ABSTRACT

WANG, LIWEI. Nonparametric Models for Longitudinal Data Using Bernstein Polynomial Sieve. (Under the direction of Sujit K. Ghosh.)

Analysis of longitudinal data within a mixed model framework becomes a challenging task when observations are subject to data irregularities like censoring and missing values. Often finite dimensional (parametric) models are found inadequate to address the complex relationship between the response and predictors. A majority of the currently available models and associated estimation methodologies are based on restrictive assumptions on the correlation structure of longitudinal data. To begin with we develop a flexible class of models based on Bernstein polynomial sieve with varying degrees and propose a model fitting mechanism assuming fully observed data. Next, we develop a new Bayesian model selection criteria based on a predictive divergence criteria which enables the selection of size of the Bernstein polynomial sieve. Various simulated data scenarios are used to illustrate the superior performance of the proposed estimation methodology and associated model selection criteria. We then extend the estimation methodology and model selection to accommodate the data irregularities using a Markov Chain Monte Carlo based approach. The newly proposed models and associated inference methodologies are illustrated using real data.
Nonparametric Models for Longitudinal Data Using Bernstein Polynomial Sieve

by

Liwei Wang

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

2013

APPROVED BY:

__________________________  _________________________
Brian Reich                Wenbin Lu

__________________________  _________________________
Arnab Maity                 Sujit K. Ghosh
Chair of Advisory Committee
DEDICATION

To my families.
BIOGRAPHY

The author was born in Hengyang, China. She received her Bachelor of Science degree in Statistics from Sun Yat-Sen University, Guangzhou, China in July 2009. After that, she arrived at the North Carolina State University in August 2009 to pursue a doctoral degree in the Department of Statistics. Her research focuses on nonparametric models for longitudinal study and Bayesian model selection.
ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude and appreciation to my advisor, Dr. Sujit K. Ghosh. He showed great patience, motivation, and enthusiasm in guiding me in exploring interesting research topics. Without his expertise and illuminating guidance, I would not have completed the research work.

I wish to thank my committee members: Drs. Brian Reich, Wenbin Lu, Arnab Maity and Denis S. Jackson for their valuable helps and insight comments. They raised lots of interesting questions and precious suggestions, which spurred me on to greater effort. Also, their positive feedback has greatly encouraged me to pursue more achievements.

My thanks also go to Drs. Pam Arroway, Jacqueline Hughes-Oliver, John Monahan and Sujit K. Ghosh for their brilliant work as Directors of Graduate Program. Drs. Pam Arroway and Jacqueline Hughes-Oliver approved my application and offered me the Teaching Assistant Scholarship. Drs. John Monahan and Sujit K. Ghosh helped a lot during my years’ stay in the program, always being patient in answering questions and taking time to discuss about study and life. Besides, I want to thank all the faculty and the staff at Department of Statistics. I feel so lucky to be part of the prestigious department.

I am also thankful for the opportunity of working with Cheryl LeSaint and all colleagues at SAS Institute Inc. They have taught me a lot on how to work in an organized and efficient way with all kinds of statistical methods. It has been an invaluable experience for me to work in a big company and collaborate with experts from different areas.

Last but not the least, I would like to thank my family and friends for their endless love and support. They make my school life at NCSU full of laughter and happiness,
make me believe in myself and never give up.
# TABLE OF CONTENTS

## LIST OF TABLES

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>v</td>
</tr>
</tbody>
</table>

## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ix</td>
</tr>
</tbody>
</table>

## Chapter 1 Mixed Effects Models for Longitudinal Study

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Gaussian Processes</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>Parameter Estimation using Gaussian Process Models</td>
<td>10</td>
</tr>
<tr>
<td>1.4</td>
<td>Organization of the Dissertation</td>
<td>13</td>
</tr>
</tbody>
</table>

## Chapter 2 Gaussian Processes Approximation with Bernstein Polynomials

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Bernstein Polynomials</td>
<td>16</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Definition and Convergency Properties</td>
<td>17</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Shape-preserving Properties</td>
<td>22</td>
</tr>
<tr>
<td>2.3</td>
<td>A Class of Linear Mixed Effects Models with Bernstein Polynomial Sieve</td>
<td>23</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Approximation of Gaussian Processes with LMM-BPS</td>
<td>23</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Priors and Starting Values</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>Convergence Properties</td>
<td>26</td>
</tr>
<tr>
<td>2.5</td>
<td>Discussion</td>
<td>34</td>
</tr>
</tbody>
</table>

## Chapter 3 Bayesian Model Selection using Predictive Divergence

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Bayesian Model Selection Criteria</td>
<td>37</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Deviance Information Criterion</td>
<td>40</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Log Pseudo Marginal Likelihood</td>
<td>42</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Bayesian Predictive Divergence Criterion</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Simulation Study</td>
<td>48</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Data Generation: Design and Setup</td>
<td>49</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Results and Discussion</td>
<td>52</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Supplementary Simulation</td>
<td>55</td>
</tr>
<tr>
<td>3.4</td>
<td>Real Data Analysis</td>
<td>61</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Growth Curve Study using Sitka Data</td>
<td>62</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Analysis of Berkeley Growth Study</td>
<td>71</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>73</td>
</tr>
</tbody>
</table>

## Chapter 4 Longitudinal Analysis Subject to Data Irregularity

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>75</td>
</tr>
</tbody>
</table>

v
4.2 Modified Bayesian Model Selection Criteria for Missing and Censored Data
   4.2.1 Modified DIC ............................................. 80
   4.2.2 Modified LPML ......................................... 81
   4.2.3 Modified BPDC ........................................... 82
4.3 Simulation Study .............................................. 84
   4.3.1 Data Generation: Design and Setup ....................... 84
   4.3.2 Results and Discussion ................................ 87
4.4 Real Data Study ............................................... 88
   4.4.1 The Schizophrenia Data with Missing Values ............ 95
   4.4.2 The ACTG 398 Study with Censored Values ............. 100
4.5 Discussion .................................................. 106

Chapter 5 Discussions and Future Work .................................. 107

REFERENCES ..................................................... 109
LIST OF TABLES

Table 3.1 Percentage of model selection decisions (Case 1-4) .......................... 56
Table 3.2 Percentage of model selection decisions (Case 5) .............................. 57
Table 3.3 Average of model selection criteria and standard errors (Case 1-3) 58
Table 3.4 Average of model selection criteria and standard errors (Case 4-5) 59
Table 3.5 Fit of Gaussian process with nonlinear mean and covariance functions 61
Table 3.6 BPDC of models on Sitka data ...................................................... 67

Table 4.1 Percentage of model selection decisions for missing data models (Case 1-3) 89
Table 4.2 Average of model selection criteria and standard errors for missing data models (Case 1-2) ................................................................. 90
Table 4.3 Average of model selection criteria and standard errors for missing data models (Case 3) ................................................................. 91
Table 4.4 Percentage of model selection decisions for censored data models (Case 1-3) ................................................................. 92
Table 4.5 Average of model selection criteria and standard errors for censored data models (Case 1-2) ................................................................. 93
Table 4.6 Average of model selection criteria and standard errors for censored data models (Case 3) ................................................................. 94
Table 4.7 Sample size across weeks of the schizophrenia study ....................... 96
Table 4.8 BPDC of models for schizophrenia study ........................................ 98
Table 4.9 BPDC of models for ACTG 398 study ............................................ 103
LIST OF FIGURES

Figure 1.1  Left: the growth curves of Sitka spruce trees. Right: the HIV viral load across weeks, where the light blue shaded area highlights those unreliable measurements. ................................. 2

Figure 1.2  First row: covariance functions; (1) \(K_{SE}(t,s) = \exp\{- (t-s)^2/2\}\), (2) \(K_S(t,s) = \exp\{- \sin^2(\pi|t-s|)/2\}\), and (3) \(K_{BB}(t,s) = \min(t,s)+ts\). Second row: three random functions drawn from corresponding centered Gaussian processes. ................................. 11

Figure 2.1  Bernstein basis polynomials of degree 2 and degree 7 ..................... 17

Figure 2.2  \(\sum_{i=0}^{m} \sqrt{\lambda_i}V_1(e_i)\) is plotted against \(m\) for the square exponential kernel. 35

Figure 3.1  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. ................................. 62

Figure 3.2  Growth curves of Sitka spruce trees over days. Black: the ozone group. Red: the control group. ................................. 63

Figure 3.3  BPDC of models for ozone and control group. ................................. 66

Figure 3.4  Left: estimated mean function over days, and original observations displayed to show the fit of the model. Right: estimated difference function and its 95% credible interval where the difference function is the mean function of ozone group minus the mean function of control group and the grey line is the reference line at \(y=0\). The fixed treatment effect obtained by Crainiceanu et al. (2005) is shown in blue dashed line at \(y=0.31\). ................................. 68

Figure 3.5  Estimated covariance function over days. Left: ozone group. Right: control group. ................................. 69

Figure 3.6  Left: estimated mean function of growth rate over days. Right: estimated difference function of growth rate and its 95% credible interval and the grey line is the reference line at \(y=0\). ................................. 70

Figure 3.7  Growth curves of females and males from age 1 to 18. Black: the female group. Red: the male group. ................................. 72

Figure 3.8  BPDC of models for female and male group. ................................. 72

Figure 3.9  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 \(y=0\). ................................. 73

Figure 3.10 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 \(y=0\). ................................. 74
Figure 4.1 Severity scores across weeks for the first 8 patients in each group. Solid segments are used to link two measurements in the neighborhood, and dashed segments are used to link two measurements with missing values in the middle.

Figure 4.2 Left: estimated mean function over weeks, and original group means by week displayed to show the fit of the model. Right: estimated difference function between placebo group and drug group along with its 95% credible interval and the reference line in grey at y=0.

Figure 4.3 Estimated covariance functions over weeks. Left: placebo group. Right: drug group.

Figure 4.4 Left: estimated first derivative of mean function over weeks. Right: estimated difference of the first derivative function between placebo group and drug group along with its 95% credible interval.

Figure 4.5 Log10 of HIV viral load (copies/ml) across weeks. The SPI group is plotted in solid black line, while DPI group is plotted in dashed red line. Horizontal reference line in blue is plotted at log_{10}(200) copies/ml where measurements below the line is inaccurate. The light blue shaded area highlights those measurements which are not reliable.

Figure 4.6 Estimated mean functions over weeks. Left: NNTRI-naive group; Right: NNTRI-experienced group.

Figure 4.7 Estimated differences between mean functions of SPI and DPI groups over days. Left: NNTRI-naive group; Right: NNTRI-experienced group.

Figure 4.8 Estimated covariance functions. Left: SPI group; Right: DPI group.
Chapter 1

Mixed Effects Models for Longitudinal Study

1.1 Introduction

In a wide variety of disciplines such as agriculture, biology, business, epidemiology, medicine, and social science, data are collected repeatedly over time on randomly selected subjects. The analysis of such repeated measurements data is known as longitudinal study. One motivating example is the Sitka spruce tree study which was first introduced and analyzed in Diggle et al. (1994). The log-sizes of spruce trees over two years at 13 irregularly spaced time points were recorded, where spruce trees were randomly assigned to the treatment group (ozone-enriched environment) and the control group (natural environment). The left panel of Figure 1.1 presents the log-sizes of the trees against the days of measurements. It is of interest to find if the environment (ozone-enriched versus natural) has any significant effect on the growth of the Sitka spruce trees. Solely based on the Figure 1.1, it is hard to tell if there is a significant difference between the two groups.
One interesting feature about this data set is that the growth curve must necessarily be nondecreasing and such shape constraints must be accommodated in statistical models. Some other longitudinal data have spontaneous shape restrictions as well, such as the hearing loss study analyzed in Davidov and Rosen (2011). The ACTG 398 data analyzed by Sun and Wu (2005) is another typical longitudinal data set which involves about 22% left-censored observations. In the right panel of Figure 1.1, it is clear that the curves tangle in a mess, where no obvious pattern can be identified simply by looking at the plot. In practice, longitudinal data are usually subject to data irregularities such as missing, censored and truncated values. This requires suitable methodologies to handle the missing/censoring/truncating mechanism, and appropriate analysis is needed for making probabilistically valid statistical inference. The two examples described above are typical real data scenarios, and in general such longitudinally observed data can be modelled within a general statistical framework.

![Sitka Data](image1.png) ![ACTG 398 Study](image2.png)

Figure 1.1: Left: the growth curves of Sitka spruce trees. Right: the HIV viral load across weeks, where the light blue shaded area highlights those unreliable measurements.
For each randomly sampled subject \( i \), let \( Y_i(t) \) denote the measured response obtained at time \( t \in [0, T] \) for \( i = 1, \ldots, I \). Let \( Z_i = (Z_{i1}, \ldots, Z_{ip})^T \) denote a vector of baseline covariates measured at time point \( t = 0 \) for the subject \( i \). More generally, we may also have time-varying covariates measured at time \( t \) denoted as \( Z_i(t) \) and the goal of the study would be exploring the relationship between \( Y(t) \) and \( Z(t) \). However, for the sake of simplicity, we first assume only baseline covariates are included in the study. For example, for the Sitka spruce trees study, we have \( p = 1 \) and the binary covariate \( Z_i \) indicating an ozone-enriched environment or natural environment. One may start with postulating the following statistical model:

\[
Y_i(t) = Z_i^T \beta + X_i(t) + \epsilon_i(t), \quad i = 1, \ldots, I, \tag{1.1}
\]

where \( \{X_i(t) : t \in [0, T]\} \) are independently and identically distributed (iid) as the stochastic process \( \{X(t) : t \in [0, T]\} \) and \( \{\epsilon_i(t) : t \in [0, T]\} \) are iid as a white noise process. More precisely, we assume \( E[\epsilon_i(t)] = 0 \) for any \( i \) and any \( t \) and \( Cov[\epsilon_i(t), \epsilon_j(t)] = \sigma^2 \delta_{ij} \delta_{ts} \) where \( \delta_{ab} = 1 \) if \( a = b \) and \( \delta_{ab} = 0 \) otherwise for \( a, b \in \mathbb{R} \). One of the main goals of the study is to model the latent stochastic process \( \{X(t) : t \in [0, T]\} \).

