol{\Phi}^q(t, \boldsymbol{z}, r_n)\right)^T$$

for $k = 0, \ldots, r_n - 1$. By equation (4.16) of Henrici (1962, page 169) we know that

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

where for $\nu = 1, 2, \ldots, q-1$

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

Now

$$\boldsymbol{k}_\rho := \boldsymbol{H}(\boldsymbol{U}^1, \boldsymbol{U}^2, \ldots, \boldsymbol{U}^q) \tag{4.2}$$

with $\boldsymbol{U}^1, \boldsymbol{U}^2, \ldots, \boldsymbol{U}^q$ being given in Table 4.1. For $\nu = 1, \ldots, q$ and $\rho = 1, \ldots, 4$, the

Table 4.1: *Arguments of $\boldsymbol{H}$ in* (4.2)

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

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

$$\gamma_{\nu 1} = \frac{(q-\nu+1)^2}{(q-\nu+2)(q-\nu+3)},$$
$$\gamma_{\nu 2} = \gamma_{\nu 3} = \frac{1(q-\nu+1)}{(q-\nu+2)(q-\nu+3)},$$
$$\gamma_{\nu 4} = \frac{1-q+\nu}{(q-\nu+2)(q-\nu+3)}.$$

By the proof of Theorem 4.2 of Henrici (1962, page 174), we have

$$\sup_{t\in[0,1]} |f_{\boldsymbol{\theta}}(t) - f_{\boldsymbol{\theta},r_n}(t)| = O(r_n^{-1}),\ \sup_{t\in[0,1]} \left\| \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(t) - \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta},r_n}(t) \right\| = O(r_n^{-1}). \tag{4.3}$$

4.3 Model description and prior specification

Now we formally describe the model. The proposed model is given by

$$Y_i = f_{\boldsymbol{\theta}}(X_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{4.4}$$

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

$$F\left(t, f_{\boldsymbol{\theta}}(t), f_{\boldsymbol{\theta}}^{(1)}(t), \ldots, f_{\boldsymbol{\theta}}^{(q)}(t), \boldsymbol{\theta}\right) = 0. \tag{4.5}$$

Let for a fixed $\boldsymbol{\theta}$, $F \in C^{m-1}((0,1), \mathbb{R}^{q+1})$ 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 $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. We assume that $f_0 \in C^m([0,1])$. Let ε_i be independently and identically 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 $X_i \overset{iid}{\sim} G$ with density g.

Let us denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{X} = (X_1, \ldots, X_n)^T$. The true joint distribution of (X_i, ε_i) is denoted by P_0.

4.4 Methodology

Now we describe the three different approaches of inference on $\boldsymbol{\theta}$ used in this paper.

4.4.1 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

$$p_{\boldsymbol{\gamma},n}(X_i, Y_i) = (\sqrt{2\pi}\sigma)^{-1} \exp\{-(2\sigma^2)^{-1}|Y_i - f_{\boldsymbol{\theta},r_n}(X_i)|^2\}g(X_i). \tag{4.6}$$

We also denote

$$p_{\boldsymbol{\gamma}}(X_i, Y_i) = (\sqrt{2\pi}\sigma)^{-1} \exp\{-(2\sigma^2)^{-1}|Y_i - f_{\boldsymbol{\theta}}(X_i)|^2\}g(X_i). \tag{4.7}$$

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 natural requirement that if errors are normally distributed and $f_{\boldsymbol{\theta}_0}$ is the true regression function, then $\boldsymbol{\gamma}_0 = (\boldsymbol{\theta}_0, \sigma_0^2)$, where $\boldsymbol{\sigma}_0^2$ is the true value of the error variance. We denote by $\ell_{\boldsymbol{\gamma}}$ and $\ell_{\boldsymbol{\gamma},n}$ the log-likelihoods with respect to (4.7) and (4.6) respectively. If $\boldsymbol{\gamma}_0$ is the unique maximizer of the right hand side above, we get

$$\int_0^1 \dot{f}_{\boldsymbol{\theta}_0}^T(t)\left(f_0(t) - f_{\boldsymbol{\theta}_0}(t)\right) g(t)dt = \mathbf{0},\ \sigma_*^2 = \sigma_0^2 + \int_0^1 |f_0(t) - f_{\boldsymbol{\theta}_0}(t)|^2 g(t)dt. \tag{4.8}$$

We assume that the sub-matrix of the Hessian matrix of $-P_0 \log p_{\boldsymbol{\gamma}}$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ given by

$$\int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}^T(t)\dot{f}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial \boldsymbol{\theta}} \left(\dot{f}_{\boldsymbol{\theta}}^T(t)\left(f_0(t) - f_{\boldsymbol{\theta}_0}(t)\right)\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) g(t)dt \tag{4.9}$$

is positive definite. The prior measure on $\boldsymbol{\Theta}$ is assumed to have a Lebesgue-density continuous and positive on a neighborhood of $\boldsymbol{\theta}_0$. The prior distribution of $\boldsymbol{\theta}$ is assumed to be independent of that of σ^2. The joint prior measure is denoted by Π with corresponding density π. We obtain the posterior of $\boldsymbol{\gamma}$ using the approximate likelihood given by (4.6).

4.4.2 Runge-Kutta Two-step Bayesian Method (RKTB)

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

$$\boldsymbol{Y} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.10}$$

where $\boldsymbol{X}_n = (\!(N_j(X_i))\!)_{1\leq i\leq n, 1\leq j\leq k_n+m-1}$, $\{N_j(\cdot)\}_{j=1}^{k_n+m-1}$ being the B-spline basis functions of order m with $k_n - 1$ interior knots. We assume for a given σ^2

$$\boldsymbol{\beta} \sim N_{k_n+m-1}(\boldsymbol{0}, \sigma^2 n^2 k_n^{-1} \boldsymbol{I}_{k_n+m-1}). \tag{4.11}$$

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

$$N_{k_n+m-1}\left(\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}, \sigma^2\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + n^{-2}k_n\boldsymbol{I}_{k_n+m-1}\right)^{-1}\right). \tag{4.12}$$

By model (4.10), the expected response at a point $t \in [0,1]$ is given by $\boldsymbol{\beta}^T\boldsymbol{N}(t)$, where $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_{k_n+m-1}(\cdot))^T$. Let us denote for a given parameter $\boldsymbol{\eta}$

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

where $f(t) = \boldsymbol{\beta}^T\boldsymbol{N}(t)$. Now we define $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} R_{f,n}(\boldsymbol{\eta})$ and induce posterior distribution on $\boldsymbol{\Theta}$ through the posterior of $\boldsymbol{\beta}$ given by (4.12). Also let us define $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} R_{f_0}(\boldsymbol{\eta})$. 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 assume that

$$\text{for all}\, \epsilon > 0, \quad \inf_{\boldsymbol{\eta}: \|\boldsymbol{\eta}-\boldsymbol{\theta}_0\| \geq \epsilon} R_{f_0}(\boldsymbol{\eta}) > R_{f_0}(\boldsymbol{\theta}_0). \tag{4.13}$$

4.4.3 Two-step Bayesian Method

Here we use the same nonparametric model and prior specifications as in RKTB, but $\boldsymbol{\theta}$ is defined as

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

where $w(\cdot)$ is a nonnegative sufficiently smooth weight function, whose derivatives up to $(q-1)^{th}$ order vanish at 0 and 1. We also assume that for all $\epsilon > 0$

$$\begin{aligned} \inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} & \left\| F\left(\cdot, f_0(\cdot), f_0^{(1)}(\cdot), \ldots, f_0^{(q)}(\cdot), \boldsymbol{\eta}\right)\right\|_w \\ & > \left\| F\left(\cdot, f_0(\cdot), f_0^{(1)}(\cdot), \ldots, f_0^{(q)}(\cdot), \boldsymbol{\theta}_0\right)\right\|_w . \end{aligned} \tag{4.14}$$

4.5 Main results

The main results of our work are given by Theorems 4.1, 4.3 and 4.7.

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*

$$\left\|\Pi_n\left(\sqrt{n}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0) \in \cdot|\boldsymbol{X},\boldsymbol{Y}\right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}, \sigma_*^2 \boldsymbol{V}_{\boldsymbol{\gamma}_0}^{-1})\right\|_{TV} \xrightarrow{P_0} 0$$

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

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

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

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 $\sigma_*^2\boldsymbol{V}_{\boldsymbol{\gamma}_0}^{-1}$ respectively. We also get the following important corollary.

Corollary 4.2. *When the regression model* (4.4) *is correctly specified and also the error is Gaussian, the Bayes estimator based on* Π_n *is asymptotically efficient.*

In RKTB we assume that the matrix

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

is nonsingular. Let us denote $\boldsymbol{C}(t) = (\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\left(\dot{f}_{\boldsymbol{\theta}_0}(t)\right)^T$ and $\boldsymbol{G}_n^T = \int_0^1 \boldsymbol{C}(t)\boldsymbol{N}^T(t)g(t)dt$. Also, we denote the posterior probability measure of RKTB by Π_n^*. Now we have the following result.

