Nonparametric Models for Longitudinal Data Using Bernstein Polynomial Sieve

Liwei Wang and Sujit K. Ghosh

Department of Statistics, North Carolina State University

NC State Department of Statistics Technical Report # 2651

Abstract

We develop a new nonparametric approach to the analysis of irregularly observed longitudinal data using a sieve of Bernstein polynomials within a Gaussian process framework. The proposed methodology has a number of novel features: (i) both the mean function and the covariance function can be estimated simultaneously under a set of mild regularity conditions; (ii) the derivative of the response process can be analyzed without additional modeling assumptions; (iii) shape constraint of the mean and covariance functions (e.g. nonnegativity, monotonicity and convexity) can be handled in a straightforward way; and (iv) the L_p approximation of the Gaussian process using the Bernstein polynomial sieve is established rigorously under mild regularities conditions. Further, in order to choose the appropriate order of the Bernstein sieve, a new Bayesian model selection criterion is proposed based on a predictive cross-validation criterion. Superior performance of the proposed nonparametric model and model selection criterion is demonstrated using both synthetic and real data examples.

Key Words: Bernstein polynomial sieve, Gaussian process, longitudinal data, MCMC methods, mixed effects model.

1 Introduction

In a wide variety of disciplines such as agriculture, biology, business, epidemiology, medicine, and social science, data are collected repeatedly at a sequence of time points on randomly selected subjects. One typical example involves observations from the sample path of a

growth curve (see Figure 3), where height/weight of a subject is measured repeatedly over time. In addition, some other related predictors such as gender and treatment group are also recorded at the baseline. Longitudinal models target at modeling the relationship between the response curves and predictors. Meanwhile, the mean of response curves is sometimes required to satisfy certain shape constraints such as non-negativity, monotonicity and convexity. For example, a growth curve must be nondecreasing, and recorded height/weight must be positive. In order to obtain a realistic estimate of the response curve, the natural shape constraints should also be considered and preserved for any finite sample estimator.

Longitudinal study involves temporal order of exposure and outcome, which requires the modelling of the auto-correlation function within the subject. Common practice assumes that the response is the additive effects of three parts: the fixed effects, which capture the underlying trend; the random effects, which capture the heterogeneity across subjects; and the noise part, which captures the observational or measurement errors. Linear mixed effects model (LMM; Laird and Ware 1982) is widely used such that both the fixed effects and the random effects are linear functions of some predictors, and the the vector of the random effects follows a multivariate Gaussian distribution. The noises are also assumed independently of the random effects to follow some Gaussian distribution with mean 0. Many standard software packages are available (e.g. PROC MIXED in SAS, lme in R, etc.) which allows the estimation of the parameters using maximum likelihood (ML; Robinson 1991), restricted maximum likelihood (REML; Patterson and Thompson 1971; Harville 1977), or expectation-maximization (EM) algorithm (Lindstrom and Bates 1988). But LMM is based on the linearity assumption that relates responses to predictors and also the covariance matrix are often chosen parametrically assuming regularly spaced sequence of time points. When the relationship between the response and predictors are nonlinear and data are observed irregularly, LMMs may often fail to reveal the true underlying trend and the

correlation structure.

Nonlinear mixed effects models (NLMMs) are proposed to handle the complex relationship between response and predictors. Davidian and Giltinan (1995), Davidian and Giltinan (2003), and Serroyen et al. (2009) provided comprehensive reviews on NLMM with various applications in practice. When we can identify the form of nonlinear relationship from the mechanistic theory, the parametric NLMM is employed. Some typical applications include the pharmacokinetics and pharmacodynamic modeling (Davidian and Giltinan 1995, Chpater 5), HIV dynamics modelling (Han and Chaloner 2004), and prostate specific antigen modeling (Morrell et al. 1995). Moreover, classical NLMM can be implemented directly in some popular statistical software, such as PROC NLMIXED in SAS and nlme in R. However, the assumption that the functional form of the nonlinear relationship is known may often turn out to be restrictive. Misspecified nonlinear relationship is likely to cause improper use of NLMM. Ke and Wang (2001) proposed a semiparameteric NLMM and applied their model to AIDS data, where only the mean function is modelled nonparametrically. A nonparametric NLMM was proposed by Lindstrom (1995), which replaced the nonlinear function with a free-note spline shape function. But the resulting covariance structure fitted in this way is not easy to interpret. In general, although NLMM provides a greater flexibility in capturing the possible relationships between the responses and predictors, the model is still based on an assumed nonlinear functional form which may only be a crude approximation to the true relation between the response and predictors. Moreover, in practice the correlation function is often modelled parametrically.

Assuming every response curve is a realization of a Gaussian process at certain time points, functional principal component analysis (FPCA) is an alternative way to tackle such longitudinal data problem owing to the Karhunen-Loève (K-L) expansion of a Gaussian process. Chiou et al. (2003) proposed a class of nonparametric functional regression models

with linear combination of eigenfunctions. Di et al. (2009) presented method of moments and smoothing techniques to obtain a smooth estimator to the covariance function and its corresponding eigenfunctions. Staicu et al. (2010) suggested the use of a restricted likelihood ratio test which is a step-wise testing for zero variance components in order to select the number of eigenfunctions. Crainiceanu and Goldsmith (2010) gave an example applying the functional principal component analysis on the sleep EEG data. However, to implement FPCA, in practice one often use empirical estimates of eigenfunctions in the first place which may cause problem when sample size is not large and when observations are very sparse or missing or censored. Also in Crainiceanu and Goldsmith (2010)'s example, they used the sample mean at every time point to estimate the mean function, which is too rough and may not satisfy a required shape constraints. One of the most common limitations of the FPCA is that these methods are primarily based on two stages of estimations: (i)the eigenfunctions are estimated based on estimated residuals and then (ii) these estimated functions are plugged-in to estimate the mean function or to predict the future observations assuming the estimated functions are "known" functions. These plugged-in estimated functions often lead to the underestimation of the overall uncertainty of the predictive process.

In the present paper, we propose a class of linear mixed effects models using Bernstein polynomial sieve to overcome some of the limitations mentioned above. With our proposed model, under a set of mild regularity conditions on the Gaussian process, we first establish the uniform approximation of an arbitrary Gaussian process by a sieve of Bernstein polynomials. Convergence properties of our proposed model are established in terms of L_2 and more generally for L_p norms under some mild regularity conditions. Additionally, we can easily incorporate shape restrictions of the mean curves by utilizing the attractive properties of Bernstein polynomials which allows us to model various popular shape constraints such as monotone, concave/convex, and various combinations of such shapes. In order to use the

method for finite sample, we still need to "choose" the order of the Bernstein sieve. A new Bayesian model selection criterion based on predictive divergence is proposed. Several simulation studies are presented to illustrate its superior performance compared to some popular Bayesian model selection criteria.

The paper is organized as follows. Section 2 presents the class of linear mixed models using Bernstein polynomial sieve to approximate the longitudinal model within a Gaussian process framework. Meanwhile, a new model selection criterion which we call the Bayesian predictive divergence criterion (BPDC) is proposed and computation details are described. In Section 3, the approximation properties of the proposed model are discussed. The corresponding proofs are provided in the (web) Appendix. Section 4 presents two simulation studies. One is to examine the performance of BPDC by comparing with a few of the popular Bayesian model selection criteria. The other is to examine the accuracy of the approximation model using Bernstein polynomials to a Gaussian process with nonlinear mean and complex covariance functions. In Section 5, our proposed model and criterion are illustrated on the Berkeley growth data. Finally, a discussion of the proposed methodology and future research work are presented in Section 6.

2 Gaussian Processes Approximation with Bernstein Polynomials

Let $Y_i(t)$ denote the measured response obtained at time $t \in [0, T]$ for subject i. Suppose we observe $Y_i(t)$ at selected set of time points $t_{i,1} < t_{i,2} < \ldots < t_{i,J_i}$ where $J_i \geq 2$ for $i = 1, \ldots, I$. Denote $y_{i,j} = Y_i(t_{i,j})$ for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$. To begin with, for simplicity we consider a simple longitudinal model with t as the only predictor. Additional predictors can be easily incorporated in the model. Let $GP(\mu(\cdot), K(\cdot, \cdot))$ denote a Gaussian process with mean function $\mu(\cdot)$ and covariance function $K(\cdot, \cdot)$. To start with, assume that

the underlying model for the simple longitudinal study is given by,

$$Y_i(t) = X_i(t) + \epsilon_i(t), \quad i = 1, \ldots, I, \tag{1}$$

where $X_i(\cdot) \overset{\text{iid}}{\sim} GP(\mu(\cdot), K(\cdot,\cdot))$, and $\epsilon_i(\cdot) \overset{\text{iid}}{\sim} GP(0, \sigma_\epsilon^2 I(\cdot,\cdot))$ where $I(t,t') = 1$ if $t = t'$ and 0 otherwise. In practice, we often do not have any specific knowledge to specify the functional forms, but some shape properties of $\mu(\cdot)$ may be known (e.g. growth curves are necessarily non-decreasing, etc.). To fit Model (1) without loosing much accuracy, we propose a class of linear mixed effects models using a sieve of Bernstein polynomials as an approximation, where Bernstein polynomial is employed due to its optimal property in retaining the shape information (Carnicer and Pena 1993). To start with, we give a brief introduction to some attractive properties of Bernstein polynomials.