One of the challenging aspects of estimating model (1.1) is that the response process \( Y_i(t) \) is not observed for all time points \( t \in [0, T] \) and the collection of time points could also vary with the subject \( i \). Such feature of the data prohibits the use of standard off-the-shelf multivariate analysis for growth curve models. In general, we may observe \( Y_i(t) \) measured only at a selected set of time points \( t_{i1} < t_{i2} < \ldots < t_{iJ_i} \) where \( J_i \geq 2 \) for \( i = 1, \ldots, I \). Let \( Y_{ij} = Y_i(t_{ij}) \) for \( i = 1, \ldots, I \) and \( j = 1, \ldots, J_i \), and the observed data can be presented as \( \mathbb{D} = \{(Y_{ij}, t_{ij}, Z_i) : i = 1, \ldots, I; j = 1, \ldots, J_i\} \). In such case, we can
postulate the following model:

$$Y_{ij} = Z_i^T \beta + X_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$  \hspace{1cm} (1.2)$$

where $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma^2)$ and the goal would be estimating parameter of the latent stochastic process $\{X(t) : t \in [0, T]\}$ and $(\beta, \sigma^2)$ based on the data $\mathbb{D}$. Parametric approaches assume that $X_i = (X_{i1}, \ldots, X_{ij})^T = U_ib_i$, where $U_i = (U_{i1}, \ldots, U_{iq})$ denotes a $J_i \times q$ matrix designed to capture the trend and $b_i$ denotes random effects to capture the heterogeneity across subjects. This reduces the model (1.2) to the following linear mixed effects model (LMM),

$$Y_i = Z_i^T \beta + U_ib_i + \epsilon_i, \quad i = 1, \ldots, I.$$  \hspace{1cm} (1.3)$$

Following common practice, we assume $b_i \overset{iid}{\sim} N(\mathbf{0}, \Sigma_b)$ and $\epsilon_i \overset{ind}{\sim} N(\mathbf{0}, \sigma^2 I)$. Many standard software packages are available (e.g. PROC MIXED in SAS, lme in R, etc.) which allows the estimation of the parameters $\theta = (\beta, \Sigma_b, \sigma^2)$ using maximum likelihood (ML; Robinson 1991), restricted maximum likelihood (REML; Patterson and Thompson 1971; Harville 1977), or EM algorithm (Lindstrom and Bates 1988).

Clearly, the models described in (1.2) and (1.3) are based on the linearity assumption that relates $Y_i$’s to $Z_i$’s and $U_i$’s, which usually provides a very good first order approximation. Model (1.3) has been generalized to approximate more complex relationship between the responses ($Y_i$’s), baseline covariates ($Z_i$’s) and trend variables ($U_i$’s). More generally, we may consider the following so called nonlinear mixed effects model (NLMM),

$$Y_{ij} = g(Z_i, U_{ij}, \beta, b_i) + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$  \hspace{1cm} (1.4)$$
where \( g(\cdot) \) is a known function, \( b_i \)'s are subject specific random effects to capture variations across subjects, and \( \beta \) is a vector of fixed effects capturing the baseline effects. As in model (1.3) we have similar assumptions regarding the \( b_i \)'s and \( \epsilon_i \)'s to be independent to each other for \( i = 1, \ldots, I \). See Davidian and Giltinan (1995), Davidian and Giltinan (2003), and Serroyen et al. (2009) for comprehensive reviews on NLMM with various applications in practice. When we can identify the form of \( g(\cdot) \) from the mechanistic theory, the parametric NLMM is employed. Some typical applications include the pharmacokinetics and pharmacodynamic modeling (Davidian and Giltinan 1995, Chapter 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 semiparametric 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-node spline shape function. But the resulting covariance structure fitted in this way is not easy to interpret.

In general, although model (1.4) provides a greater flexibility in capturing the possible relationships between the responses (\( Y_i \)'s), baseline covariates (\( Z_i \)'s) and trend variables (\( U_i \)'s), the model is still based on an assumed nonlinear functional form which may only be a crude approximation to the true relationship between the response and predictors. Moreover, in practice the correlation function is often modelled parametrically. One of the goals of this dissertation is to somewhat relax the assumed relationship between the response and predictors and also avoid making strong parametric assumptions. To begin
with, we focus on modeling the latent stochastic process $X(t)$ in model (1.1). A widely used model to approximate a continuous stochastic process is given by the Gaussian process which is characterized by specifying a mean function and a covariance function. Next we provide a brief review of Gaussian processes in Section 1.2 and a class of nonlinear mixed effects models based on Gaussian processes (NLMM-GP) is introduced in Section 1.3. Finally, in Section 1.4, the organization of the whole dissertation is outlined.

### 1.2 Gaussian Processes

Gaussian distribution is one of the well-known distributions in statistics and has been widely used in many practical scenarios involving measured responses subject to random errors. A finite dimensional vector $X = (X_1, \ldots, X_m)^T$ is said to have a multivariate Gaussian distribution if a linear function $a_1X_1 + \ldots + a_mX_m$ has a univariate Gaussian distribution for any $a = (a_1, \ldots, a_m)^T \in \mathbb{R}^m$ (Casella and Berger 2001, Chapter 4).

Extending this definition to a stochastic process $\{X(t); t \in [0, T]\}$, we would say that it is a Gaussian process if every finite dimensional vector $(X(t_1), \ldots, X(t_m))^T$ has a multivariate Gaussian distribution for any finite set of distinct $t_1, \ldots, t_m \in [0, T]$. It can be shown that a Gaussian process is characterized by its mean function $\mu(t) = E[X(t)]$ and covariance function $K(t, s) = Cov(X(t), X(s))$.

**Definition 1.2.1.** A stochastic process $\{X(t); t \in [0, T]\}$ is said to be a Gaussian process with mean function $\mu(t)$ and covariance function $K(t, s)$ if for any finite collection $t_1, \ldots, t_m \in [0, T]$, the vector $(X(t_1), \ldots, X(t_m))^T$ has a multivariate Gaussian distribution with mean $(\mu(t_1), \ldots, \mu(t_m))^T$ and covariance matrix $K = (K(t_i, t_j))_{i=1, \ldots, m; j=1, \ldots, m}$. 
We denote this stochastic process by the notation

\[ X(\cdot) \sim GP(\mu(\cdot), K(\cdot, \cdot)). \]

Rasmussen and Williams (2006) demonstrated that Gaussian processes can be used to approximate many well-known statistical models, including Bayesian linear models, spline models, large neural networks (under suitable conditions), and are closely related to other popular models, such as support vector machines. The linear mixed effects model (1.3) is also a special case of Gaussian processes. Assuming random variables are drawn from a Gaussian process always makes the model easy to handle and interpret. “Under the Gaussian process viewpoint, the models may be easier to handle and interpret than their conventional counterparts, such as e.g. neural networks.” (Rasmussen and Williams 2006). As Gaussian distribution can approximate many real world random phenomenon, the widespread use of Gaussian processes is always considered reasonable for longitudinal analysis, spatial analysis, and computer experiments analysis, among others. So we begin with the assumption that \( X_i(\cdot) \overset{iid}{\sim} GP(\mu(\cdot), K(\cdot, \cdot)) \) within any modeling framework as described in (1.1) One of the main stumbling block of this generalization is that we need to estimate the mean function \( \mu(\cdot) \) and the covariance function \( K(\cdot, \cdot) \) based on the observed data \( \mathbb{D} = \{(Y_{ij}, t_{ij}, Z_i) : i = 1, \ldots, I; j = 1, \ldots, J_i\} \), where \( Z_i = (Z_{i1}, \ldots, Z_{ip})^T \).

The mean function can be assumed to belong to a suitable class of smooth function (e.g. continuous functions on \([0, T]\), etc). But the covariance function needs to be a non-negative definite (nnd) function, also known as a valid kernel.

**Definition 1.2.2.** A function \( K: [0, T] \times [0, T] \to \mathbb{R} \) is said to be a non-negative definite function
(nnd) function if and only if \( K(t, s) = K(s, t) \) for any \( s, t \in [0, T] \) and

\[
\sum_{i,j=1}^{m} a_i a_j K(t_i, t_j) \geq 0 \tag{1.5}
\]

for any set of points \( \{t_1, \ldots, t_m\} \) in \([0, T]\) and for any set of real numbers \( \{a_1, \ldots, a_m\} \).

A nnd function can be characterized by its spectral decomposition by extending the idea of spectral decomposition of a nnd matrix. The characterization is given by the celebrated Mercer’s Theorem (restated by Minh et al. 2006).

**Theorem 1.2.1. Mercer’s Theorem.** Let \( K \) be a continuous nonnegative definite function on \([0, T] \times [0, T]\) satisfying

\[
\int_0^T \int_0^T \int \ K^2(s, t) dw(s) dw(t) < \infty, \tag{1.6}
\]

where \( w(\cdot) \) is an absolutely continuous weight function defined on \([0, T]\). Then, there exists a sequence of orthogonal functions \( \{e_k(t); k = 1, 2, \ldots; t \in [0, T]\} \) satisfying

\[
\int_0^T e_k(t) e_l(t) dw(t) = \delta_{kl}
\]

and a sequence of nonnegative real numbers \( \{\lambda_k : \lambda_k \geq 0; k = 1, 2, \ldots\} \) such that

\[
K(s, t) = \sum_{k=1}^{\infty} \lambda_k e_k(s) e_k(t), \tag{1.7}
\]

where the infinite series converges absolutely for each pair \((s, t) \in [0, T]^2\) and uniformly on each compact subset of \([0, T]^2\).

Remark: \( \{e_k(t); k = 1, 2, \ldots\} \) are called the eigenfunctions of \( K \), and \( \{\lambda_k; k = 1, 2, \ldots\} \)
are known as the eigenvalues of $K$.

Eigenvalues and eigenfunctions defined in Theorem 1.2.1 can be obtained by solving the integral equations $\int_0^T K(t, s)e(s)dw(s) = \lambda e(t)$ and $\int_0^T e^2(t)dw(t) = 1$. For example, the Wiener process defined on $[0, 1]$ has the kernel $K(t, s) = \min(t, s)$, whose eigenfunctions and eigenvalues are $\{e_k(t) = \sqrt{2}\sin((k - \frac{1}{2})\pi t); k = 1, \ldots\}$ and $\{\lambda_k = [(k - \frac{1}{2})^2\pi^2]^{-1}; k = 1, \ldots\}$, when the weight function is the Lebesgue measure. Also in virtue of Mercer’s Theorem, Loève (1945) and Karhunen (1947) introduced Karhunen-Loève (K-L) expansion of a stochastic process which allows us to represent a stochastic process using a countable number of real numbers and sequence of orthogonal functions.

**Theorem 1.2.2.** Karhunen-Loève (K-L) expansion of a stochastic process. Let $X(t)$ be a second-order stochastic process defined over a probability space $(\Omega, F, P)$ and indexed over a closed and bounded interval $[0, T]$ with continuous mean function $\mu(\cdot)$ and covariance function $K(\cdot, \cdot)$ satisfying the assumptions in Theorem 1.2.1. Then, $X(t)$ has a (quadratic mean) representation in terms of $L_2$ norm.

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \sqrt{\lambda_k}e_k(t)Z_k,$$  \hspace{1cm} (1.8)

where the random variables $Z_k$’s satisfying $\sqrt{\lambda_k}Z_k = \int_0^T X(t)e_k(t)dw(t)$ have zero-mean, and are uncorrelated with variance 1. In other words, $E[Z_k] = 0$ and $\text{Cov}(Z_k, Z_l) = \delta_{kl}$ for any $k$ and $l$.

Remark: If we further assume that $X(\cdot) \sim GP(\mu(\cdot), K(\cdot, \cdot))$, then random variables $\{Z_k; k = 1, 2, \ldots\}$ are iid as standard normal distribution, i.e. $Z_k \overset{\text{iid}}{\sim} N(0, 1)$.

There are several popular kernel functions including squared exponential kernel $K_{SE}(t, s) = \sigma^2 \exp(-|t - s|^2)$, exponential kernel $K_E(t, s) = \sigma^2 \exp(-|t - s|)$, seasonal
The kernel $K_S(t, s) = \sigma^2 \exp\{-k \sin^2(\pi|t - s|)\}$, and the kernel to a centered Brownian bridge $K_{BB}(t, s) = \min(t, s) + ts$, among others. Figure 1.2 displays three valid kernel functions and three random functions drawn from the corresponding $GP(0, K(\cdot, \cdot))$. It shows that with various kernel functions we can mimic a variety of shapes for random functions.

Another interesting fact about Gaussian process is that if the second order derivative $\frac{\partial^2 K(t, s)}{\partial t \partial s}$ exists for all pairs of points $(s, t) \in [0, T] \times [0, T]$, then the derivative of a Gaussian process is still a Gaussian process. This is true because differentiation is a linear operation (Solak et al. 2003). Suppose $X(t) \sim GP(\mu(\cdot), K(\cdot, \cdot))$ where $\mu(\cdot)$ and $K(\cdot, \cdot)$ are differentiable, then the derivative of $X(t)$ also forms a Gaussian process, $X'(t) \sim GP(\mu^*(\cdot), K^*(\cdot, \cdot))$, where $\mu^*(t) = \mu'(t)$ and $K^*(t, s) = \frac{\partial^2 K(t, s)}{\partial t \partial s}$. This property enables us to analyze the derivative of a Gaussian process, e.g., the growth rate (derivative of the growth curve). One the other side, we can also take advantage of the fact when we have the derivative observations, e.g., the perturbation data (Solak et al. 2003).

1.3 Parameter Estimation using Gaussian Process Models

In order to keep the modelling framework simple, we temporarily drop the baseline covariates from our models, and present a review of methodologies to estimate parameters of a Gaussian process based model. Consider a simplified version of model (1.1) by dropping
Figure 1.2: First row: covariance functions; (1) $K_{SE}(t,s) = \exp\{-\frac{(t-s)^2}{2}\}$, (2) $K_S(t,s) = \exp\{-\sin^2(\pi|t-s|)/2\}$, and (3) $K_{BB}(t,s) = \min(t,s) + ts$. Second row: three random functions drawn from corresponding centered Gaussian processes.

the baseline (or time-varying) covariates,

$$Y_i(t) = X_i(t) + \epsilon_i(t), t \in [0,T], \quad i = 1, \ldots, I, \quad (1.9)$$

$$X_i(t) \overset{iid}{\sim} GP(\mu(\cdot), K(\cdot, \cdot)),$$

$$\epsilon_i(t) \overset{iid}{\sim} WN(\sigma^2),$$

where $WN(\sigma^2)$ denotes a white noise process with mean zero and variance $\sigma^2$. For example, we may assume $\epsilon_i(t)$ to be a Gaussian process with mean function identically zero and covariance function $K(t,s) = \sigma^2 \delta_{ts}$.