Theorem 4.3. *Let*

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

If $\boldsymbol{B}$ *is non-singular, then for* $m \geq 3$ *and* $n^{1/(2m)} \ll k_n \ll n^{1/4}$,

$$\left\|\Pi_n^*\left(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot|Y\right) - N\left(\boldsymbol{\mu}_n, \sigma_0^2\boldsymbol{\Sigma}_n\right)\right\|_{TV} = o_{P_0}(1). \tag{4.15}$$

Remark 4.4. Following the steps of Lemma 3.16 it can be proved that both $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are stochastically bounded. Hence, with high true probability the posterior distribution of $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ contracts at $\boldsymbol{0}$ at $n^{-1/2}$ rate.

We also get the following important corollary.

Corollary 4.5. *When the regression model* (4.4) *is correctly specified and the true distribution of error is Gaussian, the Bayes estimator based on* Π_n^* *is asymptotically efficient.*

We denote $\boldsymbol{h} = (f, f^{(1)}, \ldots, f^{(q)})^T$ and $\boldsymbol{h}_0$ to be similar to $\boldsymbol{h}$ with f being replaced by f_0. For a function $\phi(\boldsymbol{t}, \boldsymbol{h}(t), \boldsymbol{\theta})$ let us denote $D_{\boldsymbol{h}}\phi = \frac{\partial}{\partial \boldsymbol{h}}\phi$ and $D_{\boldsymbol{\theta}}\phi = \frac{\partial}{\partial \boldsymbol{\theta}}\phi$ respectively. We denote $\boldsymbol{G}(t, \boldsymbol{h}(t), \boldsymbol{\theta}) = (D_{\boldsymbol{\theta}}F(t, \boldsymbol{h}(t), \boldsymbol{\theta}))^T F(t, \boldsymbol{h}(t), \boldsymbol{\theta})$.

Theorem 4.6. *We denote* $\boldsymbol{G}(t,\boldsymbol{h}(t),\boldsymbol{\theta}) = (D_{\boldsymbol{\theta}}F(t,\boldsymbol{h}(t),\boldsymbol{\theta}))^T F(t,\boldsymbol{h}(t),\boldsymbol{\theta})$*. Let the matrix*

$$\boldsymbol{M}(\boldsymbol{h}_0,\boldsymbol{\theta}_0) = \int_0^1 D_{\boldsymbol{\theta}}\left(\boldsymbol{G}(t,\boldsymbol{h}_0(t),\boldsymbol{\theta}_0)\right) w(t)dt$$

be nonsingular. For $m > (2q+2)$*,* $n^{1/2m} \ll k_n \ll n^{1/(4q+4)}$ *and the assumption (4.14), there exists* $E_n \subseteq \boldsymbol{\Theta} \times C^m((0,1))$ *with* $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$*, such that for* $(\boldsymbol{\theta},\boldsymbol{h}) \in E_n$*,*

$$\left\|\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - (\boldsymbol{M}(\boldsymbol{h}_0,\boldsymbol{\theta}_0))^{-1}\sqrt{n}(\boldsymbol{\Gamma}(f)-\boldsymbol{\Gamma}(f_0))\right\| \to 0, \tag{4.16}$$

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

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

$$(\boldsymbol{M}(\boldsymbol{h}_0,\boldsymbol{\theta}_0))^{-1}\boldsymbol{\Gamma}(f) = \int_0^1 \boldsymbol{A}(t)\boldsymbol{\beta}^T\boldsymbol{N}(t)d\boldsymbol{t} = \boldsymbol{H}_n^T\boldsymbol{\beta}, \tag{4.17}$$

where $\boldsymbol{H}_n^T = \int_0^1 \boldsymbol{A}(t)\boldsymbol{N}^T(t)dt$ which is a matrix of order $p\times(k_n+m-1)$. Then in order to approximate the posterior distribution of $\boldsymbol{\theta}$, it suffices to study the asymptotic posterior distribution of the linear functional of $\boldsymbol{\beta}$ given by (4.17). The next theorem describes the approximate posterior distribution of $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$.

Theorem 4.7. *Let us denote*

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

and $\boldsymbol{D} = \left(\left(\langle A_k(\cdot),A_{k'}(\cdot)\rangle_g\right)\right)_{k,k'=1,\ldots,p}$*. If* $\boldsymbol{D}$ *is non-singular, then under the conditions of Theorem 4.6,*

$$\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right) - N\left(\boldsymbol{\mu}_n^*,\sigma_0^2\boldsymbol{\Sigma}_n^*\right)\right\|_{TV} = o_{P_0}(1). \tag{4.18}$$

4.6 Simulation Study

We consider the van der Pol equation

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

with initial condition $f_\theta(0) = 2, f'_\theta(0) = 0$ to study the posterior distribution of θ. We consider the situation where the true regression function belongs to the solution set. 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 1000 replications for each case. Under each replication a sample of size 1000 is drawn from the posterior distribution of $\boldsymbol{\theta}$ using RKSB, RKTB and Bayesian two-step methods and then 95% equal tailed credible intervals are obtained. Bayesian two-step method is abbreviated as "TS" in the table. We calculate the coverage and the average length of the corresponding credible interval over these 1000 replications. The estimated standard errors of the interval length and coverage are given inside the parentheses in the table. The true parameter vector is chosen as $\theta_0 = 1$. The above system is not analytically solvable. 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. We take $m = 5$ and $m = 7$ for RKTB and Bayesian two-step method respectively. Looking at the order of k_n suggested by Theorem 4.3, k_n is chosen as 13 and 18 for $n = 100$ and $n = 500$ respectively in RKTB. In Bayesian two-step method the choices are 2 and 3 for $n = 100$ and $n = 500$ respectively. The simulation results are summarized in the Table 4.2. Not surprisingly the first two methods performs much better compared to the third one because of asymptotic efficiency obtained from Corollaries 4.2 and 4.5 respectively.

Table 4.2: *Coverages and average lengths of the Bayesian credible intervals for the three methods*

n		RKTB		RKSB		TS	
		coverage (se)	length (se)	coverage (se)	length (se)	coverage (se)	length (se)
100	θ	99.9 (0.00)	1.02 (0.05)	100.0 (0.00)	0.97 (0.02)	99.9 (0.01)	4.02 (0.88)
500	θ	100.0 (0.00)	0.24 (0.00)	99.9 (0.00)	0.24 (0.00)	99.9 (0.00)	1.22 (0.18)

4.7 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. As in Lemma 3.7 we can argue that there exists a compact subset U of $(0,\infty)$ such that $\Pi_n(\sigma^2 \in U|\boldsymbol{X},\boldsymbol{Y}) \overset{P_0}{\to} 1$. Let $\Pi_{U,n}(\cdot|\boldsymbol{X},\boldsymbol{Y})$ be the posterior distribution conditioned on $\sigma^2 \in U$. By Theorem A.2 if we can ensure that there exist stochastically bounded random variables $\boldsymbol{\Delta}_{n,\gamma_0}$ and a positive definite matrix $\boldsymbol{V}_{\gamma_0}$ such that for every compact set $K \subset \mathbb{R}^{p+1}$,

$$\sup_{h\in K}\left|\log\frac{p^{(n)}_{\boldsymbol{\gamma}_0+\boldsymbol{h}/\sqrt{n},n}}{p^{(n)}_{\boldsymbol{\gamma}_0,n}}(\boldsymbol{X},Y) - \boldsymbol{h}^T\boldsymbol{V}_{\boldsymbol{\gamma}_0}\boldsymbol{\Delta}_{n,\gamma_0} + \frac{1}{2}\boldsymbol{h}^T\boldsymbol{V}_{\boldsymbol{\gamma}_0}\boldsymbol{h}\right| \to 0, \tag{4.19}$$

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

$$P_0^{(n)}\Pi_{U,n}\left(\sqrt{n}\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| > M_n|\boldsymbol{X},Y\right) \to 0, \tag{4.20}$$

then

$$\left\|\Pi_{U,n}\left(\sqrt{n}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\in\cdot|\boldsymbol{X},Y\right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0},\boldsymbol{V}^{-1}_{\boldsymbol{\gamma}_0})\right\|_{TV} \overset{P_0}{\to} 0.$$

To show that the conditions (4.19) and (4.20) hold, we prove results similar to Lemmas 3.7

to 3.11. Following the steps of Lemma 3.8 we get $\boldsymbol{V}_{\boldsymbol{\gamma}_0} = \begin{pmatrix} \sigma_*^{-2}\boldsymbol{V}_{\theta_0} & \mathbf{0} \\ \mathbf{0} & \sigma_*^{-4}/2 \end{pmatrix}$ with $\boldsymbol{V}_{\boldsymbol{\theta}_0} = \int_0^1 \left(\dot{\boldsymbol{f}}^T_{\boldsymbol{\theta}_0}(t)\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(t) - \frac{\partial}{\partial \boldsymbol{\theta}} \left(\dot{\boldsymbol{f}}^T_{\boldsymbol{\theta}}(t) \left(\boldsymbol{f}_0(t) - \boldsymbol{f}_{\boldsymbol{\theta}_0}(t) \right) \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) g(t)dt$ and $\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0} = \boldsymbol{V}^{-1}_{\boldsymbol{\gamma}_0}\mathbb{G}_n\dot{\ell}_{\boldsymbol{\gamma}_0,n}$. We get

$$\left\| \Pi_n \left(\sqrt{n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \in \cdot | \boldsymbol{X}, Y \right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n,\boldsymbol{\gamma}_0}, \boldsymbol{V}^{-1}_{\boldsymbol{\gamma}_0}) \right\|_{TV} \xrightarrow{P_0} 0$$

since $\|\Pi_n - \Pi_{U,n}\|_{TV} = o_{P_0}(1)$ and the result follows. □

Proof of Corollary 4.2. The log-likelihood of the correctly specified model with Gaussian error is given by