2.1 Shape Restricted Bernstein Polynomials

The Bernstein polynomials (BP) were introduced by Sergei Natanovich Bernstein in 1912, and since then this class of polynomials has become one of the most popular classes of polynomials in numerical analysis. Lorentz (1953) is a complete handbook on Bernstein polynomials, including complete proofs of many interesting theorems related to Bernstein polynomial and its generalizations.

The Bernstein basis polynomials on $[0,1]$ of degree $m-1$ is defined as

$$b_{k,m}(t) = \binom{m-1}{k-1} t^{k-1}(1-t)^{m-k}, \quad k = 1, \ldots, m, t \in [0,1], m = 2, 3, \ldots. \tag{2}$$

The Bernstein polynomial sieve (BPS) of order $m-1$ is defined as any linear combination of Bernstein basis polynomials,

$$B_m(t) = \sum_{k=1}^{m} a_{k,m} b_{k,m}(t), \quad t \in [0,1], \tag{3}$$

where $a_{k,m}$ can be any real number. Notice that BPS includes iterates of BP's which enjoy higher order convergence than usual BPS (Kelisky and Rivlin 1967). Any BPS is continuous

and infinitely many differentiable on $[0,1]$. One of the nice properties of BPS is that the derivatives of a BPS is still a BPS of a lower degree. In fact, it is well known that

$$B'_m(t) = (m-1)\sum_{k=1}^{m-1}(a_{k+1,m} - a_{k,m})b_{k,m-1}(t), \tag{4}$$

and more generally, the $l-$th order derivative is still a BP and is given by

$$B_m^{(l)}(t) = l!\binom{m-1}{l}\sum_{k=1}^{m-l}\nabla^{(l)}a_{k,m}b_{k,m-l}(t), \tag{5}$$

where $\nabla^{(l)}a_{k,m} = \nabla^{(l-1)}a_{k+1,m} - \nabla^{(l-1)}a_{k,m}$ for $l = 1, 2, \ldots$, and $\nabla^{(0)}a_{k,m} = a_{k,m}$. From above, it follows that linear restrictions on the coefficients $a_{k,m}$'s induce restrictions on the derivatives of $B_m(t)$ for $t \in [0,1]$. For example, if $a_{1,m} \leq a_{2,m} \leq \ldots \leq a_{m,m}$, then $B'_m(t) \geq 0$ for $t \in [0,1]$. Thus shape constraints like nonnegativity, monotonicity, and convexities can be easily imposed by using finite dimensional linear inequality constraints on the coefficients. Gal (2008), Chapter 1, provided complete proofs to many shape-preserving properties and more interesting properties. Wang and Ghosh (2012) elaborated some of the interesting results on monotone function, convex/concave function, and monotonous convex function, and established the strong consistency property of Bernstein sieve using constrained least square estimation of the mean function only.

Meanwhile, the convergence properties of Bernstein polynomials to continuous functions have been thoroughly studied. Lorentz (1953), Chapter 1, described the Bernstein Weierstrass approximation theorem which provides the convergence properties of Bernstein polynomials in L_∞ norm. Hoeffding (1971) discussed the L_1 norm approximation error for Bernstein polynomials. Jones (1976) proved several theorems for the approximation error for Bernstein polynomials in L_2 norm. More recently, Khan (1985) generalized the convergence of BPS to L_p norms under mild regularity condtions. All these theorems assist us to understand the convergence properties of Bernstein polynomial sieves.

2.2 A Class of Linear Mixed Effects Models with Bernstein Polynomials

Suppose both mean function and covariance function are continuous and satisfy some mild regularity conditions (as stated in Theorems 1 and 2 in Section 2.3), we claim that the Model (1) can be approximated by the following class of linear mixed effects models using a sieve of Bernstein polynomials (as $m \to \infty$),

$$Y_i(t) = \sum_{k=1}^{m} b_{k,m}(t)\beta_{i,k} + \epsilon_i(t), \quad i = 1, \ldots, I, \tag{6}$$

where $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,m})^T$, $\beta_i \overset{\text{iid}}{\sim} N(\beta_0, D)$ and $\epsilon_i(t) \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$. Also, we may assume $\epsilon_i(t) \overset{\text{ind}}{\sim} N(0, \sigma_i^2)$ to represent heterogenous errors. With data $\{y_{i,j}\}$ observed, we can use Bayesian methods to fit the model for each m,

$$y_{i,j} = \sum_{k=1}^{m} b_{k,m}(t_{i,j})\beta_{i,k} + \epsilon_{i,j}, \quad i = 1, \ldots, I, \tag{7}$$

with some proper prior assigned to $\theta = (\beta_0, D, \sigma_\epsilon^2)$. Finally, the order m of the Bernstein sieve is chosen using a predictive criterion.

It is easy to see that with Model (6), we have approximated true $X_i(t)$ GP with the Gaussian process

$$GP\left(\sum_{k=1}^{m} b_{km}(t)\beta_{0,k}, \sum_{k=1}^{m}\sum_{h=1}^{m} b_{k,m}(t)b_{h,m}(s)D_{k,h}\right). \tag{8}$$

The details of the convergence result is presented in Section 3. If $\mu(\cdot)$ and $K(\cdot,\cdot)$ are assumed differentiable, and as differentiation is a linear operation (Solak et al. 2003), the derivative of a Gaussian process $X(\cdot) \sim GP(\mu(\cdot), K(\cdot,\cdot))$ is still a Gaussian process such that $X'(\cdot) \sim GP(\mu^*(\cdot), K^*(\cdot,\cdot))$, where $\mu^*(t) = \mu'(t)$ and $K^*(t,s) = \frac{\partial^2 K(t,s)}{\partial t \partial s}$. So we can also explore the the derivative of the Gaussian process $X_i'(t)$ in Model (1). Such a derivative process is of much practical interest when estimating the rate of change of the growth as we illustrated in the introduction and the hypothesized dark energy equation of state in cosmology (Holsclaw

et al. 2013). With Equation (5), it can be shown that $X_i'(t)$ is approximated with

$$GP\Bigg((m-1)\sum_{k=1}^{m-1} b_{k,m-1}(t)(\beta_{0,k+1}-\beta_{0,k}), \tag{9}$$
$$(m-1)^2\sum_{k=1}^{m-1}\sum_{h=1}^{m-1} b_{k,m-1}(t)b_{h,m-1}(s)(D_{k+1,h+1}-2D_{k+1,h}+D_{k,h})\Bigg).$$

This follows by another interesting property of BPs which states that the derivatives of the function are uniformly approximated by the derivatives of the corresponding BPs (Lorentz 1953, Page 13).

Let J be the number of unique time points in the data set. To avoid collinearity issue with large degree polynomials, m should always be chosen less than J in Model (6). Tenbusch (1997) suggested a more strict upper bound for the choice of m in nonparametric regression with Bernstein polynomials such that $m \leq [J^{3/4}]$. With $m = 1$, we can only approximate a degenerated Gaussian process. Hence, we also require m to be greater than 1. So in practice the value of m is chosen from set $\{2, \ldots, [J^{3/4}]\}$ using a predictive divergence criterion that we describe in the next subsection.