In practice, we never observe the process $Y_i(t)$ for all $t$, rather we only observe $Y_i = \ldots$
Then model (1.9) becomes

\[
Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J_i, \quad (1.10)
\]

\[
(X_i(t_{i1}), \ldots, X_i(t_{iJ_i}))^T \overset{i\!d}{\sim} N \left( (\mu(t_{i1}), \ldots, \mu(t_{iJ_i}))^T, (K(t_{ij}, t_{i'j'}))_{1 \leq j, j' \leq J_i} \right),
\]

\[
\epsilon_{ij} \overset{i\!d}{\sim} N(0, \sigma^2).
\]

With Model (1.10), we have modelled the fixed effects with \( \mu(\cdot) \), which is also the population mean function, and the random effects with \( K(\cdot, \cdot) \), which also describes the variation within and across the subjects. We only pose minimal restrictions on the mean function \( \mu(\cdot) \) and covariance function \( K(\cdot, \cdot) \), so that we can approximate any underlying Gaussian process that generates \( X_i(t) \). In addition, linear mixed effects models are special cases of model (1.10). Though the NLMM Model (1.4) may not belong to the class of nonlinear mixed effects models based on Gaussian processes (NLMM-GP), we can still approximate them as well as other stochastic processes with Model (1.10) up to the first two moment functions. Compared to Model (1.4) where we need to develop methodology for every specific model, NLMM-GP is easier to be handled by a general method for estimation and statistical inference. In general, NLMM-GP is flexible enough to approximate most of the Gaussian processes and accommodate other stochastic processes up to the first two moments. What’s more, its estimation and statistical inference is easy since it eventually reduces to finite-dimensional Gaussian distributions.

A commonly used method involves specifying the mean function \( \mu(\cdot) \) and the covariance function \( K(\cdot, \cdot) \) parametrically and then estimating the parameters with a likelihood based approach (Casella and Berger 2001, Chapter 7) or an estimating equation based approach (Godambe 1991) or a Bayesian approach (Inusa et al. 2012, Chapter 2). More specifically, suppose we assume \( \mu(t) = \sum_{k=1}^{p} \phi_k(t) \beta_k \) and \( K(t, s) = \sum_{k=1}^{q} \lambda_k \phi_k(t) \phi_k(s) \),
where \( \{ \phi_k(t); k = 1, 2, \ldots; t \in [0, T]\} \) forms an orthogonal basis with respect to the weight function \( w(t), t \in [0, T], \) say, \( \int_0^T \phi_k(t)\phi_l(t)dw(t) = \delta_{kl}. \) Then Model (1.10) can be represented as

\[
Y_{ij} = \sum_{h=1}^p \phi_h(t_{ij})\beta_h + \sum_{k=1}^q \sqrt{\lambda_k} \phi_k(t_{ij})Z_{ik} + \sigma \epsilon_{ij}, \tag{1.11}
\]

where \( Z_{ik}, \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, 1), \) \( \beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p, \) and \( \lambda = (\lambda_1, \ldots, \lambda_q)^T \in [0, \infty). \)

Note that Model (1.11) can be written in vector form,

\[
Y_i = F_i \beta + e_i, \tag{1.12}
\]

\[
e_i \overset{\text{ind}}{\sim} N(0, F_i D(\lambda) F_i^T + \sigma^2 I),
\]

where \( F_i = (\phi_k(t_{ij})) \) is a \( J_i \times p \) matrix with \((j,k)\)th entry \( \phi_k(t_{ij}). \) We can use a likelihood based method (e.g. maximum likelihood method or Bayesian method) to estimate the model parameters \( \theta = (\beta, \lambda, \sigma^2) \in \mathbb{R}^p \times [0, \infty)^q \times [0, \infty). \) Although the above framework provides a very flexible class of models to approximate a Gaussian process, there are at least couple of limitations. First, we need to suitably choose the class of orthogonal basis \( \{ \phi_k(t); k = 1, 2, \ldots; t \in [0, T]\}. \) Second, we need to select the truncation orders \( p \) and \( q. \) One of the major goals of the dissertation is to overcome the above two limitations by simultaneously estimating the \( \mu(\cdot) \) and \( K(\cdot, \cdot) \) under some mild regularity conditions.

### 1.4 Organization of the Dissertation

The dissertation is organized as follows. In Chapter 1, we introduce linear mixed effects model and nonlinear mixed effects models for longitudinal data analysis. After a
discussion on- definition and properties of the Gaussian processes, a class of nonlinear mixed effects models based on Gaussian process is introduced and compared to other NLMM’s. In Chapter 2, a class of linear mixed effect models using Bernstein polynomial sieve (LMM-BPS) is proposed to approximate the class of NLMM-GP. The convergence properties of the LMM-BPS are explored in terms of $L_p$ norms for $1 \leq p < \infty$. In Chapter 3, we propose Bayesian predictive divergence criterion (BPDC) for selecting the tuning parameter in LMM-BPS with a simulation study comparing it with other two popular Bayesian model selection criteria. Two data analysis are presented using the LMM-BPS. In Chapter 4, LMM-BPS for longitudinal data subject to data irregularities such as missing and censoring is discussed. Also, BPDC is generalized to the cases where missing and censored values exist. Various simulated data scenarios are used to illustrate the superior performance of the modified BPDC. Finally, discussions and future work are presented in Chapter 5.
Chapter 2

Gaussian Processes Approximation
with Bernstein Polynomials

2.1 Introduction

In Chapter 1, we have introduced a class of Gaussian processes which flexible to approximate most of the stochastic processes reasonably well. However, the major limitation is that we still need to select a suitable class of functions that can well approximate the mean and covariance functions simultaneously. If we want to restrict it to very few ones, we have the problem in capturing proper forms of \( \mu(\cdot) \) and \( K(\cdot,\cdot) \). In the existing literature, functional principal component analysis (FPCA) is applied to tackle the problem owing to the 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 to use 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) applied the functional principal component analysis on the sleep EEG data for 500 subjects. However, to implement FPCA, we need to 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 censoring. 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 hard to interpret. We cannot make conclusion about the population trend based on their model. One of the most common limitations of the FPCA is that these methods are primarily based on two stages of estimations: first the eigenfunctions are estimated and then these estimated functions are plugged-in to estimate the mean function or to predict the future observations. To overcome some of these limitations, first we establish a result based on Bernstein polynomial sieve to show the $L_p$–norm approximation of a Gaussian process. Then we illustrate that Bernstein polynomial sieve based approximation can be written within a linear mixed effects model framework. To start with, we provide a brief overview on the theory of Bernstein polynomial based approximations.

### 2.2 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. A recent centennial retrospective review of BP can be found in Farouki (2012).
2.2.1 Definition and Convergency Properties

The Bernstein basis polynomials of degree \( m - 1 \) is defined as

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

(2.1)

Figure 2.1 gives the plots of Bernstein basis polynomial sieve of degree \( m - 1 \) for \( m = 3 \) and \( m = 8 \) respectively. The Bernstein polynomials of degree \( m - 1 \) is defined as any linear combination of Bernstein basis polynomials,

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

(2.2)

where \( a_{k,m} \) can be any real number. Any BP is continuous and infinitely many differentiable on \([0,1]\). One of the nice properties of BP is that the derivative of a BP is still a
BP of a lower degree. In other words,

$$B'_m(t) = (m-1) \sum_{k=1}^{m-1} (a_{k+1,m} - a_{k,m}) b_{k,m-1}(t).$$  \hspace{1cm} (2.3)$$

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

$$B^{(l)}_m(t) = l! \left( \begin{array}{c} m-1 \\ l \end{array} \right) \sum_{k=1}^{m-l} \nabla^{(l)} a_{k,m} b_{k,m-l}(t),$$  \hspace{1cm} (2.4)$$

where $\nabla^{(l)} a_{k,m} = \nabla^{(l-1)} a_{k+1,m} - \nabla^{(l-1)} a_{k,m}$ for $l = 1, 2, \ldots$. 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.

One of the most celebrated results in real analysis is that the sequence of polynomials are dense for the space of continuous functions on a compact interval, which is known as the Weierstrass Approximation Theorem. Bernstein (1912) provided a constructive proof of this celebrated result which we state below.

**Theorem 2.2.1. Bernstein Weierstrass Approximation Theorem, (Lorentz, 1953, Chapter 1).** Let $f : [0,1] \to \mathbb{R}$ be a bounded function and consider the Bernstein polynomial given by

$$B_m(f,t) = \sum_{k=1}^{m} f \left( \frac{k-1}{m-1} \right) b_{k,m}(t), \hspace{0.5cm} t \in [0,1], \hspace{0.5cm} m = 2, 3, \ldots.$$  \hspace{1cm} (2.5)$$
Then,

$$\lim_{m \to \infty} B_m(f, t) = f(t) \quad (2.6)$$

holds at each point of continuity of $f(t)$; and the convergence holds uniformly on any $[a, b] \subseteq [0, 1]$ if $f(t)$ is continuous on $[a, b]$.

$B_m(f, t)$ defined in Theorem 2.2.1 is also known as the Bernstein polynomial of function $f$. With Equation (2.5), we see that the Bernstein polynomial of a function reduces the estimation of the entire function to a finite dimensional problem. It also enables the shape-preserving property of Bernstein polynomials, which we introduce later in Section 2.2.2. We also state some well-known results that provide the order of approximation.

**Theorem 2.2.2. The Degree of Approximation by Bernstein Polynomials, (Lorentz, 1953, Chapter 1).** Let $f : [0, 1] \to \mathbb{R}$ has a continuous derivative $f'(t)$ in $[0, 1]$. If $f'(t)$ satisfies the Lipschitz condition of order 1 on $[0, 1]$ such that

$$|f'(t_1) - f'(t_2)| \leq A|t_1 - t_2|, \quad t_1, t_2 \in [0, 1], \quad (2.7)$$

where $A$ is a constant. Then,

$$|B_m(f, t) - f(t)| \leq C(m - 1)^{-1}, \quad (2.8)$$

where $C$ is a constant.

Bernstein polynomials can accommodate functions with support set other than $[0, 1]$. 
If a function defined on \([a, b]\) is continuous, we can define

\[
B_m(f, x) = \sum_{k=1}^{m} f \left( a + \frac{(b - a)(k - 1)}{m - 1} \right) b_{k,m} \left( \frac{x - a}{b - a} \right).
\]  

(2.9)

Equation (2.9) is obtained through the linear transformation \( t = (x - a)/(b - a) \) such that \( t \in [0, 1] \). Babu et al. (2002) suggested to use transformation \( t = x/(1 + x) \) and \( t = 1/2 + (1/\pi) \tan^{-1}(x) \) to map \([0, \infty)\) and \((-\infty, \infty)\) to \([0, 1]\) respectively. In practice we may use \( t = (x - a)/(b - a) \) where \( a \) and \( b \) are “estimated” from the observed values of \( x \)’s. After the transformation, the support set is always condensed or enlarged to \([0, 1]\), which is of length 1.

Theorem 2.2.1 provides the convergence properties of Bernstein polynomials in \( L_\infty \) norm. Hoeffding (1971) discussed the \( L_1 \) norm approximation error for Bernstein polynomials. Jones (1976) proved several theorems for the approximation error of Bernstein polynomials in \( L_2 \) norm. Khan (1985) proved some results in terms of \( L_p \) norm for \( p \geq 1 \). Some important results are listed below.

**Theorem 2.2.3. The \( L_1 \) norm of the Bernstein polynomials approximation, Hoeffding (1971).** Let \( f \) be of bounded variation in \([0,1]\). Then

\[
\int_0^1 |B_m(f, t) - f(t)| dt \leq C_m J(f) + m^{-1} Var_{[0,1]}(f),
\]  

(2.10)

where \( J(f) = \int_0^1 t^{1/2}(1-t)^{1/2}|f(t)| dt \) and \( C_m = 2^{1/2}(m-1/2)^{m-1/2}m^{-m} < \sqrt{2/[(m-1)e]} \).

**Conditions 2.2.1. Condition T, Jones (1976).** We say a function \( f(t) \) satisfies condition T at a point \( p \in [0, 1] \) if (a) for any \( \epsilon > 0 \) there exist \( \tau > 0 \) and \( 0 < q < 1 \)

\[1Var_{[0,1]}(f)\] is the total variation of \( f \) on \([0,1]\), defined as \( \sup_P \sum_{i=1}^{m-1} |f(t_{i+1}) - f(t_i)| \), where \( P = \{t_0, \ldots, t_m\} \) is any partition of \([0,1]\).

20
such that for any \( t_1, t_2 \in (0, 1) \) satisfying \( 0 < q(p - t_2) < p - t_1 < p - t_2 \leq \tau \) or \( 0 < q(t_2 - p) < t_1 - p < t_2 - p \leq \tau \), \( |f(t_1)/f(t_2) - 1| \leq \epsilon \) holds, and (b) there exist \( \gamma > 0, M > 0, a \geq 0 \) such that for \( 0 < |p - t| < \gamma, t \in (0, 1) \), \( f(t) \geq M|u - p|^a \) holds.

**Theorem 2.2.4.** The \( L_2 \) norm of the Bernstein polynomials approximation, Jones (1976). Let \( f \) be an absolutely continuous function on \((0, 1)\) with derivative \( f' \). Suppose \( f \) and \( f' \) satisfy the conditions

(i) \( f \) is bounded on \([0, 1]\) and there exist \( \delta > 0 \) and \( t_0 \in (0, 1) \) such that \( |f(t) - f(0)| \leq Kt^\delta, |f(t) - f(1)| \leq K(1 - t)^\delta \) for \( 0 \leq t \leq t_0 \), and some \( K > 0 \),

(ii) \( f' \) is continuous on \((0, 1)\), satisfies Condition 2.2.1, part (a) at 0 and 1,

(iii) \( J_2(f) = \int_0^1 [f'(t)]^2 t(1 - t) dt < \infty \).

Then

\[
\int_0^1 [B_m(f, t) - f(t)]^2 dt \leq m^{-1} J_2(f) + o(m^{-1}).
\]

**Theorem 2.2.5.** The \( L_p \) norm of the Bernstein polynomials approximation, Khan (1985). Let \( f(t) \) be a continuous function on \([0, 1]\) with continuous derivative \( f'(t) \). Define \( D_m^p(f, t) = \sum_{k=1}^m |f(k/m - 1) - f(t)|^p b_{m,k}(t) \), and \( V_p(f) = \int_0^1 t^{p/2}(1 - t)^{p/2} |f'(t)|^p dt \).

Then

\[
m^{p/2} D_m^p(f, t) \rightarrow C_p t^{p/2}(1 - t)^{p/2} |f'(t)|^p \text{ as } m \rightarrow \infty,
\]

and

\[
\lim_{m \rightarrow \infty} m^{p/2} \int_0^1 D_m^p(f, t) dt = C_p V_p(t),
\]

for \( 1 \leq p < \infty \), where \( C_p = 2^{p/2} \Gamma((p + 1)/2)/\sqrt{\pi} \).
With Theorems 2.2.1, 2.2.3, 2.2.4, and 2.2.5 we are able to explore convergence properties of Bernstein polynomials approximation for Gaussian processes.