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

Thus $\frac{\partial}{\partial \boldsymbol{\theta}_0}\ell_{\boldsymbol{\gamma}_0}(X, Y) = \sigma_0^{-2}\left(\dot{\boldsymbol{f}}_{\boldsymbol{\theta}_0}(X)\right)^T (Y - f_{\boldsymbol{\theta}_0}(X))$ 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

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

Thus $\boldsymbol{V}^{-1}_{\boldsymbol{\gamma}_0} = (\boldsymbol{I}(\boldsymbol{\gamma}_0))^{-1}$ if the regression function is correctly specified and the true error distribution is $N(0, \sigma_0^2)$. □

Proof of Theorem 4.3. We have for $f(\cdot) = \boldsymbol{\beta}^T\boldsymbol{N}(\cdot)$

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

where $\boldsymbol{G}_n^T = \int_0^1 \boldsymbol{C}(t)\boldsymbol{N}^T(t)g(t)dt$ which is a matrix of order $p \times (k_n + m - 1)$. Consequently, the asymptotic variance of the conditional posterior distribution of $\boldsymbol{G}_n^T\boldsymbol{\beta}$ is $\sigma^2\boldsymbol{G}_n^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n + \frac{k_n}{n^2}\boldsymbol{I}_{k_n+m-1}\right)^{-1}\boldsymbol{G}_n$. We can derive the posterior consistency of σ^2 similar to Lemma 3.17. Following result similar to Lemma 3.15, it suffices to show that for any neighborhood $\mathscr{N}$ of σ_0^2,

$$\sup_{\sigma^2\in\mathscr{N}} \left\| \Pi_n^* \left(\sqrt{n} \boldsymbol{G}_n^T \boldsymbol{\beta} - \sqrt{n} \left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1} \int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}(t)\right)^T f_0(t) g(t) \in \cdot | \boldsymbol{X}, \boldsymbol{Y}, \sigma^2 \right) \right. \\ \left. - N(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Sigma}_n) \right\|_{TV}$$

is $o_{P_0}(1)$. The rest of the proof follows from that of Theorem 3.3. □

Proof of Corollary 4.5. The log-likelihood of the correctly specified model is given by

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

Thus $\dot{\ell}_{\boldsymbol{\theta}_0}(X,Y) = -\sigma_0^{-2} \left(\dot{f}_{\boldsymbol{\theta}_0}(X)\right)^T (Y - f_{\boldsymbol{\theta}_0}(X))$ and the Fisher information is given by $\boldsymbol{I}(\boldsymbol{\theta}_0) = \sigma_0^{-2} \int_0^1 \left(\dot{f}_{\boldsymbol{\theta}_0}(X)\right)^T \dot{f}_{\boldsymbol{\theta}_0}(t) g(t) dt$. Following the proof of Lemma 3.16 we get

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

This limit is equal to $\left(\boldsymbol{I}(\boldsymbol{\theta}_0)\right)^{-1}$ under the correct specification of the regression function as well as the likelihood. □

Proof of Theorem 4.6. By definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ we have

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

Subtracting the second equation from the first and applying the Mean-value Theorem, we get

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

Now we will show that $\int_0^1 D_{\boldsymbol{h}_0} \left(\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\right) w(t) (\boldsymbol{h}(t) - \boldsymbol{h}_0(t)) dt$ is a linear functional of $f - f_0$. Note that $\int_0^1 D_{\boldsymbol{h}_0}\left(\boldsymbol{G}(t, \boldsymbol{h}_0(t), \boldsymbol{\theta}_0)\right) w(t)(\boldsymbol{h}(t) - \boldsymbol{h}_0(t))dt$ can be written as

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

We shall show that every term of this sum is a linear functional of $f - f_0$. We observe that

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

Proceeding this way we get

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

We also define

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

where $\epsilon_n \to 0$. Using the steps of the proofs of Lemmas 2.12 and 2.13, we can prove the posterior consistency of $\boldsymbol{\theta}$. Hence, there exists a sequence $\{\epsilon_n\}$ so that $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$. Hence, on E_n

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

By result similar to Lemma 2.14, $\boldsymbol{\Gamma}(f - f_0)$ assigns most of its mass inside a large compact set. Now using result similar to Lemma 2.12, we can finally assert that inside the set E_n, the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is given by that of

$$\sqrt{n}\left((\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0))^{-1}\right) \boldsymbol{\Gamma}(f - f_0) = \sqrt{n}\left((\boldsymbol{M}(\boldsymbol{h}_0, \boldsymbol{\theta}_0))^{-1}\right) (\Gamma(f) - \Gamma(f_0)).$$

□

Proof of Theorem 4.7. By Theorem 4.6 and (4.17), it suffices to show that for any σ^2 in a neighborhood of σ_0^2,

$$\left\|\Pi\left(\sqrt{n}\boldsymbol{H}_n^T\boldsymbol{\beta} - \sqrt{n}\left(\boldsymbol{M}(\boldsymbol{g}_0,\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{\Gamma}(f_0) \in \cdot|\boldsymbol{Y}\right) - N(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Sigma}_n)\right\|_{TV} = o_{P_0}(1). \tag{4.21}$$

Note that the posterior distribution of $\boldsymbol{H}_n^T\boldsymbol{\beta}$ is a normal distribution with mean vector

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

and dispersion matrix

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

respectively. We calculate the Kullback-Leibler divergence between two Gaussian distributions and show that it converges in P_0-probability to zero to prove the assertion. The rest of the proof is similar to that of Theorem 2.7. □

Chapter 5

Bayesian inference for partial differential equation models

5.1 Introduction

In some applications, a multivariate regression model $\boldsymbol{Y} = \boldsymbol{f}_{\boldsymbol{\theta}}(t_1, t_2) + \boldsymbol{\varepsilon}$ is described by partial differential equations (PDE) such as of the form

$$\boldsymbol{F}\left(t_1, t_2, \frac{\partial \boldsymbol{f}_{\boldsymbol{\theta}}(t_1,t_2)}{\partial t_1}, \frac{\partial \boldsymbol{f}_{\boldsymbol{\theta}}(t_1,t_2)}{\partial t_2}, \frac{\partial^2 \boldsymbol{f}_{\boldsymbol{\theta}}(t_1,t_2)}{\partial t_1^2}, \frac{\partial^2 \boldsymbol{f}_{\boldsymbol{\theta}}(t_1,t_2)}{\partial t_2^2}, \frac{\partial^2 \boldsymbol{f}_{\boldsymbol{\theta}}(t_1,t_2)}{\partial t_1 \partial t_2}, \boldsymbol{\theta}\right) = 0,$$

where $\boldsymbol{F}$ is an appropriately smooth vector valued function and $\boldsymbol{\theta}$ is the unknown parameter of interest which has to be estimated from the noisy data. PDE models are often encountered in fields like biology and finance. In most of the situations, these systems are not analytically solvable. Müller and Timmer (2004) applied the multiple shooting approach. In the same paper they also used a two step approach similar to Varah (1982) using splines to estimate the parameters. Xun et al. (2013) used a two-step optimization process called parameter cascading method. In the first step a nonparametric B-spline model is fit using penalized least squares approach for a given parameter vector. The parameter vector is estimated by least squares method in the second step. The asymptotic normality as well as $\sqrt{n}$-consistency of the estimator were established in their work. In the Bayesian framework, the Bayesian P-splines approach has been used by Xun et al.

(2013). But the theoretical aspects of Bayesian estimation for PDE models have not been explored yet. We extend the approach discussed in Chapter 2 for PDE models of a given order and given number of independent variables. We put a prior on the coefficients of the B-spline basis functions and establish a Bernstein-von Mises theorem for the posterior distribution of the parameter with the $n^{-1/2}$ contraction rate.

The rest of the chapter is organized as follows. The model assumptions and prior specifications are given in Section 5.2. The main results are given in Section 5.3. In Section 5.4 we have carried out a simulation study. Proofs of the main results are given in Section 5.5.

5.2 Model description and prior specification

Suppose we have a partial differential equation of order α given by

$$\boldsymbol{F}\left(\boldsymbol{t},\left(D^{\boldsymbol{r}}\boldsymbol{f}_{\boldsymbol{\theta}}:|\boldsymbol{r}|\leq\alpha\right),\boldsymbol{\theta}\right)=0,$$

where $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, which is a compact subset of $\mathbb{R}^p$ and $\boldsymbol{t}=(t_1,\ldots,t_s)^T$. Denoting by Y_i the observation taken at the point $\boldsymbol{x}_i=(x_{i1},\ldots,x_{is})$, $i=1,\ldots,n$, we propose the model

$$\boldsymbol{Y}_i=\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}_i,\ i=1,\ldots,n,$$

whereas the data is generated by the model

$$\boldsymbol{Y}_i=\boldsymbol{f}_0(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}_i,\ i=1,\ldots,n.$$

The true mean $\boldsymbol{f}_0$ does not necessarily lie in the family $\{\boldsymbol{f}_{\boldsymbol{\theta}}:\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$. We assume both $\boldsymbol{F}$ and the regression function to be one dimensional for the sake of simplicity in notation. Extension to multidimensional case is straightforward. Let us assume $\varepsilon_i\overset{iid}{\sim}P_0$, which is a probability distribution with mean zero and finite variance σ_0^2 as well as finite fourth order moment.