2.3 Bayesian Model Selection using Predictive Divergence

In this subsection, we propose a new Bayesian model selection criterion based on predictive divergence for the purpose of choosing the tuning parameter m. However, the proposed criterion is not restricted to the choice of Bernstein sieve model and it can be applied for general model selection purpose among many competing models. The new cross-validation model selection criterion, Bayesian predictive divergence criterion (BPDC) is motivated by Davies et al. (2005) and Geisser and Eddy (1979). We generalize the predictive divergence defined in Davies et al. (2005) to Bayesian inferential frame work. Suppose the candidate Bayesian model is $y \sim f(\cdot|\theta)$ with a prior distribution $\theta \sim \pi(\theta)$. We define Bayeisan predictive discrepancy for the i-th independent subject as the expectation of $-2\log f(y_i|\theta)$

with respect to the posterior distribution of θ depending on the leave-one-out data y_{-i}, say

$$d_i^B(y, f) = \int -2 \log f(y_i|\theta) p(\theta|y_{-i}) d\theta, \tag{10}$$

where $p(\theta|y_{-i}) = f(y_{-i}|\theta)p(\theta)/\int f(y_{-i}|\theta)p(\theta)d\theta$ denotes the posterior distribution of θ given the data $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_I)$. The total Bayeisan predictive discrepancy over all subjects is then defined as

$$d_{BPDC}(y, f) = \sum_{i=1}^{I} d_i^B(y, f). \tag{11}$$

Taking expectation of $d_{BPDC}(y, f)$ with respect to the true model, we get

$$\begin{aligned} \Delta_{BPDC}(f) &= E_y[d_{BPDC}(y, f)], & (12) \\ &= E_y[\sum_{i=1}^{I} \int -2 \log f(y_i|\theta) p(\theta|y_{-i}) d\theta]. & (13) \end{aligned}$$

Our target is to find an unbiased estimator to $\Delta_{BPDC}(f)$. Note that in Equation (13), the term inside expectation is a function of y. Thus, the term inside is an unbiased estimate of $\Delta_{BPDC}(f)$, and we can then define the Bayesian predictive divergence criterion (BPDC) as

$$\begin{aligned} \text{BPDC} &= \sum_{i=1}^{I} \int -2 \log f(y_i|\theta) p(\theta|y_{-i}) d\theta, & (14) \\ &= \sum_{i=1}^{I} -2 E_\theta[\log f(y_i|\theta)|y_{-i}]. \end{aligned}$$

With simple models, we can compute $\int \log f(y_i|\theta)p(\theta|y_{-i})d\theta$ directly for each i. But for most of complex Bayesian models with high dimensional θ, it is often not possible to integrate out θ. In this situation, without losing much accuracy, ideally we can use Monte Carlo integration to calculate $E_\theta[\log f_i(y_i|\theta)|y_{-i}]$ by generating samples $\theta_i^{(l)} \sim p(\theta|y_{-i})$ for $l = 1, \ldots, L$. So one option for computing BPDC in this case is to generate $\theta_i^{(l)}$ by MCMC simulation with data that excludes the ith subject and repeat the procedure for all I subjects. This means that we have to run the MCMC simulation I times to get the value of BPDC, which is computationally expensive, especially when we have lots of subjects in the data set.

Importance sampling (IS) provides a solution to this computation problem. Gelfand and Dey (1994), Peruggia (1997), and Vehtari and Lampinen (2002) advocated the use of IS in computing the expectation with respect to the case-deletion posterior.

Suppose we want to compute the expectation $E_{p_i}[g(\theta)] = \int g(\theta)p_i(\theta)d\theta$ with respective to the density $p_i(\theta)$ for $i = 1, \ldots, I$ (e.g. $p_i(\theta) = p(\theta|y_{-i})$, etc.). Instead of generating samples from $p_i(\theta)$ and repeat the procedure I times, we can obtain samples from a candidate density $p(\theta)$ and use the identity $E_{p_i}[g(\theta)] = \int p_i(\theta)g(\theta)d\theta = \int \frac{p_i(\theta)}{p(\theta)}g(\theta)p(\theta)d\theta = E_p[\frac{p_i(\theta)}{p(\theta)}g(\theta)]$. Now if $p(\theta) = q(\theta)/C$ and $p_i(\theta) = q_i(\theta)/C_i$ are known only by their kernel functions $q(\theta)$ and $q_i(\theta)$ respectively, then $E_{p_i}[g(\theta)]$ can be estimated consistently (as $L \to \infty$) by

$$\bar{g}_L = \sum_{l=1}^{L} g(\theta^{(l)})\bar{w}_l(p_i, p),$$

where $\theta^{(l)} \overset{\text{iid}}{\sim} p(\theta)$, $\bar{w}_l(p_i, p) = w_l(p_i, p)/\sum_{h=1}^{L} w_h(p_i, p)$, and $w_l(p_i, p) = q_i(\theta^{(l)})/q(\theta^{(l)})$ (Peruggia 1997). The strong law of large number for Markov chains always implies the consistency of the estimator, but its performance depends critically on the variance of the IS weight $w_l(p_i, p)$. For a standard Bayesian linear regression model, Peruggia (1997) proved necessary and sufficient conditions for finite variance of IS weights.

As the above discussion attests, we can apply IS method to compute BPDC. Suppose we are given a Bayesian model where $\pi(\theta)$ denotes the prior density function of the parameter θ, a data set $y = (y_1, \ldots, y_I)$ where y_i's are mutually independent vectors given θ, and MCMC samples $\theta^{(l)}$, $l = 1, \ldots, L$, based on the full data y, i.e. $\theta^{(l)} \sim p(\theta|y)$. Letting $p(\theta) = p(\theta|y)$, $p_i(\theta) = p(\theta|y_{-i})$, $C = m(y)$, and $C_i = m(y_{-i})$, we have $w_l(p_i, p) = 1/f(y_i|\theta^{(l)})$, and hence

$$\bar{w}_l(p_i, p) = \frac{1/f(y_i|\theta^{(l)})}{\sum_{h=1}^{L} 1/f(y_i|\theta^{(h)})} = \left(\sum_{h=1}^{L} \frac{f(y_i|\theta^{(l)})}{f(y_i|\theta^{(h)})}\right)^{-1}.$$

Finally, using the above defined IS we can approximate BPDC by

$$\begin{aligned} \text{BPDC}_a &= -2\sum_{i=1}^{I}\sum_{l=1}^{L}\log f(y_i|\theta^{(l)})\bar{w}_l(p_i,p), \\ &= -2\sum_{i=1}^{I}\sum_{l=1}^{L}\log f(y_i|\theta^{(l)})\left(\sum_{h=1}^{L}\frac{f(y_i|\theta^{(l)})}{f(y_i|\theta^{(h)})}\right)^{-1}. \end{aligned} \tag{15}$$

We use the above approximation to select the sieve order m of our proposed model.

3 Convergence Properties

In this section, we present convergence properties of the finite dimensional approximation of the $GP(\mu(\cdot), K(\cdot,\cdot))$ by a class of random BPS of the form $\sum_{k=1}^{m} b_{km}(t)\beta_{km}$, where $\beta_m = (\beta_{1m},\ldots,\beta_{mm})^T \sim N(\mu_m, D_m)$ as $m \to \infty$.

Theorem 1. *Consider a Gaussian process $X(t)$ defined on $[0,1]$ with continuous mean function $\mu(t)$ and continuous nonnegative definite covariance function $K(t,s)$. Suppose λ_i's and $e_i(\cdot)$'s are eigenvalues and eigenfunctions of K, where the first derivatives of the eigenfunctions exist and are continuous. Also assume that*

$$\sum_{i=1}^{\infty}\lambda_i\int_0^1 t(1-t)|e_i'(t)|^2dt < \infty$$

Then, there exists $X_m(t) = B_m^T(t)\beta_m$ where $B_m(t) = (b_{1m}(t),\ldots,b_{mm}(t))^T$, $b_{km}(t) = \binom{m-1}{k-1}t^{k-1}(1-t)^{m-k}$, and $\beta_m \sim N(\mu_m, D_m)$ for some μ_m and D_m, such that

$$\lim_{m\to\infty} E\|X - X_m\|_2^2 = \lim_{m\to\infty} E\int_0^1 |X(t) - X_m(t)|^2dt = 0,$$

i.e., the sequence of stochastic processes $\{X_m(t) : t \in [0,1], m = 1,2,\ldots\}$ converges in mean square to the stochastic process $\{X(t) : t \in [0,1]\}$.

If we make further assumptions on eigenvalues and eigenfunctions of the covariance function, we can get convergence property in L_p norm for $p \geq 1$.

Theorem 2. *Let $1 \leq p < \infty$. Consider a Gaussian process $X(t)$ defined on $[0,1]$ with continuous mean function $\mu(t)$ and continuous nonnegative definite function $K(t,s)$. Suppose λ_i's and $e_i(\cdot)$'s are eigenvalues and eigenfunctions of K, and continuous derivatives $e_i'(t)$ exists for all i's. Assume that*

(i) $\sum_{i=1}^{\infty} \sqrt{\lambda_i}\{V_p(e_i)\}^{1/p} < \infty$, *where* $V_p(e_i) = \int_0^1 t^{p/2}(1-t)^{p/2}|e_i'(t)|^p dt$, *and*

(ii) $\sum_{i=1}^{\infty} \sqrt{\lambda_i}\|e_i\|_p < \infty$, *where* $\|e_i\|_p = \{\int_0^1 |e_i(t)|^p dt\}^{1/p}$.

Then, there exists $X_m(t) = B_m^T(t)\beta_m$ *where* $B_m(t) = (b_{1m}(t), \ldots, b_{mm}(t))^T$, $b_{km}(t) = \binom{m-1}{k-1} t^{k-1}(1-t)^{m-k}$, *and* $\beta_m \sim N(\mu_m, D_m)$ *for some* μ_m *and* D_m, *such that*

$$\lim_{m\to\infty} E\|X - X_m\|_p = \lim_{m\to\infty} E\{\int_0^1 |X(t) - X_m(t)|^p dt\}^{1/p} = 0, \tag{16}$$

i.e., the sequence of stochastic processes $\{X_m(t) : t \in [0,1], m = 1, 2, \ldots\}$ *converges in* L_p *norm to the stochastic process* $\{X(t) : t \in [0,1]\}$.