2.2.2 Shape-preserving Properties

Another attractive property of Bernstein polynomials is its optimal shape restriction property among all polynomials (Carnicer and Pena 1993). Chapter 1, Gal (2008) provided complete proofs to many shape-preserving properties and some other interesting properties. Wang (2012) also elaborated some of the interesting results on convex functions, quasi-convex functions, monotone functions, and $u$-monotone function where $u(t) = t^\lambda$, $\lambda \in (0, 1)$. They showed that to estimate a shape-restricted function with certain pattern using Bernstein polynomials we only need to put some restrictions on the coefficients of the Bernstein polynomials. For example, if $f(t)$ is an unknown continuous increasing function on $[0, 1]$, then $B_m(f, t)$ is also a continuous increasing function. The coefficients of $B_m(f, t)$ are $\{f(0), f\left(\frac{1}{m-1}\right), \ldots, f(1)\}$, which is also an increasing sequence. Therefore, we just need to restrict the Bernstein polynomial coefficients to satisfy $a_{1,m} < a_{2,m} < \ldots < a_{m,m}$ if we try to find a Bernstein polynomial that approximates $f(t)$. If $f(t)$ is an unknown nonnegative function, then $B_m(f, t)$ is also a nonnegative function with nonnegative coefficients. Hence, we require $a_{k,m} \geq 0$ for $k = 1, \ldots, m$. If $f(t)$ is a convex function such that $2f\left(\frac{t_1+t_2}{2}\right) \leq f(t_1) + f(t_2)$, then $B_m(f, t)$ is also a convex function. In this case, we enforce $a_{k,m} - 2a_{k+1,m} + a_{k+2,m} \geq 0$ for $k = 1, \ldots, m - 2$. There is a huge literature of the use of BP for shape restricted regression methods. Examples include Chak et al. (2005), Chang et al. (2005), Chang et al. (2007), Curtis and Ghosh (2011), and Wang and Ghosh (2012), all of which used the shape-preserving properties and applied the Bernstein polynomials to density estimation, survival analysis, model
selection, and other areas.

2.3 A Class of Linear Mixed Effects Models with Bernstein Polynomial Sieve

2.3.1 Approximation of Gaussian Processes with LMM-BPS

As we discussed in Section 2.1, we have problem in identifying the mean and covariance functions in NLMM-GP. It is well known that the FPCA has its drawbacks in estimating the eigenfunctions based on sample covariance matrix. To get around of this annoying problem, we can use an approximation model with Bernstein polynomials instead. In Section 2.2, advantages of using Bernstein polynomials in approximating a continuous function have been stated. First of all, the consistency of Bernstein polynomials assures that we can use Bernstein polynomial sieve to approximate the mean and kernel functions. Next with some tricks, we can approximate both of them in one step. Secondly, the coefficients of Bernstein polynomials have very nice interpretation, which makes the inference of the model straightforward and reveal the underlying trend. Thirdly, with Bernstein polynomials we can handle the shape-restricted model by simply putting constraints on the coefficients. This is quite helpful when we have some knowledge on the shape of trend curves (e.g. growth curves are nondecreasing, etc.), and makes the model more reasonable. So we propose a class of linear mixed effects models with Bernstein
polynomials (LMM-BPS) to replace Model (1.10),

\[
Y_{ij} = \sum_{k=1}^{m} b_{km}(t_{ij})\beta_{ik} + \epsilon_{ij}, i = 1, \ldots, I; j = 1, \ldots, J_i, \tag{2.11}
\]

\[
\beta_i \iid \sim N(\beta_0, D),
\]

\[
\epsilon_i \iid \sim N(0, \sigma_i^2 \mathbf{I}),
\]

where \( \beta_i = (\beta_{i1}, \ldots, \beta_{im})^T \), \( \beta_0 = (\beta_{01}, \ldots, \beta_{0m})^T \) and \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T \). With Model (2.11), we have approximated \( X(t) \) with Gaussian process

\[
\text{GP}\left( \sum_{k=1}^{m} b_{km}(t)\beta_{0k}, \sum_{k=1}^{m} \sum_{h=1}^{m} b_{km}(t)b_{hm}(s)D_{kh} \right).
\]

According to the discussion about the derivative of a Gaussian process in Section 1.2, Model (2.11) indicates the derivative of the response process \( X'(t) \) can be approximated with Gaussian process

\[
\text{GP}\left( (m-1) \sum_{k=1}^{m-1} b_{k,m-1}(t)(\beta_{0,k+1} - \beta_{0,k}), \tag{2.12}
\right.
\]

\[
(m-1)^2 \sum_{k=1}^{m-1} \sum_{h=1}^{m-1} b_{k,m-1}(t)b_{h,m-1}(s)(D_{k+1,h+1} - 2D_{k+1,h} + D_{k,h}) \right).
\]

The above formula can be easily shown using Equation (2.4).

Denote \( Y_i = (Y_{i1}, \ldots, Y_{iJ_i})^T \) as the \( J_i \times 1 \) response vector, \( t_i = (t_{i1}, \ldots, t_{iJ_i})^T \) as the \( J_i \times 1 \) vector of time points, and \( B_i = (b_{1m}(t_i) \ldots b_{mm}(t_i)) \) as the \( J_i \times m \) matrix of basis
functions. Now the Model (2.11) can be written as

\[
Y_i = B_i \beta_i + \epsilon_i, \quad i = 1, \ldots, I, \quad \tag{2.13}
\]

\[
\beta_i \overset{iid}{\sim} N(\beta_0, D),
\]

\[
\epsilon_i \overset{inid}{\sim} N(0, \sigma_i^2 I).
\]

Then, at \( t_i \) the mean of the Gaussian process is approximated by \( B_i \beta \) and covariance is approximated by \( B_i DB_i^T \).

Let \( J \) be the number of distinct \( t_{ij} \)'s. To avoid collinearity issue with large degree polynomials, \( m \) should always be less than \( J \) in Model (2.11). Tenbusch (1997) suggested a more strict upper bound for the choice of \( m \) in nonparametric approximation with Bernstein polynomials such that \( m \leq \lfloor J^{3/4} \rfloor \). For the case \( m = 1 \), we can only approximate a degenerated Gaussian process. Hence, we also require \( m \) to be greater than 1. So the value of \( m \) is chosen from \( \{2, \ldots, \lfloor J^{3/4} \rfloor \} \).

### 2.3.2 Priors and Starting Values

To estimate Model (2.13), we can compute MLE (maximum likelihood estimate) or RMLE (restricted maximum likelihood estimate) by solving some estimating equations. Also, setting \( \beta_i = \beta_0 + b_i \), we can treat \( (b_1, b_2, \ldots, b_I) \) as missing, and apply EM algorithm to obtain estimates of \( (\beta_0, D, \sigma_1^2, \ldots, \sigma_I^2) \). Another way is to specify proper prior distributions for the parameters to form a Bayesian model, and estimate the model using MCMC methods.

As many longitudinal data involve missing observations and Bayesian methodologies are particularly well suited to properly handle imputation procedure, we choose to adopt...
a Bayesian modelling framework with the following set of prior distributions,

\[
\beta_0 \sim N(\tilde{\beta}_0, \sigma_0^2 I), \quad \text{(2.14)}
\]
\[
D \sim IWish^2(R, v),
\]
\[
\sigma^2_i \overset{iid}{\sim} IG(a, b),
\]

where \(\tilde{\beta}_0, \sigma_0^2, R, v, a, \) and \(b\) are pre-specified constants.

Stack \(Y_i\)'s and \(B_i\)'s such that \(Y = (Y_1^T, \ldots, Y_I^T)^T\) and \(B = (B_1^T, \ldots, B_I^T)^T\). Initial values of the parameters are obtained from naive least square estimates.

\[
\beta_0^{(0)} = (B^T B)^{-1} B^T Y, \quad \text{(2.15)}
\]
\[
b_i^{(0)} = (B_i^T B_i)^{-1} B_i^T (Y_i - B_i \beta^{(0)}),
\]
\[
\sigma^2_i^{(0)} = \frac{1}{J_i - 1} (Y_i - B_i \beta^{(0)} - B_i b_i^{(0)})^T (Y_i - B_i \beta^{(0)} - B_i b_i^{(0)}),
\]
\[
D^{(0)} = \frac{1}{I - 1} \sum_{i=1}^I b_i^{(0)} b_i^{(0)T}.
\]

With (2.14) and (2.15), we can estimate Model (2.13) by drawing posterior samples using Gibbs sampling via software like WinBUGS/OpenBUGS.

### 2.4 Convergence Properties

In this section, we present a finite dimensional approximation of the \(GP(\mu(\cdot), K(\cdot, \cdot))\) by random Bernstein polynomial sieve, \(\sum_{k=1}^m b_{km}(t) \beta_{km}\), where \(\beta_m = (\beta_{1m}, \ldots, \beta_{mm})^T \sim \]

\(^2\)IWish stands for inverse Wishart distribution. If a \(m \times m\) positive definite random matrix \(D\) follows \(IWish(R, v)\), its density function is \(f(D|R, v) = \frac{|R|^{v/2} |D|^{-\frac{v+m+1}{2}} \exp\{-\frac{1}{2} tr(RD^{-1})\}}{2^{mv} \Gamma_m(v/2) \Gamma_m(v/2)}\), where \(\Gamma_m(\cdot)\) is the multivariate gamma function.

\(^3\)B\(^{-}\) denotes the MoorePenrose pseudoinverse of \(B\).
$N(\mu_m, D_m)$ as $m$ goes to infinity. More specifically, given $X(\cdot) \sim GP(\mu(\cdot), K(\cdot, \cdot))$ we construct a sequence of stochastic processes $X_m(\cdot) = \sum_{k=1}^{m} b_{km}(\cdot)\beta_{km}$ and show that $E\|X(\cdot) - X_m(\cdot)\|_p \to 0$ as $m \to \infty$ for $p \geq 1$.

**Theorem 2.4.1.** 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 $\lambda_i$'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

$$\sum_{i=1}^{\infty} \lambda_i \int_{0}^{1} t(1-t)|e_i'(t)|^2 dt < \infty$$

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 $\mu_m$ and $D_m$, such that

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

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 $\{X(t) : t \in [0, 1]\}$.

**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 $X(t) = \mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i$ in mean square sense, where $Z_i \iid N(0, 1)$.

$X_m(t)$ 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

\[
E\|X - X_m\|_2^2 = E\left[ \int_0^1 (X(t) - X_m(t))^2 dt \right],
\]

\[
= \int_0^1 E\left[ \left\{ \mu(t) - B_m^T(t) \gamma_m(\mu) \right\}^2 \right],
\]

\[
+ \sum_{i=1}^{m} \lambda_i \| e_i(t) - B_m^T(t) \gamma_m(e_i) \|_2^2 dt \left( E(Z_i^2) + \int_0^1 \sum_{i=m+1}^{\infty} \lambda_i e_i^2(t) dt \right),
\]

\[
= \| \mu - B_m^T \gamma_m(\mu) \|_2^2 + \sum_{i=1}^{m} \lambda_i \| e_i - B_m^T \gamma_m(e_i) \|_2^2 + \sum_{i=m+1}^{\infty} \lambda_i \| e_i \|_2^2.
\]

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, using Cauchy-Schwarz inequality, we have \( \sum_{i=1}^{m} \lambda_i \| e_i - B_m^T \gamma_m(e_i) \|_2^2 \leq \sum_{i=1}^{\infty} \lambda_i \| e_i - B_m^T \gamma_m(e_i) \|_2^2 = m^{-1} \sum_{i=1}^{\infty} \lambda_i m \| e_i - B_m^T \gamma_m(e_i) \|_2^2 \). Define \( Q_m^2(f) = \int_0^1 \sum_{k=1}^{m} |f(k^{-1}) - \)
\[ f(t)^2 b_{km}(t) dt. \] Since \( \sum_{k=1}^{m} b_{km}(t) = 1, \]

\[ |e_i - B_{m}^{T} \gamma_m(e_i)|^2 = \left| \sum_{k=1}^{m} e_i \left( \frac{k-1}{m-1} \right) b_{km}(t) - e_i(t)^2 \right|^2, \]

\[ = \left| \sum_{k=1}^{m} \{e_i \left( \frac{k-1}{m-1} \right) - e_i(t)\} b_{km}(t) \right|^2, \]

\[ \leq \sum_{k=1}^{m} \{|e_i \left( \frac{k-1}{m-1} \right) - e_i(t)\} \sqrt{b_{km}(t)} \right|^2 \sum_{k=1}^{m} \sqrt{b_{km}(t)} , \]

\[ = \sum_{k=1}^{m} \left| e_i \left( \frac{k-1}{m-1} \right) - e_i(t) \right|^2 b_{km}(t). \]

Therefore,

\[ \|e_i - B_{m}^{T} \gamma_m(e_i)\|_2^2 = \int_{0}^{1} \left| \sum_{k=1}^{m} e_i \left( \frac{k-1}{m-1} \right) b_{km}(t) - e_i(t)^2 \right|^2 dt, \]

\[ \leq \int_{0}^{1} \sum_{k=1}^{m} \left| e_i \left( \frac{k-1}{m-1} \right) - e_i(t) \right|^2 b_{km}(t) dt = Q_m^2(e_i). \]

Using Theorem 1 in Khan (1985), we have

\[ \lim_{m \to \infty} m \|e_i - B_{m}^{T} \gamma_m(e_i)\|_2^2 \leq \lim_{m \to \infty} m Q_m^2(e_i) = C_2 V_2(e_i), \]

where \( V_2(e_i) = \int_{0}^{1} t(1-t)|e_i'(t)|^2 dt. \) Then, with assumption that \( \sum_{i=1}^{\infty} \lambda_i V_2(e_i) < \infty \) and Fatou’s lemma,

\[ \lim_{m \to \infty} \sum_{i=1}^{\infty} \lambda_i m \|e_i - B_{m}^{T} \gamma_m(e_i)\|_2^2 \leq \sum_{i=1}^{\infty} \lambda_i \lim_{m \to \infty} m Q_m^2(e_i), \]

\[ = C_2 \sum_{i=1}^{\infty} \lambda_i V_2(e_i) < \infty. \]

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,

$$\lim_{m \to \infty} \sum_{i=m+1}^{\infty} \lambda_i \| e_i \|^2 = \lim_{m \to \infty} \sum_{i=m+1}^{\infty} \lambda_i = 0.$$ 

Altogether, we have proved $\lim_{m \to \infty} E \| X - X_m \|^2 = 0$.