Since the expression of f_θ is not usually available, the proposed model is embedded in the nonparametric regression model

$$\boldsymbol{Y} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.1}$$

where $\boldsymbol{X}_n = ((\prod_{k=1}^{s} N_{j_k}(x_{ik})))_{1\le i\le n, 1\le j_1,\ldots,j_s\le k_n+m-1}$, $\{N_{j_k}(\cdot)\}_{j=1}^{k_n+m-1}$ being the B-spline basis functions of order m with k_n-1 interior knots. Also we consider P_0 to be unknown and use $N(0,\sigma^2)$ as the working distribution for the error. 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. Let us denote by $\xi_{1,j},\ldots,\xi_{k_n-1,j}$ the set of interior knots with $\xi_{l,j} = l/k_n$ for $l = 1,\ldots,k_n-1$ and $j = 1,\ldots,s$. Denoting by Q_n, the empirical distribution function of $\boldsymbol{x}_i$, $i = 1,\ldots,n$, we assume that

$$\sup_{t\in[0,1]^s} |Q_n(\boldsymbol{t}) - \prod_{j=1}^{s} t_j| = o(k_n^{-s}),$$

$Q_n(\cdot)$ being the empirical distribution function of $\{\boldsymbol{X}_i, i = 1,\ldots,n\}$ and $Q(\cdot)$ is a distribution function with positive continuous density. We assume

$$\boldsymbol{\beta}|\sigma^2 \overset{iid}{\sim} N_{k_n+m-1}(\boldsymbol{0}, \sigma^2 n k_n^{-s}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}). \tag{5.2}$$

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

$$\boldsymbol{\beta}|\boldsymbol{Y},\sigma^2 \sim N_{k_n+m-1}\left(c_n^{-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}, c_n^{-1}\sigma^2(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\right), \tag{5.3}$$

where $c_n = (1 + k_n^s/n)$. By model (5.1), the expected response vector at a point $\boldsymbol{t}\in[0,1]^s$ is given by $\boldsymbol{\beta}^T\boldsymbol{N}(\boldsymbol{t})$, where $\boldsymbol{N}(\cdot) = \{\prod_{k=1}^{s} N_{j_k}(\cdot)\}_{1\le j_1,\ldots,j_s\le k_n+m-1}$.

Let $w(\cdot)$ be a weight function whose mixed partial derivatives up to $(\alpha-1)^{th}$ order vanish at 0 and 1. We define

$$\begin{aligned} R_f(\boldsymbol{\eta}) &= \left\| F\left(\boldsymbol{t}, (D^{\boldsymbol{r}} f : |\boldsymbol{r}| \le \alpha), \boldsymbol{\theta}\right)\right\|_w, \\ \psi(f) &= \arg\min_{\boldsymbol{\eta}\in\boldsymbol{\Theta}} R_f(\boldsymbol{\eta}). \end{aligned}$$

It is easy to check that $\boldsymbol{\psi}(f_{\boldsymbol{\eta}}) = \boldsymbol{\eta}$ for all $\boldsymbol{\eta}\in\Theta$. Thus the map $\boldsymbol{\psi}$ extends the definition of the parameter $\boldsymbol{\theta}$ beyond the model. Let us define $\boldsymbol{\theta}_0 = \boldsymbol{\psi}(f_0)$. We assume that $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}$. From now on, we shall write $\boldsymbol{\theta}$ for $\boldsymbol{\psi}(f)$ and treat it as the parameter of interest. A posterior is induced on $\boldsymbol{\Theta}$ through the mapping $\boldsymbol{\psi}$ acting on $f(\cdot) = \boldsymbol{\beta}^T \boldsymbol{N}(\cdot)$ and the posterior of $\boldsymbol{\beta}$ is given by (5.3).

5.3 Results

Our objective is to study the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. The main result of this chapter is given by Theorem 5.2. Before that the asymptotic representation of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is given by the next theorem under the assumption that

$$\text{for all}\, \epsilon > 0, \quad \inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}_0\|\geq\epsilon} R_{\boldsymbol{f}_0}(\boldsymbol{\eta}) > R_{\boldsymbol{f}_0}(\boldsymbol{\theta}_0). \tag{5.4}$$

The above condition implies that $\boldsymbol{\theta}_0$ is the unique point of minimum of $R_{\boldsymbol{f}_0}(\cdot)$ and $\boldsymbol{\theta}_0$ should be a well-separated point of minimum. We denote $\boldsymbol{g} = (D^{\boldsymbol{r}} f)^T$ and $\boldsymbol{g}_0 = (D^{\boldsymbol{r}} f_0)^T$ with $|\boldsymbol{r}| \leq \alpha$. For a function $\phi(\boldsymbol{t}, \boldsymbol{g}(\boldsymbol{t}), \boldsymbol{\theta})$ let us denote $D_{\boldsymbol{g}}\phi = \frac{\partial}{\partial \boldsymbol{g}}\phi$ and $D_{\boldsymbol{\theta}}\phi = \frac{\partial}{\partial \boldsymbol{\theta}}\phi$ respectively.

Theorem 5.1. *We denote* $\boldsymbol{G}(\boldsymbol{t}, \boldsymbol{g}(\boldsymbol{t}), \boldsymbol{\theta}) = \left(D_{\boldsymbol{\theta}} F(\boldsymbol{t}, \boldsymbol{g}(\boldsymbol{t}), \boldsymbol{\theta})\right)^T F(\boldsymbol{t}, \boldsymbol{g}(\boldsymbol{t}), \boldsymbol{\theta})$. *Let the matrix*

$$\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0) = \int_0^1 D_{\boldsymbol{\theta}} \left(\boldsymbol{G}(\boldsymbol{t}, \boldsymbol{g}_0(\boldsymbol{t}), \boldsymbol{\theta}_0)\right) w(\boldsymbol{t}) d\boldsymbol{t}$$

be nonsingular. For of $m > (2\alpha + s)$ *and* $n^{1/2m} \ll k_n \ll n^{1/(4\alpha+2s)}$ *and the assumption* (5.4), *there exists* $E_n \subseteq \boldsymbol{\Theta} \times C^m((0,1))$ *with* $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$, *such that for* $(\boldsymbol{\theta}, \boldsymbol{g}) \in E_n$,

$$\left\| \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \left(\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0)\right)^{-1} \sqrt{n}(\boldsymbol{\Gamma}(f) - \boldsymbol{\Gamma}(f_0)) \right\| \to 0, \tag{5.5}$$

where $\boldsymbol{\Gamma}(f) = -\sum_{|\boldsymbol{r}|\leq\alpha} \int_0^1 (-1)^{\sum_{j=1}^s r_j} D^{\boldsymbol{r}} \left\{D_{\boldsymbol{g}_0,\boldsymbol{r}} \left(\boldsymbol{G}(\boldsymbol{t}, \boldsymbol{g}_0(\boldsymbol{t}), \boldsymbol{\theta}_0)\right) w(\boldsymbol{t})\right\} f(\boldsymbol{t}) d\boldsymbol{t}$.

Denoting

$$-\left(\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0)\right)^{-1} \sum_{|\boldsymbol{r}|\leq\alpha} (-1)^{\sum_{j=1}^s r_j} D^{\boldsymbol{r}} \left\{D_{\boldsymbol{g}_0,\boldsymbol{r}} \left(\boldsymbol{G}(\boldsymbol{t}, \boldsymbol{g}_0(\boldsymbol{t}), \boldsymbol{\theta}_0)\right) w(\boldsymbol{t})\right\} = \boldsymbol{A}(\boldsymbol{t}),$$

we have

$$(\boldsymbol{M}(\boldsymbol{g}_0,\boldsymbol{\theta}_0))^{-1}\,\boldsymbol{\Gamma}(f) \;=\; \int_0^1 \boldsymbol{A}(\boldsymbol{t})\boldsymbol{\beta}^T\boldsymbol{N}(\boldsymbol{t})d\boldsymbol{t} = \boldsymbol{H}_n^T\boldsymbol{\beta}, \tag{5.6}$$

where $\boldsymbol{H}_n^T = \int_0^1 \boldsymbol{A}(\boldsymbol{t})\boldsymbol{N}^T(\boldsymbol{t})d\boldsymbol{t}$ which is a matrix of order $p\times(k_n+m-1)^s$. Then in order to approximate the posterior distribution of $\boldsymbol{\theta}$, it suffices to study the asymptotic posterior distribution of the linear functional of $\boldsymbol{\beta}$ given by (5.6). The next theorem describes the approximate posterior distribution of $\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$. We denote by $\lambda := \lambda_{j_1,\ldots,j_s}$ the linear functional constructed as the dual basis for $\{N_{j_k}\}_{k=1}^s$ as given in Theorem 12.5 of Schumaker (2007).