Remark: (1) If λ_i's have finitely many non-zero values, condition (i) and (ii) are trivially satisfied. (2) When $p = 2$, condition (ii) can be simplified to $\sum_{i=1}^{\infty} \sqrt{\lambda_i} < \infty$ since $\|e_i\|_2 = 1$. Ritter et al. (1995) showed that $\lambda_i \sim i^{-2r-2}$ asymptotically (as $i \to \infty$) if the covariance function satisfies the Sacks-Ylvisaker condition of order $r \geq 1$. Then, condition (ii) is very likely to be satisfied when Sacks-Ylvisaker condition meets.

Consider the popular squared exponential covariance function defined on $[0,1]$ such that

$$K(t,s) = \exp\{-\frac{(\Phi^{-1}(t) - \Phi^{-1}(s))^2}{2}\}, \qquad t, s \in [0,1], \tag{17}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of a standard normal distribution. Fasshauer and McCourt (2012) showed that the eigenvalues and the orthonormal eigenfunctions are $\lambda_i = \left(\frac{3-\sqrt{5}}{2}\right)^{i+1/2}$ and $e_i(t) = \phi_i(\Phi^{-1}(t))$, for $i = 0, 1, \ldots$, where $\phi_i(t) = \gamma_i \exp\left(-\frac{\sqrt{5}-1}{4}t^2\right) H_i(\sqrt[4]{\frac{5}{4}}t)$, $\gamma_i = \sqrt{5^{1/4}/(2^i i!)}$, and $H_i(x) = (-1)^i \exp(t^2)\frac{d^i}{dt^i}\exp(-t^2)$ (Hermite polynomial). For $p = 1$, Figure 1 shows condition (i) in Theorem 2 is satisfied, and condition (ii) is also satisfied since

$\sqrt{\frac{3-\sqrt{5}}{2}} < 1$ and $\|e_i\|_1 \leq 1$. Thus, with Theorem 2, we can show that there always exists a Bernstein polynomial sieve converges to a Gaussian process with continuous mean function and square exponential covariance function in L_1 norm.

Theorem 1 demonstrates that we can always find a sequence of models based on Bernstein polynomials that converge to the Gaussian process under some mild regularity conditions in L_2 norm. Since convergence in L_2 norm implies convergence in probability and convergence in distribution, this theorem naturally holds for the cases of convergence in probability and distribution of Bernstein polynomials approximation to Gaussian processes. Theorem 2 demonstrates an even stronger consistency of our Bernstein polynomial sieve to Gaussian processes satisfying certain regularity conditions in terms of L_p norm. Moreover, in the proof of Theorem 1 and 2, we have explicitly constructed the Bernstein polynomials estimator where μ_m and D_m preserve the shape of $\mu(\cdot)$ and $K(\cdot, \cdot)$.

Suppose a Gaussian process $X(u)$ is defined on some support set $D \subseteq \mathbb{R}$ other than $[0, 1]$, with mean function $\mu(u)$ and $K(u_1, u_2)$. We can then find a proper invertible transformation function $t = g(u)$ which maps D to $[0, 1]$. Since a Gaussian process is determined by mean and kernel functions, $X(t)$ with mean function $\mu(g^{-1}(t))$ and kernel function $K(g^{-1}(t_1), g^{-1}(t_2))$ is the equivalent Gaussian process defined on $[0, 1]$. Then under regularity conditions on the new mean and kernel functions, $\mu \circ g^{-1}$ and $K \circ g^{-1}$, we can apply Theorem 1 and 2. In this way, Theorem 1 and 2 are naturally extended to Gaussian processes with support sets other than $[0, 1]$.

Our theorems can also be generalized to any other polynomials, since there is a one-to-one mapping between Bernstein basis polynomials of degree $m - 1$ and power basis polynomials of degree $m - 1$. So, instead of Bernstein polynomials, we can use other polynomials such as linear combination of power basis, Hermite polynomials, Laguerre polynomials and Jacobi polynomials, as long as the Gaussian process satisfied the corresponding regularity conditions

listed in Theorem 1 and 2. Moreover, results of Khan (1985) can be used to generalize our results to many other operators (e.g. Feller operator).

4 Simulation Studies

In this section, various simulated data scenarios are used to compare the performance of our proposed model selection criterion, BPDC, with two popular Bayesian model selection criteria, Deviance information criterion (DIC; Spiegelhalter et al. 2002) and log pseudo marginal likelihood (LPML; Geisser and Eddy 1979). Also, a simulation study is conducted to show the approximation to a Gaussian process with nonlinear mean and covariance functions using a sieve of Bernstein polynomials.

4.1 Comparison of Bayesian Model Selection Criteria

DIC is frequently used for model selection and the computation is easy by reusing the output from Markov chain Monte Carlo (MCMC) samples obtained from the posterior distribution. Most statistical softwares provide value of DIC by default such as WinBUGS/OpenBUGS and SAS. DIC can have different definitions if parameters of interests are different. We considered two types of DIC's in our simulation study, conditional DIC and marginal DIC (page 282, Lesaffre and Lawson 2012). The leave-one-out cross-validation based criterion, LPML, is also developed in the frame work of Bayesian analysis, and can be approximately well by using with MCMC samples along with importance sampling. In our simulation study, we compared the performance of BPDC with conditional DIC, marginal DIC and LPML. In accordance with DIC and BPDC, we define LPML $= -2\sum_{i=1}^{I} \log f(y|y_{-i})$, where $f(y_i|y_{-i})$ denotes the posterior predictive density of y_i given y_{-i}.

Data were generated from Model (6). Time points $(t_{i,1}, \ldots, t_{i,J})$ were determined as an arithmetic sequence of length J between 0.01 and 0.98 for all i's. Three scenarios with the

following settings were included in the study. Case 1 is the simplest scenario where the number of observations for each subject is comparably large and random error variance is small. The number of observations for each subject is reduced from 30 to 10 in Case 2, while the random error variance is increased from 0.01 to 0.25. In Case 3, heterogeneous random errors are considered.

- Case 1: m=4, I=50, $J = 30$, $\sigma_i^2 = 0.01$ for all i's, $\beta_0 = (1, 0.5, 0.8, -0.7)^T$, and $D = 0.01\mathbf{I}_4$;

- Case 2: m=3, I=50, $J = 10$, $\sigma_i^2 = 0.25$ for all i's, and

$$\begin{aligned} \beta_0 &= (1, -0.7, 2)^T, \\ D &= \begin{pmatrix} 1 & 0 & 0.2 \\ 0 & 0.7 & 0 \\ 0.2 & 0 & 1.2 \end{pmatrix}; \end{aligned}$$

- Case 3: m=3, I=50, $J = 10$, $\sigma_i^2 \overset{\text{iid}}{\sim} IG(5, 1)$, a distribution with mean 0.25, β_0 and D are set the same as in Case 2.

We computed model selection criteria for models with m ranging from 2 up to $[J^{3/4}]$. To fit Bayesian models, WinBUGS was used to perform the MCMC methods, where the first 3000 samples were dropped as burn-in samples and remaining 3000 samples were used for posterior inference. Table 1 summarizes the percentage of times a model with a given $m \in \{2, \ldots, [J^{3/4}]\}$ is selected based on different criteria. For instance, in Case 1, the (true) model with $m = 4$ was selected 96% by BPDC and models with $m \leq 3$ and $m \geq 6$ were never selected. It clearly shows that BPDC and LPML always select the correct model with the highest percentage. On the contrary in Case 3, DIC's selects the most complex model (largest m) most of the times. This clearly indicates that DIC's potentially lead to selecting over-fitted models in this scenario. Moreover, BPDC outperforms LPML in terms of the higher percentage of model selection decision at the correct model, no matter the

random error term is large or small, homogeneous or heterogeneous. In particular, for Case 1 where we have many observations for every subject, BPDC selects the correct model with a proportion as high as 96% which is 12.5% more than the second best one. Moreover, BPDC selects the incorrect model 9% of the times while the LPML, marginal DIC and conditional DIC selects incorrect models 17.5%, 16.5%, and 36%, respectively for the Case 1 scenario. Similar conclusions can also be drawn for Case 2 and 3. This also demonstrates the superior performance of BPDC compared to DIC's and LPML.

4.2 Empirical Approximation of Nonlinear Gaussian Processes with Bernstein Polynomials

In this subsection, by using BPDC as our model selection criterion, we carried out a simulation study to explore the approximation of a Gaussian process with nonlinear mean and covariance functions using the proposed BPS.

Suppose the true model is Model (1) with $\sigma_\epsilon^2 = 0.01$. Following the real data example presented by Wu and Ding (1999), we chose mean function to be

$$\mu(t) = 0.001\{\exp(12.142 - 6.188t) + \exp(7.624 + 0.488t)\}, \qquad t \in [0, 1],$$

and the true covariance function to be the squared exponential covariance function defined in (17).