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 $Cov(Z_i, Z_j) = \delta_{ij}$.

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 2.4.2.** 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 $\lambda_i$’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

1. $\sum_{i=1}^{\infty} \lambda_i \{ V_p(e_i) \}^{1/p} < \infty$, where $V_p(e_i) = \int_0^1 |e_i'(t)|^p dt$.
2. $\sum_{i=1}^{\infty} \lambda_i \| e_i \|_p < \infty$, where $\| e_i \|_p = \left\{ \int_0^1 |e_i(t)|^p dt \right\}^{1/p}$.

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

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

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 $\{X(t) : t \in [0, 1]\}$. 

30
Proof. Using Mercer’s theorem, we have \( X(t) \overset{d}{=} \mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i \), 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,

\[
E\|X - X_m\|_p = E\{\|\mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} e_i(t) Z_i - B_m^T(t) \gamma_m(\mu) - \sum_{i=1}^{m} \sqrt{\lambda_i} B_m^T(t) \gamma_m(e_i) Z_i\|_p\},
\]

\[
\leq \|\mu - B_m^T \gamma_m(\mu)\|_p + \sum_{i=1}^{m} \sqrt{\lambda_i} \|e_i - B_m^T \gamma_m(e_i)\|_p E|Z_i| + \sum_{i=m+1}^{\infty} \sqrt{\lambda_i} \|e_i\|_p E|Z_i|.
\]

Since \( Z_i \overset{\text{iid}}{\sim} N(0,1) \), \( E|Z_i| = \sqrt{\frac{\pi}{2}} \). By Bernstein Weierstrass approximation theorem, we have \( \lim_{m \to \infty} \|\mu - B_m^T \gamma_m(\mu)\|_p = 0 \). Define \( Q_m^p(f) = \int_0^1 \sum_{k=1}^{m} |f(k-\frac{m-1}{m-1}) - f(t)|^p b_{km}(t) dt \).

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, \( \left| \sum_{k=1}^{m} (e_i(k-\frac{m-1}{m-1}) - e_i(t)) b_{km}(t) \right|^p \leq \sum_{k=1}^{m} |f(k-\frac{m-1}{m-1}) - e_i(t)|^p b_{km}(t) dt \), since \( \sum_{k=1}^{m} b_{km}(t) = 1 \). Hence,

\[
\|e_i - B_m^T \gamma_m(e_i)\|_p = \left\{ \int_0^1 \left| \sum_{k=1}^{m} e_i(k-\frac{m-1}{m-1}) b_{km}(t) - e_i(t) \right|^p dt \right\}^{1/p},
\]

\[
\leq \left\{ \int_0^1 \sum_{k=1}^{m} |e_i(k-\frac{m-1}{m-1}) - e_i(t)|^p b_{km}(t) dt \right\}^{1/p},
\]

\[
= \{Q_m^p(e_i)\}^{1/p}.
\]
Then,
\[
\lim_{m \to \infty} \sum_{i=1}^{m} \sqrt{\lambda_i} \|e_i - B_m^T \gamma_m(e_i)\|_p \leq \lim_{m \to \infty} m^{-1/2} \sum_{i=1}^{m} \sqrt{\lambda_i \sqrt{m}} \{Q_m^p(e_i)\}^{1/p},
\]
\[
\leq \lim_{m \to \infty} m^{-1/2} \sum_{i=1}^{\infty} \sqrt{\lambda_i \sqrt{m}} \{Q_m^p(e_i)\}^{1/p}.
\]
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 \(\lim_{m \to \infty} E \|X - X_m\|_2^2 = 0\).

Remark: (1) If \(\lambda_i\)’s have finitely many non-zero values, conditions (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 is met.

Consider the popular squared exponential covariance function defined on \([0, 1] \times [0, 1]\) such that
\[
K(t, s) = \exp\left\{ - \frac{(\Phi^{-1}(t) - \Phi^{-1}(s))^2}{2} \right\}, \quad t, s \in [0, 1],
\] (2.17)
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
\[
\lambda_i = \left(\frac{3 - \sqrt{5}}{2}\right)^{i+\frac{1}{2}} \quad \text{and} \quad e_i(t) = \phi_i(\Phi^{-1}(t)),
\]
for \(i = 0, 1, \ldots\), where \(\phi_i(t) = \gamma_i \exp\left(-\frac{\sqrt{5}-1}{4}t^2\right)H_i\left(\sqrt{\frac{5}{2}}t\right)\), \(\gamma_i = \sqrt{5^{i/4}/(2i!)}\), and \(H_i(x) = (-1)^i \exp(t^2) \frac{d^i}{dt^i} \exp(-t^2)\) (Hermite polynomial). For \(p = 1\), Figure 2.2 shows condition (i) in Theorem 2.4.2 is satisfied, and condition (ii) is also satisfied since \(\frac{3 - \sqrt{5}}{2} < 1\) and \(\|e_i\|_1 \leq 1\). Thus, with Theorem 2.4.2, we can show that there always exists a Bernstein polynomial sieve that converges to a Gaussian process with continuous mean function and square exponential covariance function in \(L_1\) norm.

Theorem 2.4.1 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 2.4.2 demonstrates an even stronger convergence of our Bernstein polynomial sieve to Gaussian processes satisfying certain regularity conditions in terms of \(L_p\) norm. Moreover, in the proof of Theorem 2.4.1 and 2.4.2, we have explicitly constructed the Bernstein polynomials estimator where \(\mu_m\) 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 covariance 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 covariance functions, $\mu \circ g^{-1}$ and $K \circ g^{-1}$, we can apply Theorems 2.4.1 and 2.4.2. In this way, Theorems 2.4.1 and 2.4.2 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 Theorems 2.4.1 and 2.4.2. Moreover, results of Khan (1985) can be used to generalize our results to many other operators (e.g. Feller operator). Every type of polynomial has its own advantage. We still favor the use of a Bernstein polynomials estimator to approximate a Gaussian process not only because we have the theorems originally proved for it but also because it can preserve the original shape information of the Gaussian process.

2.5 Discussion

We have proposed a class of linear mixed effects models with Bernstein polynomial sieve (LMM-BPS) to approximate NLMM-GP in this chapter. Background knowledge about Bernstein polynomials is presented to help understand the advantages and properties of LMM-BPS. Choices of the tuning parameter $m$ are suggested as any integer between 2 and $\lfloor J^{3/4} \rfloor$, where $J$ is the number of unique time points in the data set. More discussion on choosing the best $m$ is made in next chapter. Priors and initial values are suggested to form and estimate our LMM-BPS in a Bayesian framework. Some convergence properties are proved to demonstrate that the proposed LMM-BPS can approximate the mean and
Figure 2.2: $\sum_{i=0}^{m} \sqrt{\lambda_i} V_1(e_i)$ is plotted against $m$ for the square exponential kernel.

covariance functions of Gaussian processes simultaneously in terms of $L_p$ norm, which
also imply consistency in distribution and in probability. Moreover, the optimal shape
restriction properties of Bernstein polynomials are preserved in our model, which allows
us to incorporate predetermined knowledge into our models.
Chapter 3

Bayesian Model Selection using Predictive Divergence

3.1 Introduction

Statistical model selection is an important part of any data analysis. We may not ever know the true model, but we can always choose the best model among a given set of models. In the preceding chapter, we have introduced a class of linear mixed effects models with Bernstein polynomial sieve. But the problem of selecting the tuning parameter \( m \) remains. In this chapter, we focus on the tuning parameter selection problem and real data analysis. In Section 3.2, we give a brief review of several popular model selection methods, especially the deviance information criterion (DIC) and log pseudo marginal likelihood (LPML). Besides those frequently used methods, we propose the Bayesian predictive divergence criterion (BPDC) by extending the definition of predictive divergence criterion (PDC) in Davies et al. (2005) to Bayesian models. In Section 3.3, a simulation study that compares the three Bayesian model selection criteria, DIC, LPML, and
BPDC, for linear mixed models is described. The results from simulation studies show the performance of BPDC is superior to the other two in selecting the tuning parameter $m$. It should work better at least for other simpler models, such as generalized linear models. An supplementary simulation study is presented showing how well LMM-BPS can approximate NLMM-GP. In Section 3.4, we apply our method to a Sitka spruce tree data, where we try to model the growth curves. Other models used before for the same data set are also discussed and compared to ours. Finally, the results are illustrated in plots and conclusions are made based on the estimated model.

### 3.2 Bayesian Model Selection Criteria

Consider a countable class of model $\{M_1, M_2, \ldots\}$ and the problem of model selection refers to the selection of one as the best approximation to the true data generating mechanism. To decide which model is preferred, we need a concept of the “distance” between the candidate model and the true model. A smaller “distance” implies a better model. As a statistical model is rather an abstract object, there are several ways to measure the closeness between models. One popular measure of discrepancy which is related to the likelihood function is known as the Kullback-Leibler discrepancy or divergence measure (KLD). Suppose the true density function of the data $Y$ is $g(Y)$ and $f(Y)$ is the density function postulated by a candidate model, then the KLD between the true density and the candidate density is defined as

$$KLD(f|g) = \int g(Y) \log \frac{g(Y)}{f(Y)} dY,$$

(3.1)
provided that the integral exists. Notice that \( KLD(f|g) \geq 0 \) and \( KLD(f|g) = 0 \) if and only if \( f = g \) on the set \( \{ Y : \min(f(Y),g(Y)) > 0 \} \). However, KLD is not a proper distance because it is not symmetric in the sense that \( KLD(f|g) \neq KLD(g|f) \). But this is generally not a problem as one may simply use a symmetrized version \( KLD^*(f|g) = (KLD(f|g) + KLD(g|f))/2 \). Notice that \( KLD(f|g) = - \int g(Y) \log f(Y) dY + \int g(Y) \log g(Y)dY = -E_g(\log f(Y)) + c(g) \), and hence \( E_g(\log f(Y)) \) discriminates the value of \( KLD(f|g) \) for a fixed \( g \). In fact \( KLD(f_1|g) - KLD(f_2|g) = E_g[\log \frac{f_2(Y)}{f_1(Y)}] \) and so \(-2\log f(Y) \) can be used equivalently as the Kullback-Leibler discrepancy, and hence the adequacy of a model can be measured by the expected KL discrepancy, say \(-2E_g(\log f(Y)) \). It is easy to see that the density \( f \) that minimizes \( KLD(f|g) \) would also minimize \(-2E_g(\log f(Y)) \). One difficulty of using KLD is that we do not know \( g \) and we cannot compute \( KLD(f|g) \).

Most of the model selection criteria target at obtaining an estimate related to this expected KL discrepancy. The Akaike information criterion (AIC) proposed by Akaike (1974) is widely used as a model selection criterion. Furthermore, it was proved to be an asymptotic estimator of the KL discrepancy under some regularity conditions. Along these ideas, some other popular criteria were developed afterwards, such as the Bayesian information criterion (BIC; Schwarz 1978), Hannan-Quinn information criterion (HQIC; Hannan and Quinn 1979), and the deviance information criterion (DIC; Spiegelhalter et al. 2002). The DIC was shown to closely resemble the AIC for a selected class of models (Spiegelhalter et al. 2002).

In contrast to using information criteria like AIC, BIC, etc, another popular class of model selection criteria is based on cross-validation methods when the main target of analysis is prediction. Three types of cross-validation are generally adopted (Stone 1974; Geisser 1975): (i)K-fold cross-validation, (ii)repeated random sub-sampling cross-
validation, (iii) leave-one-out cross-validation, where the leave-one-out cross-validation is shown to be asymptotically related to AIC (Stone 1974). Criteria such as log pseudo marginal likelihood (LPML; Geisser and Eddy 1979) and predictive divergence criterion (PDC; Davies et al. 2005) based on leave-one-out cross-validation were proposed and shown to approximate AIC when some regularity conditions are satisfied.

In Bayesian analysis, Bayes factor (BF; Jeffreys 1935; Berger and Delampady 1987) is another tool for model selection which is defined as the ratio of marginal densities of two competing models evaluated at the observed data. The BF does not require alternative models to be nested and favors simpler model in the event of similar fits, but it is sensitive to the assumptions in the parametric model and the choice of prior distribution (Kass and Raftery 1995). Computation of BF can be challenging for some models. Though a series of papers have made progress in computing the BF (Newton and Raftery 1994; DiCiccio et al. 1997; Weinberg 2012), it is still numerically unstable when diffuse prior or reference priors are used and is undefined when improper priors are used. DIC is frequently used for model selection purpose 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. The leave-one-out cross-validation based criterion LPML is also developed in the frame work of Bayesian analysis, and can be approximated well by using MCMC samples along with importance sampling. Thus, from a computational advantages point of view we choose to use DIC and LPML for model selection. Other than that, a new criterion named Bayesian predictive divergence criterion (BPDC), which is extended from PDC, is proposed as an alternative. We discuss more details on DIC, LPML, and BPDC in the following sub-sections.
### 3.2.1 Deviance Information Criterion

For a given candidate model \( \{ f(\cdot, \theta) : \theta \in \Theta \} \) and a given data set denoted by \( Y \) (which could be a matrix of observed response and predictors), the DIC proposed by Spiegelhalter et al. (2002) is defined as

\[
DIC = D(Y, \bar{\theta}) + 2p_D, \tag{3.2}
\]

where \( D(Y, \theta) = -2 \log f(Y|\theta), \bar{\theta} = E(\theta|Y) \), and \( p_D = E(D(Y, \theta)|Y) - D(Y, \bar{\theta}) \). The term \( D(Y, \bar{\theta}) \) measures the fit of the model, and the term \( p_D \) is a measure of the complexity of the model, also called “the effective number of parameters”. For the purpose of computing, DIC can be also written as

\[
DIC = -4E(\log f(Y|\theta)|Y) + 2 \log f(Y|E(\theta|Y)). \tag{3.3}
\]

If we get samples from posterior distribution (e.g. by using MCMC), say \( \theta^{(l)} \sim p(\theta|Y) \) for \( l = 1, \ldots, L \), DIC can be estimated as

\[
\hat{DIC} = -4 \frac{1}{L} \sum_{l=1}^{L} \log f(Y|\theta^{(l)}) + 2 \log f \left( Y \mid \frac{1}{L} \sum_{l=1}^{L} \theta^{(l)} \right). \tag{3.4}
\]

In our case, candidate models are from a class of linear mixed models with general form

\[
Y_{i} = B_i \beta_i + \epsilon_i, i = 1, \ldots, I \tag{3.5}
\]

\[
\beta_i \sim N(\beta_0, D),
\]

\[
\epsilon_i \sim N(0, \sigma^2_I I).
\]
As discussed in Spiegelhalter et al. (2002), the deviance $D(Y, \theta)$ for DIC can be computed in different ways corresponding to the parameter of interest and likelihood functions. In Model (3.5), if individual level parameters ($\beta_i$’s) are our focus, we can set $\theta = (\beta_1, \ldots, \beta_I, \sigma^2_1, \ldots, \sigma^2_I)$. It follows that $Y_i|\theta \sim N(B_i\beta_i, \sigma^2_i I)$ and $f(Y|\theta) = \prod_{i=1}^{I} f(Y_i|\beta_i, \sigma^2_i)$.