Theorem 5.2. *Let us denote*

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

and $\boldsymbol{C} = (\!(\langle\lambda\circ A_k(\cdot),\lambda\circ A_{k'}(\cdot)\rangle)\!)_{k,k'=1,\ldots,p}$. *If* $\boldsymbol{C}$ *is non-singular, then under the conditions of Theorem 5.1,*

$$\left\|\Pi\left(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\in\cdot|\boldsymbol{Y}\right) - N\left(\boldsymbol{\mu}_n,\sigma_0^2\boldsymbol{\Sigma}_n\right)\right\|_{TV} = o_{P_0}(1). \tag{5.7}$$

5.4 Simulation Study

We consider the heat equation

$$\begin{aligned}\frac{\partial f_\theta(x,t)}{\partial t} &= \theta\frac{\partial^2 f_\theta(x,t)}{\partial x^2},\;\; x\in[0,1],\;\; t\in[0,5],\\ f_\theta(0,t) &= 0,\;\; f_\theta(1,t)=1,\\ f_\theta(x,0) &= \sin(\pi x).\end{aligned}$$

The true value of the parameter is chosen $\theta_0 = 0.001$. The true regression function is $f_{\theta_0}(\cdot,\cdot)$. The x and t domains are divided into equispaced grids where we obtain the observations. The weight function is chosen as $w(x,t) = x^3(1-x)^3t^3(1-t)^3$. Samples of sizes 100 and 400 are considered. We simulate 1000 replications for each case. Under each replication a sample of size 1000 is drawn from the posterior distribution of θ using Bayesian two-step method and then 95% equal tailed credible intervals are obtained. We calculate the coverage and the average length of the corresponding credible interval over these 1000 replications. The output are given in Table 5.1. The estimated standard errors of the interval length and coverage are given inside the parentheses in the table. The true distribution of error is taken either $N(0,(0.1)^2)$ or a scaled t distribution with 6 degrees of freedom, where scaling is done in order to make the standard deviation 0.1. We put an inverse gamma prior on σ^2 with shape and scale parameters being 30 and 5 respectively. We take $m = 7$. Looking at the order of k_n suggested by Theorem 5.1, k_n is chosen as 2 and 3 for $n = 100$ and $n = 400$ respectively.

Table 5.1: *Coverages and average lengths of the Bayesian credible interval*

n		$N(0,(0.1)^2)$		scaled t_6	
		coverage (se)	length (se)	coverage (se)	length (se)
100	θ	100.0 (0.00)	0.007 (0.0005)	100.0 (0.00)	0.007 (0.0005)
400	θ	100.0 (0.00)	0.004 (0.0003)	100.0 (0.00)	0.004 (0.0003)

5.5 Proofs

Proof of Theorem 5.1. By the definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ we have that

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

Subtracting the second equation from the first and applying the Mean-value Theorem, we obtain

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

Now we shall show that $\int_0^1 D_{\boldsymbol{g}_0}\left(\boldsymbol{G}(\boldsymbol{t},\boldsymbol{g}_0(\boldsymbol{t}),\boldsymbol{\theta}_0)\right)w(\boldsymbol{t})(\boldsymbol{g}(\boldsymbol{t})-\boldsymbol{g}_0(\boldsymbol{t}))d\boldsymbol{t}$ is a linear functional of $f-f_0$. Note that $\int_0^1 D_{\boldsymbol{g}_0}\left(\boldsymbol{G}(\boldsymbol{t},\boldsymbol{g}_0(\boldsymbol{t}),\boldsymbol{\theta}_0)\right)w(\boldsymbol{t})(\boldsymbol{g}(\boldsymbol{t})-\boldsymbol{g}_0(\boldsymbol{t}))d\boldsymbol{t}$ can be written as

$$\sum_{|\boldsymbol{r}|\leq\alpha}\int_0^1 D_{\boldsymbol{g}_0,\boldsymbol{r}}\left(\boldsymbol{G}(\boldsymbol{t},\boldsymbol{g}_0(\boldsymbol{t}),\boldsymbol{\theta}_0)\right)w(\boldsymbol{t})(D^{\boldsymbol{r}}f(\boldsymbol{t})-D^{\boldsymbol{r}}f_0(\boldsymbol{t}))d\boldsymbol{t}.$$

We shall prove that every term of this sum is a linear functional of $f-f_0$. We observe that

$$\begin{aligned}&\int_0^1 D_{\boldsymbol{g}_0,\boldsymbol{r}}\left(\boldsymbol{G}(\boldsymbol{t},\boldsymbol{g}_0(\boldsymbol{t}),\boldsymbol{\theta}_0)\right)w(\boldsymbol{t})(D^{\boldsymbol{r}}f(\boldsymbol{t})-D^{\boldsymbol{r}}f_0(\boldsymbol{t}))d\boldsymbol{t}\\ &=\ (-1)^{\sum_{j=1}^s r_j}\int_0^1 D^{\boldsymbol{r}}\left\{D_{\boldsymbol{g}_0,\boldsymbol{r}}\left(\boldsymbol{G}(\boldsymbol{t},\boldsymbol{g}_0(\boldsymbol{t}),\boldsymbol{\theta}_0)\right)w(\boldsymbol{t})\right\}(f(\boldsymbol{t})-f_0(\boldsymbol{t}))d\boldsymbol{t}.\end{aligned}$$

Proceeding this way we get

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

We also define

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

where $\epsilon_n \to 0$. Applying Theorem A.8 we get the posterior consistency of

$$\sup_{\boldsymbol{t}\in[0,1]^s} \sqrt{n}\|\boldsymbol{g}(\boldsymbol{t}) - \boldsymbol{g}_0(\boldsymbol{t})\|^2$$

at zero. Then using the steps of the proof of Lemma 2.13, we can prove the posterior consistency of $\boldsymbol{\theta}$. Hence, there exists a sequence $\{\epsilon_n\}$ so that $\Pi(E_n^c|\boldsymbol{Y}) = o_{P_0}(1)$. Hence, on E_n

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

By Lemma 5.3, $\boldsymbol{\Gamma}(f - f_0)$ assigns most of its mass inside a large compact set. Thus, we can assert that inside the set E_n, the asymptotic behavior of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is given by that of

$$\sqrt{n}\left((\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0))^{-1}\right)\boldsymbol{\Gamma}(f - f_0) = \sqrt{n}\left((\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0))^{-1}\right)(\Gamma(f) - \Gamma(f_0)).$$

□

Proof of Theorem 5.2. By Theorem 5.1 and (5.6), it suffices to show that

$$\sup_{\sigma^2\in\mathscr{N}}\left\|\Pi\left(\sqrt{n}\boldsymbol{H}_n^T\boldsymbol{\beta} - \sqrt{n}\,(\boldsymbol{M}(\boldsymbol{g}_0, \boldsymbol{\theta}_0))^{-1}\boldsymbol{\Gamma}(f_0) \in \cdot|\boldsymbol{Y}, \sigma^2\right) - N(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Sigma}_n)\right\|_{TV} = o_{P_0}(1)$$

for any neighborhood $\mathscr{N}$ of σ_0^2. Note that the posterior distribution of $\boldsymbol{H}_n^T\boldsymbol{\beta}$ is a normal distribution with mean vector

$$(1 + k_n^s/n)^{-1}\boldsymbol{H}_n^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}$$

and dispersion matrix

$$\sigma^2(1 + k_n^s/n)^{-1}\boldsymbol{H}_n^T(\boldsymbol{X}_n^T X_n)^{-1}\boldsymbol{H}_n$$

respectively. We calculate the Kullback-Leibler divergence between two Gaussian distributions and show that it converges in P_0-probability to zero to prove the assertion. The rest of the proof is similar to that of Theorem 2.7. □

Lemma 5.3. *Under the conditions of Theorem 5.1, the eigenvalues of* $\mathbf{Var}_0(\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y}))$ *are of the order* n^{-1} *and*

$$\max_{1\leq k\leq p}\left|[\mathbf{E}_0(\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y},\sigma^2))]_k - \int_0^1 A_k(\boldsymbol{t})f_0(\boldsymbol{t})d\boldsymbol{t}\right| \;=\; o\left(n^{-1/2}\right),$$

where $A_k(\boldsymbol{t})$ *denotes the* k^{th} *element of the vector* $\boldsymbol{A}(\boldsymbol{t})$ *as defined after the statement of Theorem 5.1 for* $k = 1, \ldots, p$.

Proof. We note that

$$\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y}) = \left(1+\frac{k_n^s}{n}\right)^{-1}\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}.$$

Hence,

$$\mathbf{Var}_0(\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y})) = \sigma_0^2\left(1+\frac{k_n^s}{n}\right)^{-2}\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{H}_n.$$