We generated 200 data sets, where each data set contains 50 subjects with 10 observations for each subject. Time points are determined as an arithmetic sequence of length 10 between 0.02 and 0.98 for all subjects. We fit Model (6) for every trial with tuning parameter m ranging between 2 and 9, and selected the model with the lowest BPDC. WinBUGS was used to perform the MCMC methods, and the final 3000 out of 6000 MCMC samples were kept for inference.

The performance of the fitted model is evaluated by the integrated total bias for both

mean function and covariance function defined as following:

$$\begin{aligned}
\text{IBias}(\hat{\mu}) &= \int_0^1 \{\mu(t) - \hat{\mu}(t)\}dt = \int_0^1 \mu(t)dt - \int_0^1 \hat{\mu}(t)dt, \\
\text{IBias}(\hat{K}) &= \int_0^1 \int_0^1 \{K(t,s) - \hat{K}(t,s)\}dtds = \int_0^1 \int_0^1 K(t,s)dtds - \int_0^1 \int_0^1 \hat{K}(t,s)dtds,
\end{aligned}$$

where $\hat{\mu}(t) = \sum_{k=1}^m \hat{\beta}_{km} b_{km}(t)$, $\hat{K}(t,s) = \sum_{k_1=1}^m \sum_{k_2=1}^m \hat{D}_{k_1k_2} b_{k_1m}(t) b_{k_2m}(s)$, and $\hat{\beta}_{km}$'s and $\hat{D}_{k_1k_2}$'s are posterior means of model parameters. To calculate $IBias(\hat{\mu})$ and $IBias(\hat{K})$, we just need to compute $\int_0^1 \hat{\mu}(t)dt$ and $\int_0^1 \int_0^1 \hat{K}(t,s)dtds$ since $\int_0^1 \mu(t)dt$ and $\int_0^1 \int_0^1 K(t,s)dtds$ are fixed once the true mean function and covariance function are determined. With the formula of derivatives of a Bernstein polynomial, we can compute the integration of a Bernstein polynomial quickly. For our case $\int_0^1 \hat{\mu}(t)dt = \sum_{k=1}^m \hat{\beta}_{km}/m$ and $\int_0^1 \int_0^1 \hat{K}(t,s)dtds = \sum_{k_1=1}^m \sum_{k_2=1}^m \hat{D}_{k_1k_2}/m^2$. The integration of the true mean and covariance function can be obtained numerically using R, and we have

$$\begin{aligned}
\text{IBias}(\hat{\mu}) &= 31.902 - \sum_{k=1}^m \hat{\beta}_{km}/m, \\
\text{IBias}(\hat{K}) &= 0.577 - \sum_{k_1=1}^m \sum_{k_2=1}^m \hat{D}_{k_1k_2}/m^2.
\end{aligned}$$

Table 2 presents the Monte Carlo means and Monte Carlo standard errors of these integrated biases. The mean integrated bias computed based on 200 trials for the mean function and covariance function, both of which are very close to 0. We also calculated the p-value based on a one-sample t test for testing whether the integrated bias is significantly different from 0. For both mean function and the covariance function, the p-values are larger than 0.1, which demonstrates that there is no significant difference between the true Gaussian process and the approximated Gaussian process using Bernstein polynomials in terms of integrated bias at significance level of 0.1. In Figure 2, the true mean function, the estimated mean function, and the 95% pointwise credible interval are overlaid. Though we used different line color to show them, it is hard to tell them apart since they are so close to each other. Actually,

at every evaluation point the true mean function is always lying between the 95% upper bound and lower bound. This demonstrates that our Bernstein polynomial approximation also works very well in terms of pointwise fit.

5 Berkeley Growth Data Analysis

An interesting data set is the Berkeley growth data which monitor the growth of children from age 1 to age 18. In this study, the heights of 39 males and 54 females were collected at irregularly spaced time points. It is accessible on public domain with data name "growth" in package "fda" in R.

Figure 3 shows the growth curves (heights) of the two groups, female and male. Our goal is to find if there is a significant difference between the growth of females and males in their youth. In this study, both growth curves and growth rates are of interest. In addition, we need to consider the nondecreasing constraint in the mean function since human beings cannot grow shorter during their youth. To model the monotone mean function, we fit the following model for female and male respectively.

$$\begin{aligned} y_{i,j} &= \sum_{k=1}^{m} b_{k,m}(t_{i,j})\beta_{i,k} + \sigma_i \epsilon_{i,j}, \\ \beta_i &\overset{\text{iid}}{\sim} N(\beta_0, D), \\ \epsilon_{i,j} &\overset{\text{iid}}{\sim} N(0,1), \end{aligned} \tag{18}$$

with priors

$$\begin{aligned} \alpha_0 &\sim LogNormal(\mathbf{0}, 100\mathbf{I}), \\ D_0 &\sim InvWishart(100\mathbf{I}, m+2), \\ \sigma_i^2 &\overset{\text{iid}}{\sim} InvGamma(0.01, 0.01), \end{aligned}$$

where $\beta_i = (\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,m})^T$,$\beta_0 = (\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,m})^T$, $\alpha_0 = (\alpha_{0,1}, \alpha_{0,2}, \ldots, \alpha_{0,m})^T$, $\alpha_{0,1} = \beta_{0,1}$, and $\alpha_{0,k} = \beta_{0,k} - \beta_{0,k-1}$ for $k = 2, \ldots, m$. The linear transformation $t = age/20$

was used to convert $age \in [1, 18]$ to $t \in [0.05, 0.9] \subseteq [0, 1]$. Model selection criteria BPDC were computed for $m = 2, \ldots, 14$, shown in Figure 4. It is clear that BPDC reach its minimum at m=7 for the female group and m=14 for the male group. Finally, the prediction based on our fitted model is good demonstrated by a high coefficient of determination,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i,j}(Y_{ij} - \hat{Y}_{ij})^2}{\sum_{i,j}(Y_{ij} - \bar{Y})^2} = 0.999,$$

where $\hat{Y}_{ij}$ denotes the posterior predictive mean of Y_{ij} and $\bar{Y}$ denotes the overall mean.

Figure 5 displays the estimated mean growth curves on the left panel and the difference function along with 95% credible interval on the right panel. Males older than 3 years are significantly taller than females at the same age, and females younger than 3 years appear to be taller than males. Figure 6 shows the estimated growth rates as well as the difference of growth rates between the two groups. It is not surprising to find that females are growing significantly faster than males between age 10 to 14. However, before age 10 and from age 15 to age 18, males grow significantly faster.

6 Discussion

We proposed a class of linear mixed effects model using Bernstein polynomial sieve for fitting longitudinal model under the assumption of Gaussian process. This nonparametric model only requires very limited regularity conditions on the Gaussian process, and can incorporate the shape restriction into the framework. With the proposed model, the derivative of the response process can be easily obtained and its shape property can be retained as well. Convergence properties of our proposed approximation are presented under a set of mild regularity conditions. We also proposed a leave-one-out cross validation based Bayesian model selection criterion, BPDC, which has been shown to have an explicit formula for computation using the important sampling. Simulation studies were carried out to compare BPDC with some of the popular Bayesian model selection criteria, DIC and LPML. It shows

that BPDC outperforms the other two in terms of selecting the correct model out of a class of mixed effects models when a BPS is used. In the real data analysis, we applied our methodology to the growth data of males and females. The nondecreasing constraint of the growth curve was imposed in our proposed model using a linear transformation. The coefficient of determination is demonstrated to be as high as 0.99, for the Berkeley growth study. Interesting findings were revealed in looking at the growth curve estimates as well as the growth rate estimates.

In the future, we would like to explore the longitudinal data subject to missing and censored observations. Also, in our proposed model, we have concentrated on the simple case where predictor z is not allowed to vary with time. We would like to investigate the case where we have a time-varying $Z(t)$ as our predictors.

References

Carnicer, J. and Pena, J. (1993). Shape preserving representations and optimality of the Bernstein basis. *Advances in Computational Mathematics*, 1(2):173–196.

Chiou, J., Müller, H., and Wang, J. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):405–423.

Crainiceanu, C. M. and Goldsmith, A. J. (2010). Bayesian functional data analysis using winbugs. *Journal of Statistical Software*, 32(11):1–33.

Davidian, M. and Giltinan, D. (1995). *Nonlinear models for repeated measurement data*, volume 62. Chapman & Hall/CRC.

Davidian, M. and Giltinan, D. (2003). Nonlinear models for repeated measurement data: an

overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387–419.

Davies, S. L., Neath, A. A., and Cavanaugh, J. E. (2005). Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. *Statistical Methodology*, 2(4):249–266.

Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3(1):458–488.

Fasshauer, G. E. and McCourt, M. J. (2012). Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762.