The DIC computed with this Deviance is called conditional DIC in our paper, and denoted as $DIC_c$. This is also the one that is routinely provided by WinBUGS/OpenBUGS. If the population level parameter is of our interest, we should set $\theta = (\beta_0, D, \sigma^2_I)$. It leads to $Y_i|\theta \sim N(B_i\beta_0, B_iDB_i^T + \sigma^2_i I)$ and $f(Y|\theta) = \prod_{i=1}^{I} f(Y_i|\beta_0, D, \sigma^2_i)$. Here we have integrated out $\beta_i$’s, so we name the DIC computed in this way as marginal DIC and denote it $DIC_m$. We compute both $DIC_c$ and $DIC_m$ in the simulation study to see which one works better in selecting the correct model from a set of candidate models.

DIC has been widely used in real data analysis, and is suggested to work well. Spiegelhalter et al. (2002) provided some justification mainly considering generalized linear regression models, showing DIC is analogous to AIC. But its theoretical foundations have not been thoroughly explored in the literature (see, e.g. Celeux et al. 2006 and Plummer 2008). Plummer (2008) stated that two necessary assumptions are needed to use DIC. First, data can be broken into components that are independent conditional on parameters in focus. Second, the term $p_D$ must be much smaller than the number of independent observations, i.e. $p_D \ll I$. When either of the assumptions does not hold, DIC tends to under-penalize the complexity of a model. In Model (3.5), when individual level parameter is our concern, the second assumption here may not be satisfied since the size of parameters grows manyfold with the number of subjects in our case. Therefore, we should be careful in using DIC as our model selection criterion.
3.2.2 Log Pseudo Marginal Likelihood

Geisser and Eddy (1979) proposed the leave-one-out cross-validation predictive density, which is also known as conditional predictive ordinate (CPO; Geisser 1993; Gelfand 1995). The CPO for subject $i$ is defined as

$$CPO_i = \int f(Y_i|\theta)p(\theta|Y_{-i})d\theta = p(Y_i|Y_{-i}), \quad (3.6)$$

where $Y_{-i}$ denotes the vector of all observations except the $i$th one. A larger value of $CPO_i$ implies a better fit of the model at $Y_i$. $CPO_i$ can also be a tool for identifying outliers or influential points. If $CPO_i$ is very low compared to other subjects in the data set, the $i$th subject is likely to be an outlier or an influential point. $CPO_i$ is easy to compute with MCMC approximation. Assuming that $Y_i$’s are mutually independent given $\theta$, note that

$$p(\theta|Y) = \frac{f(Y_i|\theta,Y_{-i})f(Y_{-i}|\theta)p(\theta)}{p(Y_{-i})p(Y_i|Y_{-i})},$$

$$= \frac{p(Y_i|\theta,Y_{-i})}{p(Y_i|Y_{-i})}p(\theta|Y_{-i}).$$

So we have

$$\int \frac{p(\theta|Y)}{f(Y_i|\theta,Y_{-i})}d\theta = \frac{1}{p(Y_i|Y_{-i})},$$

$$p(Y_i|Y_{-i}) = \left\{ E_\theta \left[ \frac{1}{f(Y_i|\theta,Y_{-i})} | Y \right] \right\}^{-1},$$

$$= \left\{ E_\theta \left[ \frac{1}{f(Y_i|\theta)} | Y \right] \right\}^{-1}.$$
With posterior samples from MCMC simulation, say $\theta^{(l)} \sim p(\theta|Y), l = 1, \ldots, L$, we can estimate $CPO_i$ with

$$
\hat{CPO}_i = \left[ \frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(Y_i|\theta^{(l)})} \right]^{-1}.
$$  

(3.7)

Geisser and Eddy (1979) then introduced log pseudo marginal likelihood (LPML) as the sum of log $CPO_i$’s across all subjects. To match its scale with DIC and other information criteria, we use the LPML defined as

$$
LPML = -2 \sum_{i=1}^{I} \log(CPO_i),
$$

(3.8)

$$
= 2 \sum_{i=1}^{I} \log E_{\theta} \left[ \frac{1}{f(Y_i|\theta)} \right].
$$

Then we can estimate LPML with

$$
\hat{LPML} = -2 \sum_{i=1}^{I} \log(\hat{CPO}_i) = 2 \sum_{i=1}^{I} \log \left[ \frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(Y_i|\theta^{(l)})} \right].
$$

(3.9)

A smaller value of LPML indicates a better model. This estimate has been implemented in DPpackage in R.

Stone (1977) showed the asymptotic equivalence between LPML and AIC from frequentists’ view. According to the discussion in Gelfand and Dey (1994), LPML implicitly includes a similar penalty term as AIC. LPML can be used for comparing complex models, and does not require models to be nested. Even with improper prior, LPML is well defined. Another advantage of using LPML (Carlin and Louis 2008, Chapter 4; Christensen 2010, Chapter 4) is that it does not directly depend on estimated parameters since we have them integrated out. This is quite helpful when it is not clear what parameters
are of interest as in mixed effects models. However, Ando and Tsay (2010) pointed out LPML is hard to apply to data without mutual independence such as time series, and LPML can go wild when data contain outliers.

3.2.3 Bayesian Predictive Divergence Criterion

Davies et al. (2005) proposed predictive divergence criterion (PDC) for linear regression from a frequentists’ perspective. Analogous to the construction of AIC, given the true model $g(\cdot|\theta_0)$ and any candidate model $f(\cdot|\theta)$ they defined predictive discrepancy to be

$$d_i(\hat{\theta}_{-i}, \theta_0) = E_{Y|\theta_0}[−2 \log f(Y_i|\theta)|_{\theta=\hat{\theta}_{-i}},$$

(3.10)

$$= \left[ \int −2 \log f(Y_i|\theta)g(Y|\theta_0)dy \right]_{\theta=\hat{\theta}_{-i}},$$

(3.11)

and the overall predictive divergence to be

$$d_{PDC}(Y, \theta_0) = \sum_{i=1}^{I} d_i(\hat{\theta}_{-i}, \theta_0),$$

(3.12)

where $\hat{\theta}_{-i}$ denotes the maximum likelihood estimate of $\theta$ based on the leave-one-out data $Y_{-i}$. They then took expectation of $d_{PDC}$ with respect to $g(Y|\theta_0)$ and got

$$\Delta_{PDC}(\theta_0, f) = E_{Y|\theta_0}[d_{PDC}(Y, \theta_0)],$$

(3.13)

$$= \sum_{i=1}^{I} \int \left[ \int −2 \log f(Y_i|\theta)g(Y|\theta_0)dY \right]_{\theta=\hat{\theta}_{-i}} g(Y|\theta_0)dY.$$

Based on $\Delta_{PDC}(\theta_0, f)$, they constructed the PDC by finding an unbiased estimator of
\[
PDC = \sum_{i=1}^{I} -2 \log f(Y_i|\theta = \hat{\theta}_{-i}).
\] 

(3.14)

Davies et al. (2005) also showed the asymptotic equivalence between PDC and AIC in the framework of classical linear regression.

Motivated by Davies et al. (2005), we propose another cross-validation model selection criterion, the Bayesian predictive divergence criterion (BPDC). Following the same idea, we generalize the predictive divergence to Bayesian analysis. In a Bayesian model, parameters \(\theta_0\) and \(\theta\) are not fixed any more, but are considered as random variables with a prior distribution \(p(\theta)\). So when we define Bayesian predictive discrepancy for the \(i\)th subject, instead of plugging in the MLE of \(\theta\) based on leave-one-out data, we take expectation with respect to the posterior distribution of \(\theta\) depending on \(Y_{-i}\), say

\[
d^B_{i}(Y, f) = \int -2 \log f(Y_i|\theta)p(\theta|Y_{-i})d\theta,
\]

(3.15)

where \(p(\theta|Y_{-i}) = f(Y_{-i}|\theta)p(\theta)/\int f(Y_{-i}|\theta)p(\theta)d\theta\). The total Bayesian predictive discrepancy is

\[
d_{BPDC}(Y, f) = \sum_{i=1}^{I} d^B_{i}(Y, f).
\]

(3.16)

Taking expectation of \(d_{BPDC}(Y, f)\) with respect to the true model \(g(\cdot)\), we get

\[
\Delta_{BPDC}(f) = E_{Y}[d_{BPDC}(Y, f)]
\]

(3.17)

\[
= E_{Y}\left[\sum_{i=1}^{I} \int -2 \log f(Y_i|\theta)p(\theta|Y_{-i})d\theta\right]
\]

(3.18)
Our target is to find an unbiased estimator to $\Delta_{BPDC}(f)$. Note that in the preceding equation, the term inside expectation is a function of $y$ where $Y \sim g(\cdot)$. Thus, we define the Bayesian predictive divergence criterion as

$$BPDC = \sum_{i=1}^{I} \int -2 \log f(Y_i|\theta)p(\theta|Y_{-i})d\theta,$$

(3.19)

$$= \sum_{i=1}^{I} -2E_{\theta} \log f(Y_i|\theta)|Y_{-i}].$$

By comparing (3.19) with (3.8), we can find the definition of BPDC is very close to LPML. The difference is that BPDC takes logarithm of $f(Y_i|\theta)$ inside the integral. We can imagine that BPDC should be more stable especially when we take computational error into account since log density is more stable than the density function itself. At the same time, the asymptotic property of BPDC should be similar to LPML. Actually, both BPDC and LPML are special cases of expected utility estimates for assessing the predictive ability of the model that was discussed in Vehtari and Lampinen (2002).

With simple models, we can compute $\int \log f(Y_i|\theta)p(\theta|Y_{-i})d\theta$ directly. But for most of complex Bayesian models, it is hard to integrate out $\theta$. In this situation, without losing much accuracy, ideally we can use Monte Carlo integration to calculate $E_{\theta}[\log f(Y_i|\theta)|Y_{-i}]$ by generating samples $\theta^{(l)}_i \sim p(\theta|Y_{-i})$. So one option for computing BPDC in this case is to generate $\theta^{(l)}_i$ from 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. Referring to Smith et al. (1997), the basic idea of IS is that some of the input random variables are more important on the target being estimated than others. IS method first computes a biased estimator based on a baseline distribution, and corrects it with approximate weight given by the sampling density ratio.

Suppose we want to compute the expectation \( E_{p_i}[g(\theta)] = \int g(\theta)p_i(\theta)d\theta \) with respect to the density \( p_i(\theta) \) for \( i = 1, \ldots, I \). 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 \( p_i(\theta) = q_i(\theta)/C \) are known only by their kernel functions \( q(\theta) \) and \( q_i(\theta) \) respectively, then \( E_{p_i}[g(\theta)] \) can be estimated consistently by

\[
\bar{g}_L = \sum_{l=1}^{L} g(\theta^{(l)})\bar{w}_i(p_i, p),
\]

where \( \theta^{(l)} \overset{\text{iid}}{\sim} p(\theta), \bar{w}_i(p_i, p) = w_i(p_i, p)/\sum_{h=1}^{L} w_h(p_i, p) \), and \( w_i(p_i, p) = q_i(\theta^{(l)})/q(\theta^{(l)}) \) (Peruggia 1997). The strong law of large number implies the consistency of the estimator, but its performance depends critically on the variance of the IS weight \( w_i(p_i, p) \). If the variance of the IS weight is infinite, the IS estimate become unstable. Also, IS approximation is related to case deletion posterior. If we delete a highly influential subject, the corresponding IS estimator will be adversely affected (Weiss 1996, Bradlow and Zaslavsky 1997, Weiss and Cho 1998). Particularly for a standard Bayesian linear regression model, Peruggia (1997) proved necessary and sufficient conditions for finite variance of IS weights.

With the above discussion, we can apply IS method to compute BPDC. Suppose a Bayesian model \( f(Y, \theta) \) where \( Y_i \)'s are mutually independent given \( \theta \), a data set \( Y = (Y_1, \ldots, Y_I) \) where \( Y_i \) can be a scaler or a vector, and MCMC samples \( \theta^{(l)}, l = 1, \ldots, L, \)
based on the full data $Y$, i.e. $\theta^{(l)} \sim p(\theta|Y)$. Then,

$$p(\theta|Y) = p(Y|\theta)p(\theta)/p(Y)$$
$$= p(Y_{-i}|\theta)f(Y_i|\theta)p(\theta)/p(Y),$$

$$p(\theta|Y_{-i}) = p(Y_{-i}|\theta)p(\theta)/p(Y_{-i}).$$

Let $p(\theta) = p(\theta|y)$, $p_i(\theta) = p(\theta|Y_{-i})$, $C = p(Y)$, and $C_i = p(Y_{-i})$. We get $w_l(p_i, p) = 1/f(Y_i|\theta^{(l)})$, and hence

$$w_l(p_i, p) = \frac{1/f(Y_i|\theta^{(l)})}{\sum_{h=1}^{L} 1/f(Y_i|\theta^{(h)})} = \left(\sum_{h=1}^{L} \frac{f(Y_i|\theta^{(h)})}{f(Y_i|\theta^{(l)})}\right)^{-1}. \quad (3.21)$$

Finally, we can approximate BPDC by

$$BPDC_a = -2 \sum_{i=1}^{I} \sum_{l=1}^{L} \log f(Y_i|\theta^{(l)})\bar{w}_l(p_i, p), \quad (3.22)$$

$$= -2 \sum_{i=1}^{I} \sum_{l=1}^{L} \log f(Y_i|\theta^{(l)}) \left(\sum_{h=1}^{L} \frac{f(Y_i|\theta^{(h)})}{f(Y_i|\theta^{(l)})}\right)^{-1}. \quad (3.23)$$

In the rest of this dissertation, we will compute $BPDC_a$ as the value of $BPDC$. To explore the numerical performance of BPDC, we carry out a simulation study to compare BPDC with DIC and LPML.

### 3.3 Simulation Study

In this section, various simulated data scenarios are used to compare the performance of our proposed model selection criterion, BPDC, with other two popular model selection criteria, DIC and LPML. We mainly consider linear mixed models since our target is to
select model from a class of linear mixed models. Also, a supplementary simulation is conducted to show the approximation to a Gaussian process with nonlinear mean and covariance functions using Bernstein polynomials.