If $A_k(\cdot) \in C^{m^*}((0,1))$ for some $1 \leq m^* < m$, then by Theorem 12.7 of Schumaker (2007), we have $\sup\{|A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t})| : \boldsymbol{t} \in [0,1]^s\} = O(k_n^{-s})$, where $\tilde{A}_k(\cdot) = \boldsymbol{\alpha}_k^T\boldsymbol{N}(\cdot)$ and $\boldsymbol{\alpha}_k^T = \{\lambda_{j_1,\ldots,j_s} \circ A_k\}_{1\leq j_1,\ldots,j_s\leq(k_n+m-1)^s}$. We can write

$$\begin{aligned}
\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{H}_n \;=\;& (\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n)^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}(\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n)\\
&+\tilde{\boldsymbol{H}}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\tilde{\boldsymbol{H}}_n + (\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n)^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\tilde{\boldsymbol{H}}_n\\
&+\tilde{\boldsymbol{H}}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}(\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n),
\end{aligned}$$

where $[\tilde{\boldsymbol{H}}_n^T]_{k,} = \int_0^1 \tilde{A}_k(\boldsymbol{t})(\boldsymbol{N}(\boldsymbol{t}))^T d\boldsymbol{t}$ for $k = 1, \ldots, p$. Let us denote $\tilde{\boldsymbol{A}}(\boldsymbol{t}) = \left\{\tilde{A}_k(\boldsymbol{t})\right\}_{1\leq k\leq p}$. We study the asymptotic orders of the eigenvalues of the matrices $\tilde{\boldsymbol{H}}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\tilde{\boldsymbol{H}}_n$ and $(\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n)^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}(\boldsymbol{H}_n-\tilde{\boldsymbol{H}}_n)$. It should be noted that

$$\boldsymbol{\alpha}_k^T\int_0^1 \boldsymbol{N}(\boldsymbol{t})\boldsymbol{N}^T(\boldsymbol{t})d\boldsymbol{t}\,\boldsymbol{\alpha}_k = \int_0^1 \tilde{A}_k^2(\boldsymbol{t})d\boldsymbol{t} \asymp \|\boldsymbol{\alpha}_k\|^2 k_n^{-s},$$

the last step following from the proof of Theorem A.1. Thus, the eigenvalues of the matrix $\int_0^1 \boldsymbol{N}(\boldsymbol{t})(\boldsymbol{N}(\boldsymbol{t}))^T d\boldsymbol{t}$ are of order k_n^{-s}. We introduce the notation $\lambda(\boldsymbol{C})$ to denote an arbitrary eigenvalue of the matrix $\boldsymbol{C}$. Since the eigenvalues of $\left(\boldsymbol{X}_n^T \boldsymbol{X}_n/n\right)$ are of the order k_n^{-s}, we have

$$
\begin{aligned}
\lambda\left(\tilde{\boldsymbol{H}}_n^T (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \tilde{\boldsymbol{H}}_n\right) &\asymp \frac{k_n^s}{n} \lambda\left(\tilde{\boldsymbol{H}}_n^T \tilde{\boldsymbol{H}}_n\right) \\
&= \frac{k_n^s}{n} \lambda\left(\int_0^1 \tilde{\boldsymbol{A}}(\boldsymbol{t}) \boldsymbol{N}^T(\boldsymbol{t}) dt \int_0^1 \boldsymbol{N}(\boldsymbol{t}) (\tilde{\boldsymbol{A}}(\boldsymbol{t}))^T d\boldsymbol{t}\right) \\
&= \frac{k_n^s}{n} \lambda\left(\begin{pmatrix} \boldsymbol{\alpha}_1^T \\ \vdots \\ \boldsymbol{\alpha}_p^T \end{pmatrix} \left(\int_0^1 \boldsymbol{N}(\boldsymbol{t}) \boldsymbol{N}^T(\boldsymbol{t}) dt\right)^2 (\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_p)\right) \\
&\asymp \frac{k_n^s}{n k_n^{2s}} \lambda\left(\begin{pmatrix} \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_p \\ \vdots & \ddots & \vdots \\ \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_p & \cdots & \boldsymbol{\alpha}_p^T \boldsymbol{\alpha}_p \end{pmatrix}\right) \\
&\asymp \frac{1}{n} \lambda\left(\begin{pmatrix} \langle \lambda \circ A_1(\cdot), \lambda \circ A_1(\cdot) \rangle & \cdots & \langle \lambda \circ A_1(\cdot), \lambda \circ A_p(\cdot) \rangle \\ \vdots & \ddots & \vdots \\ \langle \lambda \circ A_1(\cdot), \lambda \circ A_p(\cdot) \rangle & \cdots & \langle \lambda \circ A_p(\cdot), \lambda \circ A_p(\cdot) \rangle \end{pmatrix}\right) \\
&= \frac{1}{n} \lambda(\boldsymbol{C}) \asymp \frac{1}{n}.
\end{aligned}
$$

Let us denote by $\mathbf{1}_{(k_n+m-1)^s}$ the $(k_n+m-1)^s$-component vector with all elements 1. Then for $k = 1, \ldots, p$,

$$
\begin{aligned}
&\left[(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)^T (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} (\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)\right]_{k,k} \\
&\asymp \int_0^1 (A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t}))(\boldsymbol{N}(\boldsymbol{t}))^T d\boldsymbol{t}\, (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \int_0^1 (A_j(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t}))(\boldsymbol{N}(\boldsymbol{t})) d\boldsymbol{t} \\
&= \frac{1}{n} \int_0^1 (A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t}))(\boldsymbol{N}(\boldsymbol{t}))^T d\boldsymbol{t} \left(\boldsymbol{X}_n^T \boldsymbol{X}_n/n\right)^{-1} \int_0^1 (A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t})) \boldsymbol{N}(\boldsymbol{t}) d\boldsymbol{t} \\
&\asymp \frac{k_n^s}{n} \int_0^1 (A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t}))(\boldsymbol{N}(\boldsymbol{t}))^T dt \int_0^1 (A_k(\boldsymbol{t}) - \tilde{A}_k(\boldsymbol{t})) \boldsymbol{N}(\boldsymbol{t}) d\boldsymbol{t} \\
&\lesssim \frac{1}{n k_n^s},
\end{aligned}
$$

Thus, the eigenvalues of $(\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)^T (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} (\boldsymbol{H}_n - \tilde{\boldsymbol{H}}_n)$ are of the order $(n k_n^s)^{-1}$ or less. Hence, the eigenvalues of $\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{H}_n$ are of the order n^{-1}.

By Theorem 12.7 of Schumaker (2007), there exists a $\boldsymbol{\beta}^*$ such that

$$\sup_{\boldsymbol{t}\in[0,1]^s}|f_0(\boldsymbol{t})-\boldsymbol{N}^T(\boldsymbol{t})\boldsymbol{\beta}^*|=O(k_n^{-s}).$$

For such a $\boldsymbol{\beta}^*$ we can write

$$\begin{aligned}
&\sqrt{n}\left|[\mathbf{E}_0(\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y}))]_k-\int_0^1 A_k(\boldsymbol{t})f_0(\boldsymbol{t})d\boldsymbol{t}\right|\\
\leq\ &\sqrt{n}\left|\left(1+\frac{k_n^s}{n}\right)^{-1}[\boldsymbol{H}_n^T\boldsymbol{\beta}^*]_k-[\boldsymbol{H}_n^T\boldsymbol{\beta}^*]_k\right|\\
&+\sqrt{n}\left(1+\frac{k_n^s}{n}\right)^{-1}\left|[\boldsymbol{H}_n^T(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T(\boldsymbol{f}_0(\boldsymbol{x})-\boldsymbol{X}_n\boldsymbol{\beta}^*)]_k\right|\\
&+\sqrt{n}\left|\int_0^1 A_k(\boldsymbol{t})f_0(\boldsymbol{t})d\boldsymbol{t}-[\boldsymbol{H}_n^T\boldsymbol{\beta}^*]_k\right|,
\end{aligned}$$

where $[\boldsymbol{H}_n^T\boldsymbol{\beta}^*]_k=\int_0^1 A_k(\boldsymbol{t})f^*(\boldsymbol{t})d\boldsymbol{t}$ and $f^*(\boldsymbol{t})=\boldsymbol{N}^T(\boldsymbol{t})\boldsymbol{\beta}^*$ for $k=1,\ldots,p$. Proceeding in the same way as in the proof of last part of Lemma 2.14, we get the second assertion. □

Remark 5.4. Since the bias and standard deviation of $[\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y})]_k$ are of order $n^{-1/2}$ as an estimator of $\int_0^1 A_k(\boldsymbol{t})f_0(\boldsymbol{t})d\boldsymbol{t}$, we can assert that for $k=1,\ldots,p$,

$$\left|[\mathbf{E}(\boldsymbol{H}_n^T\boldsymbol{\beta}|\boldsymbol{Y})]_k-\int_0^1 A_k(\boldsymbol{t})f_0(\boldsymbol{t})d\boldsymbol{t}\right|=O_{P_0}(n^{-1/2}).\tag{5.8}$$

Chapter 6

Future directions

In this chapter we reflect on some possible works related to this dissertation. In Section 6.1 we discuss the use of numerical solution of PDE as a mean of parametric inference for PDE models. Section 6.2 considers the non-additive model framework in differential equation models.