Gal, S. (2008). *Shape-preserving approximation by real and complex polynomials*. Birkhäuser Boston.

Geisser and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Han, C. and Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with application to hiv dynamics. *Biometrics*, 60(1):25–33.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hoeffding, W. (1971). The l_1 norm of the approximation error for Bernstein-type polynomials. *Journal of Approximation Theory*, 4(4):347–356.

Holsclaw, T., Sansó, B., Lee, H. K., Heitmann, K., Habib, S., Higdon, D., and Alam, U. (2013). Gaussian process modeling of derivative curves. *Technometrics*, 55(1):57–67.

Jones, D. H. (1976). The l_2 norm of the approximation error for Bernstein polynomials. *Journal of Approximation Theory*, 18:305–317.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, 96(456):1272–1298.

Kelisky, R. and Rivlin, T. (1967). Iterates of Bernstein polynomials. *Pacific Journal of Mathematics*, 21(3):511–520.

Khan, R. A. (1985). On the l_p norm for some approximation operators. *Journal of Approximation Theory*, 45:339–349.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. Wiley.

Lindstrom, M. (1995). Self-modelling with random shift and scale parameters and a free-knot spline shape function. *Statistics in Medicine*, 14(18):2009–2021.

Lindstrom, M. and Bates, D. (1988). Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Lorentz, G. G. (1953). *Bernstein polynomials*. Chelsea Publishing Co., New York, second edition.

Morrell, C., Pearson, J., Carter, H., and Brant, L. (1995). Estimating unknown transition

times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association*, 90(429):45–53.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, 92(437):199–207.

Ritter, K., Wasilkowski, G. W., and Woźniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying sacks-ylvisaker conditions. *The Annals of Applied Probability*, 5(2):518–540.

Robinson, G. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32.

Serroyen, J., Molenberghs, G., Verbeke, G., and Davidian, M. (2009). Nonlinear models for longitudinal data. *The American Statistician*, 63(4):378–388.

Solak, E., Murray-Smith, R., Leithead, W., Leith, D., and Rasmussen, C. (2003). Derivative observations in gaussian process models of dynamic systems. In Becker, S., Thrun, S., and Obermayer, K., editors, *Conference on Neural Information Processing Systems*, Advances in neural information processing systems 15. MIT Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–639.

Staicu, A., Crainiceanu, C., and Carroll, R. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11(2):177–194.

Tenbusch, A. (1997). Nonparametric curve estimation with Bernstein estimates. *Metrika*, 45(1):1–30.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468.

Wang, J. and Ghosh, S. (2012). Shape restricted nonparametric regression with Bernstein polynomials. *Computational Statistics & Data Analysis*, 56(9):2729–2741.

Wu, H. and Ding, A. (1999). Population hiv-1 dynamics in vivo: Applicable models and inferential tools for virological data from aids clinical trials. *Biometrics*, 55(2):410–418.

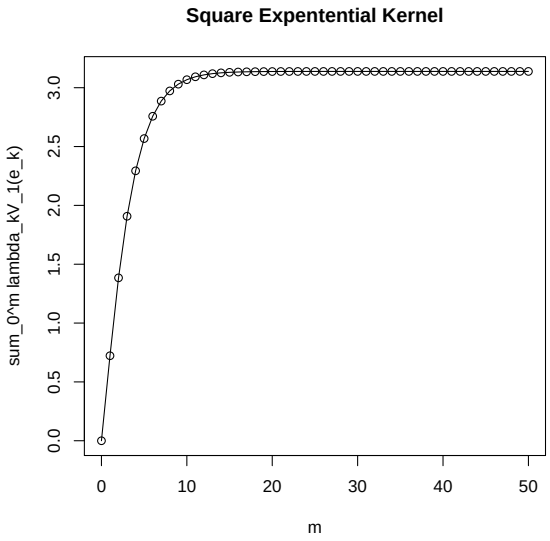

Figure 1: $\sum_{i=0}^{m} \sqrt{\lambda_i} V_1(e_i)$ is plotted against m for the square exponential kernel.

Table 1: Percentage of model selection decisions over 200 repetitions

m	BPDC	LPML	DIC_m	DIC_c
		Case 1 (m=4)		
2	0	0	0	0
3	0	0	0	0
4	**96.0**	**82.5**	**83.5**	**64.0**
5	3.5	11.5	10.0	18.0
6	0	1.5	2.5	4.5
7	0.5	2.0	2.0	6.0
8	0	2.0	1.5	2.5
9	0	0.5	0.5	2.5
10	0	0	0	1.0
11	0	0	0	0.5
12	0	0	0	1.0
13	0	0	0	0
		Case 2 (m=3)		
2	0	0	0	0
3	**76.5**	**58.5**	**59.5**	**43.0**
4	13.0	19.5	19.0	24.0
5	7.0	10.5	11.0	15.0
6	3.5	11.5	10.5	18.0
		Case 3 (m=3)		
2	0	0	0	0
3	**75.0**	**63.0**	26.5	28.5
4	15.0	18.0	19.0	22.5
5	8.0	14.0	20.0	19.5
6	2.0	5.0	**34.5**	**29.5**

[a] BPDC: Bayesian predictive divergence criterion; LPML: log pseudo marginal likelihood; DIC_m: marginal deviance information criterion; DIC_c: conditional marginal deviance information criterion.

[b] The true value of m is specified in the parentheses.

[c] The maximum value of each column for every case is bolded.

Table 2: Fit of Gaussian process with nonlinear mean and covariance functions

IBias($\hat{\mu}$)	SE of IBias($\hat{\mu}$)	p-value	IBias($\hat{K}$)	SE of IBias($\hat{K}$)	p-value
0.0046	0.0077	0.5509	-0.0115	0.0085	0.1776

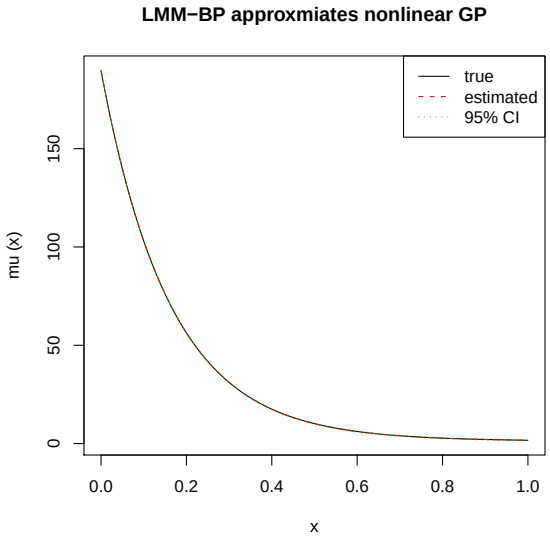

Figure 2: Estimated mean function is plotted in dashed red along with its pointwise 95% credible interval in dashed green lines. The true mean curve is displayed in solid black line.

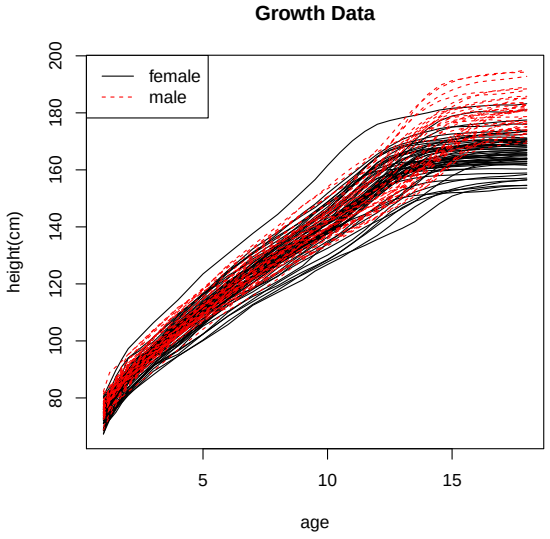

Figure 3: Growth curves of females and males from age 1 to 18. Black: the female group. Red: the male group.