### 3.3.1 Data Generation: Design and Setup

In our simulation study, we generally considered a class of Bayesian linear mixed models discussed in Chapter 2,

\[
Y_{ij} = \sum_{k=1}^{m} b_{km}(t_{ij}) \beta_{ik} + \epsilon_{ij}, \quad i = 1, \ldots, I; j = 1, \ldots, J_i
\]  

with priors,

\[
\beta_0 \sim N(\tilde{\beta}_0, \sigma_0^2 I),
\]

\[
D \sim IWish(R, v),
\]

\[
\sigma_i^2 \sim IG(a, b),
\]

where \(\{b_{km}(\cdot)\}_{k=1}^{m}\) are Bernstein polynomial basis of degree \(m-1\), \(\beta_i = (\beta_{i1}, \ldots, \beta_{im})^T\), \(\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T\), \(\beta_0\) is a \(m\)-vector of fixed effects, and \(D\) is a \(m \times m\) nonnegative matrix. Time points \((t_{i1}, \ldots, t_{iJ_i})\) are determined as an arithmetic sequence of length \(J_i\) between 0.01 and 0.98 for all \(i\)'s. For priors, we set \(\tilde{\beta}_0 = 0, \sigma_0^2 = 100, R = 0.01I, v = m + 1,\) and \(a = b = 0.01\) all through the simulation study. As we discussed in Section 2.3.1, the value of \(m\) can be ranged from 2 up to \([\max(J_i, i = 1, \ldots, I)^{3/4}]\).

Three factors that are of concern in this simulation study are sample size, number of repeated measurements for each subject, and the magnitude of random errors. Sample
size is always an important factor in simulation study. Some criteria may require larger sample size than other methods to perform well. We chose sample size $I = 50, 100, 200$. In practice, longitudinal data may have very small or large number of repeated measurements $J_i$ for each subject, where $J_i$ can remain the same or vary among subjects. To keep the simulation relatively tractable, we only considered constant $J$, with three levels $J = 5, 10, 30$. Larger $J$ implies more information, and therefore corresponds to better fit of models. However, larger $J$ is also related to more candidate models, which increases the difficulty of model selection. The random error $\sigma_i^2$ can be homogeneous or heterogeneous, and its magnitude can be small or large. We picked two levels of magnitude $\sigma_i^2 = 0.01, 0.25$, and $\sigma_i^2$ generated from two inverse gamma distributions with population mean at 0.01 and 0.25 to study the effect of the random error term.

Including all three factors in our simulation study, we decided to compare the following five cases:

- Case 1: $I=50$, $J_i = 10$ for all $i$, $\sigma_i^2 \overset{iid}{\sim} IG(5, 1)$, a distribution with mean 0.25, and

$$\beta_0 = (1, -0.7, 2)^T,$$

$$D = \begin{pmatrix}
1 & 0 & 0.2 \\
0 & 0.7 & 0 \\
0.2 & 0 & 1.2 \\
\end{pmatrix}. $$
• Case 2: I=200, \(J_i = 10\) for all \(i\), \(\sigma_i^2 \overset{\text{iid}}{\sim} IG(5, 1)\), a distribution with mean 0.25, and

\[
\beta_0 = (1, -0.7, 2)^T,
\]
\[
D = \begin{pmatrix}
1 & 0 & 0.2 \\
0 & 0.7 & 0 \\
0.2 & 0 & 1.2
\end{pmatrix}.
\]

• Case 3: I=50, \(J_i = 10\), \(\sigma_i^2 = 0.25\) for all \(i\), and

\[
\beta_0 = (1, -0.7, 2)^T,
\]
\[
D = \begin{pmatrix}
1 & 0 & 0.2 \\
0 & 0.7 & 0 \\
0.2 & 0 & 1.2
\end{pmatrix}.
\]

• Case 4: I=100, \(J_i = 5\) for all \(i\), \(\sigma_i^2 \overset{\text{iid}}{\sim} IG(101, 1)\), a distribution with mean 0.01, and

\[
\beta_0 = (1, 2)^T,
\]
\[
D = \begin{pmatrix}
0.25 & -0.01 \\
-0.01 & 0.25
\end{pmatrix}.
\]
• Case 5: \( I=50, J_i = 30, \sigma_i^2 = 0.01 \) for all \( i \), and

\[
\beta_0 = (1, 0.5, 0.8, -0.7)^T,
\]

\[
D = \begin{pmatrix}
0.01 & 0 & 0 & 0 \\
0 & 0.01 & 0 & 0 \\
0 & 0 & 0.01 & 0 \\
0 & 0 & 0 & 0.01
\end{pmatrix}.
\]

The data were generated from (3.24) with some fixed \( m, \beta_0, D, \) and \( \sigma_i^2 \)'s. We performed 200 simulations for every specific case. For fitting Bayesian models, WinBUGS14 was used to perform the MCMC methods, where the first 3000 samples were dropped as burn-in samples and 3000 samples were kept for posterior inference. Four model selection criteria were recorded: (1) conditional DIC, (2) marginal DIC, (3) LPML, and (4) BPDC. At every value of \( m \), we computed the Monte Carlo means and Monte Carlo standard errors of criteria. Also, we recorded the proportion of correct model selection based on different criteria for every trial.

### 3.3.2 Results and Discussion

We summarized the results of simulation study in Tables 3.1, 3.2, 3.3 and 3.4. Tables 3.1 and 3.2 listed the empirical probability out of 200 trials that we chose a model with tuning parameter \( m \) based on different criteria. Table 3.3 and 3.4 summarizes the average values of criteria over 200 trials and their standard errors.

From Tables 3.1 and 3.2, we can see that BPDC attains the highest probability of correct decision making for Case 1-3 and Case 5. Especially for Case 1 and Case 3 where a moderate number of observations are recorded for every subject, it outperforms the
second best one, LMPL, in terms of proportion of correct choices by about 10%. In Case 5 where the true value of $m$ is 4, the choice made based on BPDC only overfit the model to $m = 5$ and $m = 7$ with percentage of 3.5% and 0.5% respectively. Other criteria are more likely to overfit the model and choose models with higher orders up to $m = 9$ and even $m = 12$. It demonstrates that even with more candidate models involved, BPDC preserve high accuracy in choosing the correct model, and its performance is improved along with larger $J$ and smaller constant error. In Case 4, both BPDC and LPML show a pretty high ability of model recognition, where over 99% of time the correct model is chosen based on them. But comparing Case 1 and Case 2, interestingly we notice that in case of heterogenous errors increasing the sample size to a large number, e.g. 200, does not help BPDC and LPML, particularly for BPDC, whose accuracy drops a little bit.

On the other hand, the conditional DIC and marginal DIC work much worse for the model with heterogeneous large random errors. Note that in Case 1 and Case 2, the probability of choosing the correct model based on them are only around 30%, and sometimes even worse, less than 20%. Meanwhile, BPDC and LPML enjoy a high accuracy of over 60%. But conditional DIC works better in case of homogeneous random errors, say in Case 3. Its accuracy goes up to 59.5%, which is slightly higher than that of LPML, 58.5%, though still much less than that of BPDC, 76.5%. This agrees with our discussion in Section 3.2.1 that marginal DIC assumes fixed number of parameters, not going to infinity as sample size increases, when the error variance is constant. However, checking with Case 1, Case 2 and Case 4, we are surprised to find that the condition reverses, say conditional DIC outperforms marginal DIC. It is probably because the conditional DIC includes $\beta_i$ as part of the parameters which makes it more informative. Especially in Case 4, despite that only 5 measurements taken on each subject, conditional DIC has a high accuracy rate of 94%, being close to BPDC and LPML in the case. This excellent
performance may be due to the increase of sample size and limited number of model candidates.

In Tables 3.3 and 3.4, we use bold font to highlight the smallest mean value among all candidate models for every criterion. It clearly shows that BPDC and LPML always obtain their smallest value at the correct model in terms of criterion means. Meanwhile, DIC’s tend to obtain the smallest value in model with larger m, i.e. more complex model. That explains the better performance of these two cross-validation based criteria, and also warn us that DIC’s potentially lead to over-fitted models. For BPDC in Case 1 and 3, we can get convex curves with the lowest points at $m = 3$ which is the true value of m if we plot BPDC versus m. For LPML, it also forms convex curves for Case 1 and 3 but the shape is not that obvious compared to BPDC. This indicates that BPDC is better than DIC in differentiating between the two models. Besides that, the four criteria are within approximately the same range for Case 1-3 and Case 5, and the standard deviations of all the four criteria are very close, where BPDC and conditional DIC are a little higher than the other two. Especially for Case 3 and Case 5, BPDC, LPML, and marginal DIC fall in the range of $[1036.29, 1242.55]$ and $[-2394.70, -1111.26]$ respectively, with their standard deviation in $[2.35, 2.63]$ and $[3.67, 4.40]$ respectively. The marginal DIC in Case 3 and Case 5 has been proved to approximate AIC. So our simulation results numerically show the approximation of BPDC to AIC for linear mixed models. But in Case 4 where we only have 5 observations for every subject, BPDC and LPML differ a lot from marginal DIC and conditional DIC, where the former two are above 100 and the last two are below -100. In this case, the marginal DIC and conditional DIC are not reliable as they produce negative values.

With all the discussions above, we come to a few conclusions:
1. BPDC is the best among all four criteria in terms of choosing the correct model among our class of linear mixed models irrespective of homogeneous or heterogeneous errors. It approximates AIC under some classical setting, e.g., linear mixed effects models with constant error term.

2. The magnitudes of variation are very close for all four model selection criteria.

3. Cross-validation based criteria, BPDC and LPML, performs better than DIC’s for linear mixed models.

4. Marginal DIC works better for models with homogeneous errors while conditional DIC works better for models with heterogeneous errors.

### 3.3.3 Supplementary Simulation

We have demonstrated BPDC is superior to other two Bayesian criteria, DIC and LPML, in choosing the tuning parameter $m$ in LMM-BPS in the previous simulation study. In this subsection, we used BPDC as our model selection tool, and carried out a simple simulation study to show the approximation to a Gaussian process with nonlinear mean and covariance functions using Bernstein polynomial sieve.

Suppose the true model is (1.10) with $\sigma^2 = 0.01$. We choose $\mu(t) = 0.001\{\exp(12.142 - 6.188t) + \exp(7.624 + 0.488t)\}$ which is motivated by Wu and Ding (1999), and $K(t, s) = \exp\{-\frac{(\Phi^{-1}(t) - \Phi^{-1}(s))^2}{2}\}$ where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution. We have shown in Section 2.4 that GP with the squared exponential kernel satisfies all regularity conditions in Theorem 2.4.2 for $p = 1$.

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 55.
Table 3.1: Percentage of model selection decisions (Case 1-4)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC</th>
<th>LPML</th>
<th>$DIC_m$</th>
<th>$DIC_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>75.0</td>
<td>63.0</td>
<td>26.5</td>
<td>28.5</td>
</tr>
<tr>
<td>4</td>
<td>15.0</td>
<td>18.0</td>
<td>19.0</td>
<td>22.5</td>
</tr>
<tr>
<td>5</td>
<td>8.0</td>
<td>14.0</td>
<td>20.0</td>
<td>19.5</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>5.0</td>
<td>34.5</td>
<td>29.5</td>
</tr>
<tr>
<td>Case 2 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>66.0</td>
<td>65.0</td>
<td>19.0</td>
<td>32.5</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>6.5</td>
<td>6.5</td>
<td>9.5</td>
</tr>
<tr>
<td>5</td>
<td>3.0</td>
<td>2.0</td>
<td>21.5</td>
<td>18.5</td>
</tr>
<tr>
<td>6</td>
<td>27.0</td>
<td>26.5</td>
<td>53.0</td>
<td>39.5</td>
</tr>
<tr>
<td>Case 3 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>76.5</td>
<td>58.5</td>
<td>59.5</td>
<td>43.0</td>
</tr>
<tr>
<td>4</td>
<td>13.0</td>
<td>19.5</td>
<td>19.0</td>
<td>24.0</td>
</tr>
<tr>
<td>5</td>
<td>7.0</td>
<td>10.5</td>
<td>11.0</td>
<td>15.0</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>11.5</td>
<td>10.5</td>
<td>18.0</td>
</tr>
<tr>
<td>Case 4 (m=2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>99.0</td>
<td>99.5</td>
<td>65.0</td>
<td>94.0</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>0.5</td>
<td>35.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

*a* BPDC: Bayesian predictive divergence criterion; LPML: log pseudo marginal likelihood; $DIC_m$: marginal deviance information criterion; $DIC_c$: conditional marginal deviance information criterion (computed in WinBUGS).