6.1 Efficient Bayesian inference for partial differential equation models using numerical solution

There is no universal rule of numerically solving PDE's. The solution depends on the nature of the PDE. We describe the numerical techniques for different kinds of PDE's

6.1.1 Parabolic Equation

The PDE is given by

$$\begin{aligned} \frac{\partial^2 f(x,t)}{\partial x^2} &= \frac{1}{c}\frac{\partial f(x,t)}{\partial t},\ 0 < x < 1,\, t > 0, \\ f(0,t) &= f(1,t) = 0, \\ f(x,0) &= g(x), \end{aligned}$$

c being a constant. The x and t domains are discretized into grids with common differences h and k respectively as

$$\begin{aligned} x_i &= ih, \\ t_j &= jk \end{aligned}$$

for $i = 0, 1, \ldots$ and $j = 0, 1, \ldots$. Let us denote by $f_{i,j}$ the numerical approximation to $f(x_i, t_j)$. Denoting $r = ck/h^2$, the recursive rule for numerical solution is given by

$$f_{i,j+1} = rf_{i-1,j} + (1 - 2r)f_{i,j} + rf_{i+1,j}.$$

6.1.2 Hyperbolic Equation

The PDE is given by

$$\begin{aligned} \frac{\partial^2 f(x,t)}{\partial x^2} &= \frac{1}{c^2}\frac{\partial^2 f(x,t)}{\partial t^2},\ 0 < x < a,\ t > 0, \\ f(0,t) &= f(a,t) = 0, \\ f(x,0) &= \phi(x),\ \left.\frac{\partial f(x,t)}{\partial t}\right|_{t=0} = \psi(x). \end{aligned}$$

Again we discretize the two domains in a similar way. Defining $s = ck/h$, the recursion rule is given by

$$f_{i,j+1} = s^2 f_{i-1,j} + 2(1 - s^2)f_{i,j} + s^2 f_{i+1,j} - f_{i,j-1}.$$

6.1.3 Elliptic Equation

Here the boundary-initial value problem is given by

$$\begin{aligned} \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} &= 0,\ 0 < x < a,\ 0 < y < b, \\ f(0,y) &= \phi_1(y),\ f(a,y) = \phi_2(y),\ f(x,0) = \psi_1(x),\ f(x,b) = \psi_2(x). \end{aligned}$$

The x and y domains are divided into grids with common difference h. The corresponding recursion rule is given by

$$f_{i,j} = \frac{1}{4}\left(f_{i-1,j} + f_{i+1,j} + f_{i,j-1} + f_{i,j+1}\right).$$

We can construct approximate likelihood using the numerical solution. If the common differences decay rapidly with increasing sample size, the approximate likelihood should behave like the actual one for sufficiently large sample sizes. Putting a prior directly on the parameter vector, we can draw the posterior samples and hence arrive at an extension of RKSB method for PDE models. Similarly it should be possible to extend the RKTB approach for PDE models. Finally for sufficiently large sample sizes the posterior distribution of the parameter vector obtained from both methods should behave like a Gaussian distribution with dispersion matrix as inverse Fisher information. Hence, the Bayes estimators would be efficient, provided the accuracy of this numerical solution technique is sufficiently high.

6.2 Bayesian uncertainty quantification for non-additive error

Suppose the response has the probability distribution

$$p(y_i|\boldsymbol{\eta}_i, \tau) = h(y_i, \tau) \exp\left(\frac{(\boldsymbol{b}(\boldsymbol{\eta}_i))^T \boldsymbol{T}(y_i) - A(\boldsymbol{\eta}_i)}{d(\tau)}\right),$$

where $h(\cdot,\cdot), \boldsymbol{b}(\cdot), \boldsymbol{T}(\cdot), A(\cdot)$ are known and $\boldsymbol{\eta}_i = \boldsymbol{f_\theta}(x_i)$ for $i = 1, \ldots, n$. Here $\boldsymbol{f_\theta}(\cdot)$ is unknown. But we are given the ODE

$$\frac{d}{dt}\boldsymbol{f_\theta}(t) = \boldsymbol{F}(t, \boldsymbol{f_\theta}(t), \boldsymbol{\theta}),$$

$\boldsymbol{F}$ being known. The Bayesian inference on $\boldsymbol{\theta}$ would be an interesting problem under this framework. One possible approach is to first approximate $\boldsymbol{f_\theta}(\cdot)$ by B-spline basis functions. Here we cannot use the normal-normal conjugacy beacuse of the non-additivity.

Hence, the posterior distribution of the spline coefficients is not analytically available and we need to use MCMC technique. But for large samples Theorem 2.1 of Ghosal (1997) indicates that the posterior of the spline coefficients is approximately normal. Now we should be able to derive the Bernstein-von Mises theorem for the posterior distribution of $\boldsymbol{\theta}$. The extension of RKTB approach should be similar. The asymptotic evaluation of the RKSB approach should be same as for additive error since we do not need conjugacy there.

REFERENCES

Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Baake, E., Baake, M., Bock, H., and Briggs, K. (1992). Fitting ordinary differential equations to chaotic data. *Physical Review A*, **45**:5524–5529.

Bergstrom, A. R. (1983). Gaussian estimation of structural parameters in higher order continuous time dynamic models. *Econometrica: Journal of the Econometric Society*, pages 117–152.

Bergstrom, A. R. (1985). The estimation of parameters in nonstationary higher order continuous-time dynamic models. *Econometric Theory*, 1(03):369–385.

Bergstrom, A. R. (1986). The estimation of open higher-order continuous time dynamic models with mixed stock and flow data. *Econometric Theory*, 2(03):350–373.

Bontemps, D. (2011). Bernstein-von mises theorems for gaussian regression with increasing number of regressors. *The Annals of Statistics*, **39**:2557–2584.

Brunel, N. J. (2008). Parameter estimation of odes via nonparametric estimators. *Electronic Journal of Statistics*, **2**:1242–1267.

Brunel, N. J., Clairon, Q., and dAlché Buc, F. (2014). Parametric estimation of ordinary differential equations with orthogonality conditions. *Journal of the American Statistical Association*, **109**:173–185.

Campbell, D. and Steele, R. J. (2012). Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, 22(2):429–443.

Campbell, D. A. (2007). *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models*. ProQuest.

Chen, T., He, H. L., Church, G. M., et al. (1999). Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, volume **4**, page 4.

Chkrebtii, O., Campbell, D. A., Girolami, M. A., and Calderhead, B. (2013). Bayesian uncertainty quantification for differential equations. *arXiv preprint arXiv:1306.2365*.

De Boor, C. (1978). *A Practical Guide to Splines*, volume **27**. Springer-Verlag New York.

Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Oxford University Press.

Gabrielsson, J. and Weiner, D. (2006). *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*. Swedish Pharmaceutical Press.

Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, **91**:1400–1412.

Ghosal, S. (1997). Normal approximation to the posterior distribution for generalized linear models with many covariates. *Mathematical Methods of Statistics*, **6**:332–348.

Ghosh, S. K. and Goyal, L. (2010). Statistical inference for non-linear models involving ordinary differential equations. *Journal of Statistical Theory and Practice*, **4**:727–742.

Girolami, M. (2008). Bayesian inference for differential equations. *Theoretical Computer Science*, **408**:4–16.

Gugushvili, S. and Klaassen, C. A. (2012). $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli*, **18**:1061–1098.

Hairer, E., Nørsett, S., and Wanner, G. (1987). *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer-Verlag, Berlin.

Henrici, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations*, volume 1. New York: Wiley.

Jaeger, J. et al. (2012). *Functional Estimation in Systems Defined by Differential Equations Using Bayesian Smoothing Methods.* PhD thesis, Université catholique de Louvain.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, **2**:164–168.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, **11**:431–441.

Mattheij, R. and Molenaar, J. (1996). *Ordinary Differential Equations in Theory and Practice*, volume **43**. SIAM.

Müller, T. and Timmer, J. (2004). Parameter identification techniques for partial differential equations. *International Journal of Bifurcation and Chaos*, **14**:2053–2060.

Müller, T. G. and Timmer, J. (2002). Fitting parameters in partial differential equations from partially observed noisy data. *Physica D: Nonlinear Phenomena*, **171**:1–7.

Nowak, M. and May, R. M. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press.

Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics*, **38**:435–481.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**:741–796.

Rogers, S., Khanin, R., and Girolami, M. (2007). Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8**(Suppl 2):S2.

Schumaker, L. (2007). *Spline Functions*. Cambridge University Press.

Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, **3**:28–46.

Voit, E. O. and Almeida, J. (2004). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, **20**:1670–1681.

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.

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. *The Annals of Statistics*, **38**:23–51.

Xun, X., Cao, J., Mallick, B., Maity, A., and Carroll, R. J. (2013). Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, **108**:1009–1020.

Yoo, W. W. and Ghosal, S. (2014). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *arXiv preprint arXiv:1411.6716*.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**:1760–1782.

Zhou, S. and Wolfe, D. A. (2000). On derivative estimation in spline regression. *Statistica Sinica*, **10**:93–108.

APPENDIX

Appendix A

Some auxiliary results

We state some results we have used while deriving the theoretical results.

For the nonparametric regression model

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

using splines of order m with error variance σ^2, the least square estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}.$$

The estimated r^{th} derivative of the function at any $t \in [0,1]$ is given by

$$\hat{f}^{(r)}(t) = \hat{\boldsymbol{\beta}}^T\boldsymbol{N}^{(r)}(t)$$

for $r = 0, 1, \ldots, m-2$. Hence, the variance of the estimator of the r^{th} order derivative of $\hat{f}(t)$ is $\left(\boldsymbol{N}^{(r)}(t)\right)^T\left(\boldsymbol{X}_n^T\boldsymbol{X}_n\right)^{-1}\boldsymbol{N}^{(r)}(t)$. The asymptotic orders of these variances were derived in Theorem 2.1 of Zhou et al. (1998) and Lemma 5.4 of Zhou and Wolfe (2000) which we state in the next theorem.

Theorem A.1 (Theorem 2.1 (Zhou et al., 1998), Lemma 5.4 (Zhou and Wolfe, 2000))**.** *For any $0 \leq r \leq m-2$, there exist constants $L_{\max} > L_{\min} > 0$ such that*

$$\frac{L_{\min}\sigma^2k_n^{2r+1}}{n} \leq \operatorname{Var}\left(\hat{f}^{(r)}(t)\right) \leq \frac{L_{\max}\sigma^2k_n^{2r+1}}{n}.$$

Now we mention a number of results from Kleijn and van der Vaart (2012) which have been used in Chapters 3 and 4. In this paper the working density of the observations are taken as $p_{\boldsymbol{\theta}}(\cdot)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, whereas the true density is $p_0(\cdot)$ with corresponding probability measure P_0. They denote

$$Q_{\boldsymbol{\theta}}(\cdot) = P_0 \left(\frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_0}} \mathbb{1}(\cdot) \right).$$

The main result of the paper is given in Theorem 2.1 which we state below.