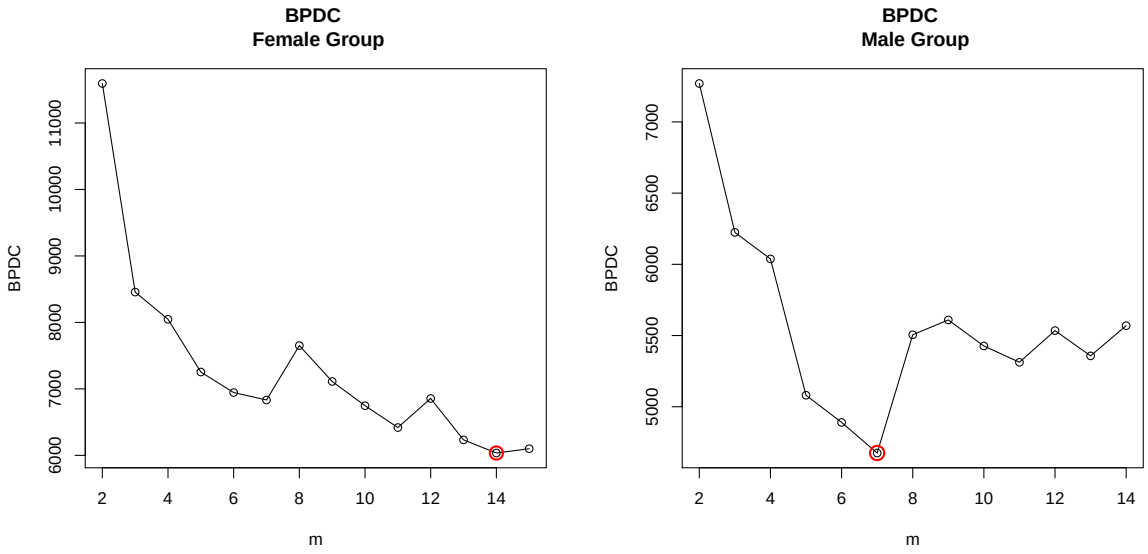

Figure 4: BPDC of models for female and male group.

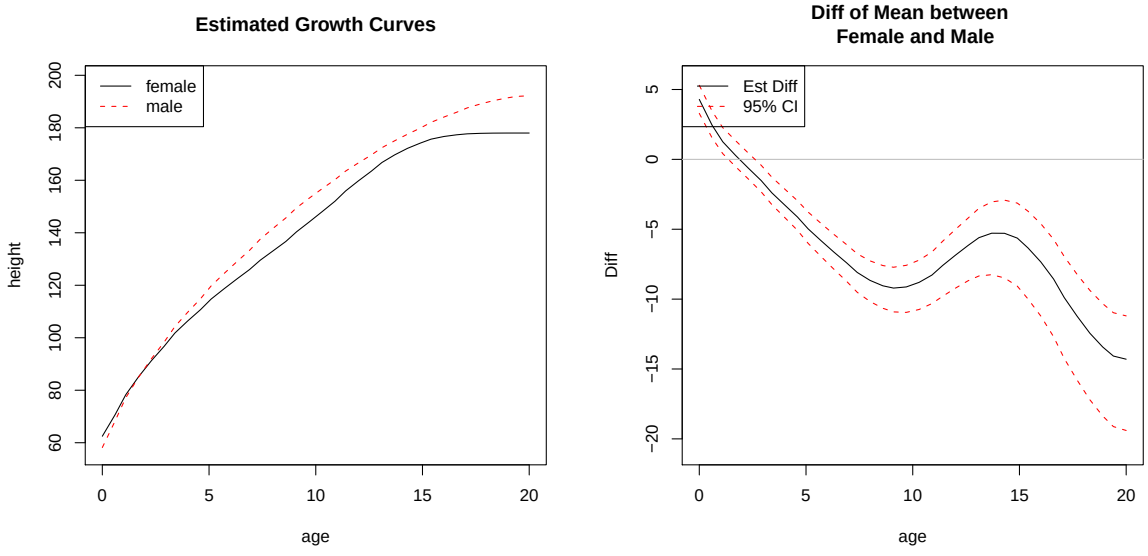

Figure 5: Left: estimated mean growth function of children over ages. Right: estimated difference function between females and males and its 95% credible interval. The grey line is the references line at y=0.

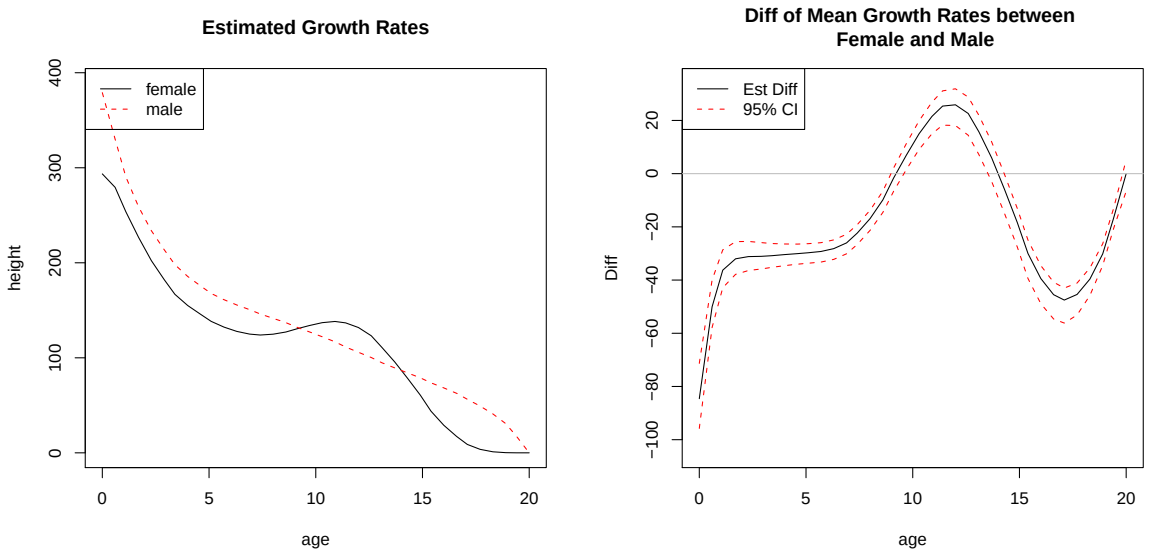

Figure 6: Left: estimated mean function of growth rate over ages. Right: estimated difference function of growth rate and its 95% credible interval and the grey line is the references line at y=0.

Appendices

A Proof of Theorem 1

Proof. By K-L expansion, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and corresponding eigenfunctions $\{e_i(t), i = 1, 2, \ldots\}$ the Gaussian process $X(t)$ can be represented as an infinite series such that $X(t) = \mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i$ in mean square sense, where $Z_i \overset{\text{iid}}{\sim} N(0, 1)$.

$X_m(t)$ has the form of $B_m^T(t)\beta_m$, where $\beta_m \sim N(\mu_m, D_m)$. Thus, we can write it in an equivalent way such that $X_m(t) = B_m^T(t)\mu_m + B_m^T(t)D_m^{1/2}Z$, where $Z = (Z_1, \ldots, Z_m)^T$. By Bernstein Weierstrass approximation theorem, for any continuous function f, there exists $f-$related $m \times 1$ vector $\gamma_m(f) = (f(0), f(\frac{1}{m-1}), \ldots, f(1))^T$ such that $B_m^T(t)\gamma_m(f)$ converges to $f(t)$ on $[0, 1]$ uniformly. We then construct $X_m(t)$ by letting $\mu_m = \gamma_m(\mu)$ and $D_m^{1/2} =$

$(\sqrt{\lambda_1}\gamma_m(e_1), \ldots, \sqrt{\lambda_m}\gamma_m(e_m))$. Then, $X_m(t) = B_m^T(t)\gamma_m(\mu) + \sum_{i=1}^m \sqrt{\lambda_i} B_m^T(t)\gamma_m(e_i)Z_i$. So

$$
\begin{aligned}
& E\|X - X_m\|_2^2 \\
= \; & E[\int_0^1 (X(t) - X_m(t))^2 dt], \\
= \; & \int_0^1 E\{[\mu(t) - B_m^T(t)\gamma_m(\mu)] + [\sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i - \sum_{i=1}^{m} \sqrt{\lambda_i} B_m^T(t)\gamma_m(e_i)Z_i]\}^2 dt, \\
= \; & \int_0^1 [\mu(t) - B_m^T(t)\gamma_m(\mu)]^2 dt \\
+ \; & \int_0^1 E\{\sum_{i=1}^{m} \sqrt{\lambda_i}[e_i(t) - B_m^T(t)\gamma_m(e_i)]Z_i + \sum_{i=m+1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i\}^2 dt, \\
= \; & \int_0^1 [\mu(t) - B_m^T(t)\gamma_m(\mu)]^2 dt \\
+ \; & \int_0^1 \sum_{i=1}^{m} \lambda_i |e_i(t) - B_m^T(t)\gamma_m(e_i)|^2 dt E(Z_i^2) + \int_0^1 \sum_{i=m+1}^{\infty} \lambda_i e_i^2(t) dt E(Z_i^2), \\
= \; & \|\mu - B_m^T\gamma_m(\mu)\|_2^2 + \sum_{i=1}^{m} \lambda_i \|e_i - B_m^T\gamma_m(e_i)\|_2^2 + \sum_{i=m+1}^{\infty} \lambda_i \|e_i\|_2^2.
\end{aligned}
$$