Theorem A.2 (Theorem 2.1 (Kleijn and van der Vaart, 2012))**.** *Suppose there exist stochastically bounded random vectors $\boldsymbol{\Delta}_{n,\boldsymbol{\theta}_0}$ and a positive definite matrix $\boldsymbol{V}_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that for every compact set $K \subset \mathbb{R}^p$,*

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

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

$$P_0^{(n)} \Pi_n \left(\sqrt{n} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \right\| > M_n | \boldsymbol{Y} \right) \to 0, \tag{A.2}$$

then

$$\left\| \Pi_n \left(\sqrt{n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \cdot | \boldsymbol{Y} \right) - \boldsymbol{N}(\boldsymbol{\Delta}_{n,\boldsymbol{\theta}_0}, \boldsymbol{V}_{\boldsymbol{\theta}_0}^{-1}) \right\|_{TV} \overset{P_0}{\to} 0.$$

We also state some other results from Kleijn and van der Vaart (2012).

Theorem A.3 (Theorem 3.1 (Kleijn and van der Vaart, 2012))**.** *Let the function $\boldsymbol{\theta} \mapsto \log(p_{\boldsymbol{\theta}}(Y_1))$ is differentiable at $\boldsymbol{\theta}_0$ in P_0-probability with derivative $\dot{\ell}_{\boldsymbol{\theta}_0}(Y_1)$ and*
(i) there is an open neighborhood U and a square integrable function $m_{\boldsymbol{\theta}_0}(\cdot)$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U$

$$\left| \log \frac{p_{\boldsymbol{\theta}_1}}{p_{\boldsymbol{\theta}_2}} \right| \leq m_{\boldsymbol{\theta}_0} \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\|, \ (P_0\text{-}a.s.),$$

(ii) the Kullback-Leibler divergence with respect to P_0 has a second order Taylor expansion

around $\boldsymbol{\theta}_0$*:*

$$-P_0 \log \frac{p_{\boldsymbol{\theta}_1}}{p_{\boldsymbol{\theta}_2}} = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{V}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)$$

for some positive definite matrix $\boldsymbol{V}_{\boldsymbol{\theta}_0}$*. Also assume that* $P(p_{\boldsymbol{\theta}}/p_{\boldsymbol{\theta}_0}) < \infty$ *for all* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$ *and* $P_0(\exp(sm_{\boldsymbol{\theta}_0})) < \infty$ *for some* $s > 0$*. Additionally assume that the prior possesses a density that is continuous and positive in a neighborhood of* $\boldsymbol{\theta}_0$*. Furthermore, let* $P_0\left(\dot{\ell}_{\boldsymbol{\theta}_0}\dot{\ell}^T_{\boldsymbol{\theta}_0}\right)$ *is invertible and that for every* $\epsilon > 0$ *there exists a sequence of tests* $\{\phi_n\}$ *such that*

$$P_0\phi_n \to 0, \quad \sup_{\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\epsilon\}} Q^n_{\boldsymbol{\theta}}(1-\phi_n) \to 0. \tag{A.3}$$

Then the posterior converges at rate $n^{-1/2}$*, that is, for every sequence* $\{M_n\}$ *with* $M_n \to \infty$*,*

$$\Pi_n\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq M_n/\sqrt{n}|\boldsymbol{Y}\right) \stackrel{P_0}{\to} 0.$$

Theorem A.4 (Theorem 3.2 (Kleijn and van der Vaart, 2012))**.** *Assume that* $\boldsymbol{\Theta}$ *is compact and that* $\boldsymbol{\theta}_0$ *is a unique point of minimum of* $-P_0\log(p_{\boldsymbol{\theta}})$*. Furthermore assume that* $P_0(p_{\boldsymbol{\theta}}/p_{\boldsymbol{\theta}_0}) < \infty$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and that the map*

$$\boldsymbol{\theta} \mapsto P_0\left(\frac{p_{\boldsymbol{\theta}}}{p^s_{\boldsymbol{\theta}_1}p^{1-s}_{\boldsymbol{\theta}_0}}\right)$$

is continuous at $\boldsymbol{\theta}_1$ *for every* s *in a left neighborhood of 1 for every* $\boldsymbol{\theta}_1$*. Then there exist tests* $\{\phi_n\}$ *satisfying* (A.3)*. A sufficient condition is that for every* $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ *the maps* $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}/p_{\boldsymbol{\theta}_1}$ *and* $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}/p_{\boldsymbol{\theta}_0}$ *are continuous in* $L_1(P_0)$ *at* $\boldsymbol{\theta} = \boldsymbol{\theta}_1$*.*

Theorem A.5 (Theorem 3.3 (Kleijn and van der Vaart, 2012))**.** *Under the conditions of Theorem A.3, for every sequence* $\{M_n\}$ *such that* $M_n \to \infty$ *there exists a sequence of tests* $\{\omega_n\}$ *such that for some constants* $D > 0, \epsilon > 0$ *and large enough* n

$$P^n_0\omega_n \to 0, \;\; Q^n_{\boldsymbol{\theta}}(1-\omega_n) \leq \exp\left\{-nD\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \wedge \epsilon^2\right)\right\}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *such that* $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq M_n/\sqrt{n}$*.*

Now we state some of the results given in Van der Vaart (1998).

Theorem A.6 (Theorem 19.4 (Van der Vaart, 1998)). *Every class $\mathscr{F}$ of measurable functions such that for every $\epsilon > 0$, $N_{[\,]}(\epsilon, \mathscr{F}, L_1(p))$, the minimum number of ϵ-brackets to cover the class is finite, is P-Glivenko-Cantelli.*

Lemma A.7 (Lemma 19.31 (Van der Vaart, 1998)). *For each $\boldsymbol{\theta}$ in an open subset of Euclidean space let $x \mapsto m_{\boldsymbol{\theta}}(x)$ be a measurable function such that the map $\boldsymbol{\theta} \mapsto m_{\boldsymbol{\theta}}(x)$ is differentiable at $\boldsymbol{\theta}_0$ for almost every x (or in probability) with derivative $\dot{m}_{\boldsymbol{\theta}_0}(x)$ and such that, for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}_0$, and for a measurable function $\dot{m}(\cdot)$ such that $P\dot{m}^2 < \infty$,*

$$\|m_{\boldsymbol{\theta}_1}(x) - m_{\boldsymbol{\theta}_2}(x)\| \leq \dot{m}(x) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Then we get that

$$\mathbb{G}_n \left(r_n(m_{\boldsymbol{\theta}_0 + \tilde{\boldsymbol{h}}_n/r_n} - m_{\boldsymbol{\theta}_0}) - \tilde{\boldsymbol{h}}_n^T \dot{m}_{\boldsymbol{\theta}_0} \right) \overset{P_0}{\to} 0$$

for every sequence $r_n \to \infty$ and every stochastically bounded random sequence $\tilde{\boldsymbol{h}}_n$.

Now we state a result from Yoo and Ghosal (2014) used in Chapter 5. Suppose the true regression model is given by $Y = f_0(\boldsymbol{x}) + \varepsilon$, where $\boldsymbol{x} \in [0,1]^s$. For a given vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s)^T$ of positive integers let $f_0 \in \mathscr{H}^{\boldsymbol{\alpha}}([0,1]^s)$, that is, $\partial^{\alpha_k}/\partial x_k^{\alpha_k}$ is uniformly bounded for $k = 1, \ldots, s$. The spline approximation of $f_0(\cdot)$ is given by $f(\cdot) = \boldsymbol{\beta}^T \boldsymbol{N}(\cdot)$, where $\boldsymbol{N}(\cdot) = \{\prod_{k=1}^s N_{j_k}(\cdot)\}_{1 \leq j_1 \leq J_1, \ldots, 1 \leq j_s \leq J_s}$. Let us denote $\alpha^* = s/\sum_{k=1}^s \alpha_k^{-1}$. Using Gaussian distribution as the working model for error and putting Gaussian prior on $\boldsymbol{\beta}$, we have the following result.

Theorem A.8 (Theorem 4.4 of Yoo and Ghosal (2014)). *If $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+s)\}}$ for $k = 1, \ldots, s$, then*

$$\mathrm{E}_0 \Pi \left(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty > M_n \epsilon_{n,\infty} | \boldsymbol{Y} \right) \to 0$$

for any $M_n \to \infty$, where $\epsilon_{n,\infty} = (\log n/n)^{\alpha^*\{1 - \sum_{k=1}^s (r_k/\alpha_k)\}/(2\alpha^*+s)}$.

Theorem A.9 (Binomial Inverse Theorem). *For square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with $\boldsymbol{A}$ non-singular,*

$$(\boldsymbol{A} + \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} \left(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{A}^{-1} \right)^{-1} \boldsymbol{B}\boldsymbol{A}^{-1},$$

$\boldsymbol{I}$ *being the identity matrix.*

The above result can be directly verified by multiplying the right side of the above display with $\boldsymbol{A} + \boldsymbol{B}$.