By Bernstein Weierstrass approximation theorem, we have $\|\mu - B_m^T\gamma_m(\mu)\|_2^2 \leq \|\mu - B_m^T\gamma_m(\mu)\|_\infty^2$ goes to 0 as m goes to infinity. Since $\lambda_i\|e_i - B_m^T\gamma_m(e_i)\|_2^2$ is always nonnegative, $\sum_{i=1}^m \lambda_i \|e_i - B_m^T\gamma_m(e_i)\|_2^2 \leq \sum_{i=1}^\infty \lambda_i \|e_i - B_m^T\gamma_m(e_i)\|_2^2 = m^{-1}\sum_{i=1}^\infty \lambda_i m \|e_i - B_m^T\gamma_m(e_i)\|_2^2$. Define $Q_m^2(f) = \int_0^1 \sum_{k=1}^m |f(\frac{k-1}{m-1}) - f(t)|^2 b_{m,k}(t) dt$, where $b_{m,k}(t) = \binom{m-1}{k-1} t^{k-1}(1-t)^{m-k}$. Then since $\sum_{k=1}^m b_{m,k}(t) = 1$,

$$
\begin{aligned}
|e_i - B_m^T\gamma_m(e_i)|^2 &= |\sum_{k=1}^{m} e_i(\frac{k-1}{m-1}) b_{m,k}(t) - e_i(t)|^2, \\
&= |\sum_{k=1}^{m} \{e_i(\frac{k-1}{m-1}) - e_i(t)\} b_{m,k}(t)|^2, \\
&\leq \sum_{k=1}^{m} |\{e_i(\frac{k-1}{m-1}) - e_i(t)\}\sqrt{b_{m,k}(t)}|^2 \sum_{k=1}^{m} |\sqrt{b_{m,k}(t)}|^2, \\
&= \sum_{k=1}^{m} |e_i(\frac{k-1}{m-1}) - e_i(t)|^2 b_{m,k}(t),
\end{aligned}
$$

using Cauchy-Schwarz inequality. Therefore,

$$\begin{aligned}\|e_i - B_m^T\gamma_m(e_i)\|_2^2 &= \int_0^1 |\sum_{k=1}^m e_i(\frac{k-1}{m-1})b_{m,k}(t) - e_i(t)|^2 dt, \\ &\leq \int_0^1 \sum_{k=1}^m |e_i(\frac{k-1}{m-1}) - e_i(t)|^2 b_{m,k}(t)dt = Q_m^2(e_i).\end{aligned}$$

Using Theorem 1 in Khan (1985), we have $\lim_{m\to\infty} m\|e_i - B_m^T\gamma_m(e_i)\|_2^2 \leq \lim_{m\to\infty} mQ_m^2(e_i) = C_2V_2(e_i)$, where $V_2(e_i) = \int_0^1 t(1-t)|e_i'(t)|^2dt$. Then, with assumption that $\sum_{i=1}^\infty \lambda_i V_2(e_i) < \infty$ and Fatou's lemma,

$$\begin{aligned}\lim_{m\to\infty}\sum_{i=1}^\infty \lambda_i m\|e_i - B_m^T\gamma_m(e_i)\|_2^2 &\leq \lim_{m\to\infty}\sum_{i=1}^\infty \lambda_i m Q_m^2(e_i), \\ &= C_2\sum_{i=1}^\infty \lambda_i V_2(e_i) < \infty.\end{aligned}$$

Thus, it follows that $\lim_{m\to\infty}\sum_{i=1}^m \lambda_i\|e_i - B_m^T\gamma_m(e_i)\|_2^2 = 0$. Meanwhile, with continuous covariance function K, we have $\sum_{i=1}^\infty \lambda_i = \int_0^1 K(t,t)dt < \infty$. Therefore, $\lim_{m\to\infty}\sum_{i=m+1}^\infty \lambda_i\|e_i\|_2^2 = \lim_{m\to\infty}\sum_{i=m+1}^\infty \lambda_i = 0$. Altogether, we have proved $\lim_{m\to\infty} E\|X - X_m\|_2^2 = 0$. □

Remark: Notice that in the above proof the theorem can be extended even if $X(t)$ is not assumed to be a Gaussian process. Any second order process can be approximated using $X_m(t)$ as long as we know the distribution of the uncorrelated sequence of Z_i's satisfying $E(Z_i) = 0$ and $Cov(Z_i, Z_{i'}) = \delta_{ii'}$.

B Proof of Theorem 2

Proof. Using Mercer's theorem , we have $X(t) \stackrel{\mathrm{d}}{=} \mu(t) + \sum_{i=1}^\infty \sqrt{\lambda_i}e_i(t)Z_i$, where $Z_i \overset{\text{iid}}{\sim} N(0,1)$. Then we construct $X_m(t)$ in the same way as we do in the proof of Theorem 1, say, $X_m(t) = B_m^T(t)\mu_m + B_m^T(t)D_m^{1/2}Z$, where $Z = (Z_1, \ldots, Z_m)^T$, $\mu_m = \gamma_m(\mu)$ and $D_m^{1/2} =$

$(\sqrt{\lambda_1}\gamma_m(e_1), \ldots, \sqrt{\lambda_m}\gamma_m(e_m))$. Then,

$$\begin{aligned} & E\|X - X_m\|_p \\ = \; & E\{\|\mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i - B_m^T(t)\gamma_m(\mu) - \sum_{i=1}^{m} \sqrt{\lambda_i} B_m^T(t)\gamma_m(e_i) Z_i\|_p\}, \\ \leq \; & \|\mu - B_m^T\gamma_m(\mu)\|_p + \sum_{i=1}^{m} \sqrt{\lambda_i}\|e_i - B_m^T\gamma_m(e_i)\|_p E|Z_i| + \sum_{i=m+1}^{\infty} \sqrt{\lambda_i}\|e_i\|_p E|Z_i|. \end{aligned}$$

Since $Z_i \sim N(0,1)$, $E|Z_i| = \sqrt{\frac{\pi}{2}}$. By Bernstein Weierstrass approximation theorem, we have $\lim_{m\to\infty} \|\mu - B_m^T\gamma_m(\mu)\|_p = 0$. Define $Q_m^p(f) = \int_0^1 \sum_{k=1}^m |f(\frac{k-1}{m-1}) - f(t)|^p b_{m,k}(t)dt$, where $b_{m,k}(t) = \binom{m-1}{k-1} t^{k-1}(1-t)^{m-k}$. Theorem 1 in Khan (1985) implies that $\lim_{m\to\infty} \sqrt{m}\{Q_m^p(e_i)\}^{1/p} = C_p\{V_p(e_i)\}^{1/p}$ for some constant C_p which only depends on p. Because $E|X| \leq \{E|X|^p\}^{1/p}$ for $p \geq 1$, we have $|E(X)|^p \leq E|X|^p$. So, $|\sum_{k=1}^m \{e_i(\frac{k-1}{m-1}) - e_i(t)\} b_{m,k}(t)|^p \leq \sum_{k=1}^m |f(\frac{k-1}{m-1}) - e_i(t)|^p b_{m,k}(t)dt$, since $\sum_{k=1}^m b_{m,k}(t) = 1$. Hence,

$$\begin{aligned} \|e_i - B_m^T\gamma_m(e_i)\|_p &= \{\int_0^1 |\sum_{k=1}^m e_i(\frac{k-1}{m-1}) b_{m,k}(t) - e_i(t)|^p dt\}^{1/p}, \\ &\leq \{\int_0^1 \sum_{k=1}^m |e_i(\frac{k-1}{m-1}) - e_i(t)|^p b_{m,k}(t)dt\}^{1/p}, \\ &= \{Q_m^p(e_i)\}^{1/p}. \end{aligned}$$

Then,

$$\begin{aligned} \lim_{m\to\infty} \sum_{i=1}^{m} \sqrt{\lambda_i}\|e_i - B_m^T\gamma_m(e_i)\|_p &\leq \lim_{m\to\infty} m^{-1/2} \sum_{i=1}^{m} \sqrt{\lambda_i}\sqrt{m}\{Q_m^p(e_i)\}^{1/p}, \\ &\leq \lim_{m\to\infty} m^{-1/2} \sum_{i=1}^{\infty} \sqrt{\lambda_i}\sqrt{m}\{Q_m^p(e_i)\}^{1/p}. \end{aligned}$$

Hence, with condition (i) such that $\sum_{i=1}^{\infty} \sqrt{\lambda_i}\{V_p(e_i)\}^{1/p} < \infty$, we have $\lim_{m\to\infty} \sum_{i=1}^{\infty} \sqrt{\lambda_i}\sqrt{m}\|e_i - B_m^T\gamma_m(e_i)\|_p \leq \sum_{i=1}^{\infty} \sqrt{\lambda_i}\{V_p(e_i)\}^{1/p} < \infty$. Therefore, $\lim_{m\to\infty} \sum_{i=1}^{m} \sqrt{\lambda_i}\|e_i - B_m^T\gamma_m(e_i)\|_p = 0$. Together with condition (ii) such that $\sum_{i=1}^{\infty} \sqrt{\lambda_i}\|e_i\|_p < \infty$, we have proved $\lim_{m\to\infty} E\|X - X_m\|_2^2 = 0$. □