*b* The true values of $m$ are specified in parentheses besides the case number.
Table 3.2: Percentage of model selection decisions (Case 5)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC</th>
<th>LPML</th>
<th>$DIC_m$</th>
<th>$DIC_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>96.0</td>
<td>82.5</td>
<td>83.5</td>
<td>64.0</td>
</tr>
<tr>
<td>5</td>
<td>3.5</td>
<td>11.5</td>
<td>10.0</td>
<td>18.0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1.5</td>
<td>2.5</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>0.5</td>
<td>2.0</td>
<td>2.0</td>
<td>6.0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2.0</td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>2.5</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*BPDC: Bayesian predictive divergence criterion; LPML: log pseudo marginal likelihood; $DIC_m$: marginal deviance information criterion; $DIC_c$: conditional marginal deviance information criterion (computed in WinBUGS).* 

*a The true values of $m$ are specified in parentheses besides the case number.*
Table 3.3: Average of model selection criteria and standard errors (Case 1-3)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>$DIC_m$ (se)</th>
<th>$DIC_c$ (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1487.36 (2.83)</td>
<td>1296.24 (2.73)</td>
<td>1253.32 (2.73)</td>
<td>1168.60 (2.89)</td>
</tr>
<tr>
<td>3</td>
<td>1300.28 <strong>1108.21</strong> (2.80)</td>
<td>1059.87 (2.62)</td>
<td>926.44 (2.96)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1320.13 (2.94)</td>
<td>1114.32 (2.72)</td>
<td>1059.69 (2.66)</td>
<td>927.35 (2.98)</td>
</tr>
<tr>
<td>5</td>
<td>1329.35 (2.97)</td>
<td>1116.94 (2.70)</td>
<td>1059.87 (2.66)</td>
<td>927.49 (2.97)</td>
</tr>
<tr>
<td>6</td>
<td>1347.76 (3.12)</td>
<td>1121.46 (2.70)</td>
<td><strong>1057.86</strong> (2.63)</td>
<td><strong>924.58</strong> (2.97)</td>
</tr>
</tbody>
</table>

| Case 2 (m=3) | | | | |
| 2 | 5865.30 (5.56) | 5101.81 (5.42) | 4926.51 (5.46) | 4576.64 (6.07) |
| 3 | **5036.06** **4281.44** (4.87) | 4097.36 (4.70) | 3481.00 (5.12) |
| 4 | 5089.19 (5.00) | 4294.25 (4.72) | 4096.38 (4.72) | 3481.63 (5.13) |
| 5 | 5099.60 (5.23) | 4299.04 (4.83) | 4094.43 (4.77) | 3480.73 (5.16) |
| 6 | 5061.90 (5.07) | 4288.41 (4.72) | **4092.32** (4.77) | **3477.94** (5.15) |

| Case 3 (m=3) | | | | |
| 2 | 1242.55 (2.63) | 1234.73 (2.61) | 1233.65 (2.60) | 1133.26 (2.87) |
| 3 | **1050.63** **1037.81** **1036.29** (2.39) | 1039.50 (2.36) | 1037.81 (2.35) | 882.84 (2.77) |
| 4 | 1055.16 (2.44) | 1039.50 (2.39) | 1037.81 (2.37) | 882.27 (2.77) |
| 5 | 1057.37 (2.41) | 1040.50 (2.37) | 1038.98 (2.37) | 884.45 (2.77) |
| 6 | 1060.74 (2.43) | 1041.55 (2.37) | 1039.98 (2.35) | 883.10 (2.76) |

\*The smallest value of each criterion among all candidate models is in bold font.
Table 3.4: Average of model selection criteria and standard errors (Case 4-5)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>DIC&lt;sub&gt;m&lt;/sub&gt; (se)</th>
<th>DIC&lt;sub&gt;c&lt;/sub&gt; (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>658.39 (2.22)</td>
<td>192.73 (2.02)</td>
<td>-118.02 (2.33)</td>
<td>-312.30 (2.06)</td>
</tr>
<tr>
<td>3</td>
<td>727.58 (2.41)</td>
<td>216.40 (2.03)</td>
<td>-117.24 (2.29)</td>
<td>-305.07 (2.00)</td>
</tr>
</tbody>
</table>

Case 4 (m=2)

<table>
<thead>
<tr>
<th>m</th>
<th>DIC&lt;sub&gt;m&lt;/sub&gt; (se)</th>
<th>DIC&lt;sub&gt;c&lt;/sub&gt; (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-1111.26 (4.23)</td>
<td>-1117.35 (4.22)</td>
</tr>
<tr>
<td>3</td>
<td>-2066.70 (3.86)</td>
<td>-2078.12 (3.85)</td>
</tr>
<tr>
<td>4</td>
<td>-2378.27 (3.68)</td>
<td>-2393.49 (3.68)</td>
</tr>
<tr>
<td>5</td>
<td>-2374.06 (3.68)</td>
<td>-2391.51 (3.68)</td>
</tr>
<tr>
<td>6</td>
<td>-2369.37 (3.67)</td>
<td>-2389.21 (3.67)</td>
</tr>
<tr>
<td>7</td>
<td>-2367.12 (3.68)</td>
<td>-2388.40 (3.69)</td>
</tr>
<tr>
<td>8</td>
<td>-2363.33 (3.69)</td>
<td>-2386.48 (3.70)</td>
</tr>
<tr>
<td>9</td>
<td>-2359.04 (3.69)</td>
<td>-2384.31 (3.70)</td>
</tr>
<tr>
<td>10</td>
<td>-2356.25 (3.69)</td>
<td>-2383.00 (3.70)</td>
</tr>
<tr>
<td>11</td>
<td>-2353.12 (3.70)</td>
<td>-2381.58 (3.70)</td>
</tr>
<tr>
<td>12</td>
<td>-2351.42 (3.67)</td>
<td>-2380.81 (3.68)</td>
</tr>
<tr>
<td>13</td>
<td>-2348.11 (3.67)</td>
<td>-2379.04 (3.69)</td>
</tr>
</tbody>
</table>

Case 5 (m=4)

*a The smallest value of each criterion among all candidate models is in bold font.*
We fit Model (3.24) 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:

\[
\text{IBias}(\hat{\mu}) = \int_0^1 \{\mu(t) - \hat{\mu}(t)\} dt = \int_0^1 \mu(t) dt - \int_0^1 \hat{\mu}(t) dt,
\]

\[
\text{IBias}(\hat{K}) = \int_0^1 \int_0^1 \{K(t,s) - \hat{K}(t,s)\} dtds = \int_0^1 \int_0^1 K(t,s) dtds - \int_0^1 \int_0^1 \hat{K}(t,s) dtds,
\]

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 \( \text{IBias}(\hat{\mu}) \) and \( \text{IBias}(\hat{K}) \), only \( \int_0^1 \hat{\mu}(t) dt \) and \( \int_0^1 \int_0^1 \hat{K}(t,s) dtds \) vary with data 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, we have \( \int_0^1 \hat{\mu}(t) dt = \sum_{k=1}^m \hat{\beta}_{km} / m \) 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

\[
\text{IBias}(\hat{\mu}) = 31.902 - \sum_{k=1}^m \hat{\beta}_{km} / m,
\]

\[
\text{IBias}(\hat{K}) = 0.577 - \sum_{k_1=1}^m \sum_{k_2=1}^m \hat{D}_{k_1k_2} / m^2.
\]

Table 3.5 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

10 between 0.02 and 0.98 for all subjects.
Table 3.5: Fit of Gaussian process with nonlinear mean and covariance functions

<table>
<thead>
<tr>
<th>IBias(\hat{\mu})</th>
<th>SE of IBias(\hat{\mu})</th>
<th>p-value</th>
<th>IBias(\hat{K})</th>
<th>SE of IBias(\hat{K})</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0046</td>
<td>0.0077</td>
<td>0.5509</td>
<td>-0.0115</td>
<td>0.0085</td>
<td>0.1776</td>
</tr>
</tbody>
</table>

function and covariance function, both of which are very close to 0. We also calculated the p-value based on 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 3.1, 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. On average \( \hat{m} = 8.445 \) was selected by BPDC in this example.

### 3.4 Real Data Analysis

We have illustrated our methodology in previous sections and chapters. In this section, we apply our method to Sitka growth study and the Berkeley growth study.
Figure 3.1: 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.

### 3.4.1 Growth Curve Study using Sitka Data

The Sitka data was first introduced and analyzed in Diggle et al. (1994), and can be accessed in package “MASS” in R,

```r
> library(MASS)
> data(Sitka)
> data(Sitka89)
> Sitka.full <- rbind(Sitka,Sitka89)
```

The Sitka data contain the repeated measurements on 79 Sitka spruce trees, 54 of which were grown in ozone-enriched chambers (ozone group) and 25 of which were in a natural atmosphere (control group) over two growing seasons, Year 1988 and Year 1989. During the study, trees were visited at irregularly spaced time, at day 152, 174, 201, 227, 258, 469, 496, 528, 556, 579, 613, 639, and 674, with a big gap between Day 258 and Day 469.
The logarithm of product of diameter squared with height was recorded on every visit as the log-size of the tree. Figure 3.2 displays growth curves of the 79 spruce trees where the ozone group is in black and the control group in red.

Figure 3.2: Growth curves of Sitka spruce trees over days. Black: the ozone group. Red: the control group.

Moser et al. (2004) suggested to fit the nonlinear von Bertalanffy or Mitscherlich growth model (Schabenberger and Pierce 2001) with an additional subject-specific parameter. Their model is

\[ Y_{ij} = (A + b_i)[1 - \exp(-k(t_{ij} - t_0))] + \epsilon_{ij}, \]  
\[ b_i \sim N(0, \sigma_b^2), \]  

where \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})^T \) has a spatial power correlated errors structure. With model (3.25), the mean function of the growth curve is \( \mu(t) = A[1 - \exp(-k(t - t_0))] \). It is
a positive nondecreasing function as long as $A > 0, k \geq 0$. Though it is reasonable for growth curves, no solid evidence can support the Sitka spruce trees in both groups followed this class of models. In other words, their model is not flexible enough, and they may miss other better options. Also, they only considered the random effects on the baseline function. There could be random effects in $k$, say $k + k_i$ for different subject.

Crainiceanu et al. (2005) analyzed the Sitka data with an additive mixed model such that

$$Y_{ij} = U_i + \alpha Z_i + h(t_{ij}) + \epsilon_{ij},$$

$$U_i \overset{iid}{\sim} N(0, \sigma_u^2),$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$

where $U_i$ is random intercept for each tree, $Z_i$ is the ozone exposure indicator, and $h(\cdot)$ is modeled using low-rank thin-plate splines. They obtained the estimate of growth rate difference $\hat{\alpha} = -0.31$ with 95% credible interval $[-0.61, -0.007]$. Noticing that the curve we are fitting is growth curve, we need a monotone nondecreasing constraint on the mean function. However, Crainiceanu et al. (2005) did not consider this monotonicity in their model (3.26).

Wang (2012) utilized Bernstein polynomials instead of the low-rank thin-plate splines to pose the shape restriction on $h(\cdot)$. The model is

$$Y_{ij} = \sum_{k=1}^{m} b_{km}(t_{ij})(\beta_{0k} + Z_i \beta_{1k} + b_{ik}) + \sigma_i \epsilon_{ij},$$

$$b_i = (b_{i1}, \ldots, b_{im})^T \overset{iid}{\sim} N(0, D),$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, 1),$$

64
where some constraints are put on $\beta_{0k}$’s and $\beta_{1k}$’s to keep the nondecreasing shape of growth curves. But with Model (3.27), Wang (2012) assumed the covariance function of the two groups to be the same, which may not be true.

Our model for analyzing this Sitka data is more flexible, allowing different tuning parameters and covariance functions for different groups. The target LMM-BPS model is

$$Y_{ij} = \sum_{k=1}^{m_1} b_{km_1}(t_{ij}) \beta_{ik}^{(1)} Z_i + \sum_{h=1}^{m_0} b_{hm_0}(t_{ij}) \beta_{ih}^{(0)} (1 - Z_i) + \sigma_i \epsilon_{ij}, \quad (3.28)$$

$$\beta_i^{(a)} \overset{iid}{\sim} N(\beta^{(a)}, D^{(a)}), \quad a = 0, 1,$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, 1),$$

where we need to put nondecreasing constraints on $\beta^{(0)}$ and $\beta^{(1)}$. Equivalently, we fit the following model based on data from ozone group and control group separately.

$$Y_{ij} = \sum_{k=1}^{m} b_{km}(t_{ij}) \beta_{ik} + \sigma_i \epsilon_{ij}, \quad (3.29)$$

$$\beta_i \overset{iid}{\sim} N(\beta_0, D),$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, 1),$$

with priors

$$\alpha_0 \sim LogNormal(0, 100 \mathbf{I}),$$

$$D_0 \sim IWish(100 \mathbf{I}, m + 2),$$

$$\sigma_i^2 \sim IG(0.01, 0.01),$$

65
where $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{im})^T, \beta_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0m})^T, \alpha_0 = (\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0m})^T, \alpha_{01} = \beta_{01}$, and $\alpha_{0k} = \beta_{0k} - \beta_{0k-1}$ for $k = 2, \ldots, m$. Note that in Model (3.29), we did not put priors on $\beta_0$. Instead, we take difference between the $\beta_{0j}$ and $\beta_{0j-1}$, and put a prior defined on $(0, \infty)$ on the difference. In this way, the Bayesian estimates to the differences must be nonnegative, which guarantees that $\beta_0$, the mean of coefficients, is nondecreasing. With our discussion in Chapter 2, this reparameterization ensures that the mean function is nondecreasing. To meet the requirement that the support set of Bernstein polynomials is $[0, 1]$, a linear transformation was taken on “day” such that $t = (\text{day} - 150)/530$, so that $t \in [0, 1]$. For each group, a tuning parameter $m$ is chosen from $\{2, 3, \ldots, 8\}$. The parameter estimates and model selection criterion BPDC were computed based on 5000 MCMC samples, where we dropped the first 15000 as burn-in samples.

![BPDC for Ozone Group](image1)

![BPDC for Control Group](image2)

Figure 3.3: BPDC of models for ozone and control group
Table 3.6: BPDC of models on Sitka data

<table>
<thead>
<tr>
<th>Group</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ozone</td>
<td>1178.0</td>
<td>1155.2</td>
<td>1125.9</td>
<td>1109.0</td>
<td>1054.3</td>
<td><strong>1034.9</strong></td>
<td>1051.2</td>
</tr>
<tr>
<td>Control</td>
<td>538.2</td>
<td>525.5</td>
<td>544.2</td>
<td>513.7</td>
<td>506.3</td>
<td>512.2</td>
<td>531.8</td>
</tr>
</tbody>
</table>

Figure 3.3 plots the value of BPDC versus m for ozone group and control group. In table 3.6, BPDC reaches its minimum 1034.9 at m=7 for ozone group, and has the smallest value at m=6 with value of 506.3 for control group. So we fit Model (3.29) with $m = 7$ for ozone group and $m = 6$ for control group. Taking the sample mean of posterior samples as our Bayesian estimates, we obtained

\[
\hat{\beta}_{\text{ozone}} = (4.23, 5.60, 5.61, 5.62, 5.65, 6.17, 6.30)^T
\]

and

\[
\hat{\beta}_{\text{control}} = (4.40, 6.06, 6.10, 6.12, 6.25, 6.65)^T
\]

Left panel of Figure 3.4 gives the plots of the estimated mean functions of growth curves,

\[
\hat{E}[Y|t] = \text{median}\{\sum_{k=1}^{m} b_{km}(t)\beta_{0k}, l = 1, \ldots, 5000\},
\]

along with observations. It is clear that the mean growth function of ozone group is always below that of the control group. This plot also suggests that our estimated mean functions successfully capture the trends in the data. We also computed the coefficient of determination, which is defined as

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

where $\hat{Y}_{ij}$ is the posterior predictive mean of $Y_{ij}$ and $\bar{Y}$ denotes the overall mean. With our fitted models, we have $R^2 = 0.971$, which confirms good fit with the data. The plot in the right panel of Figure shows the estimated mean difference function between the two groups and its pointwise 95% credible interval. The fixed treatment effect obtained
in Crainiceanu et al. (2005) is also added to the plot as a dashed horizontal reference line. Compared with their result, ours is around the same level as theirs, but has some ups and downs along the time line, which demonstrates the fact that the mean functions of two group can never be parallel all the time. The difference between the growth curves from the two groups is statistically significant between Day 300 and Day 550. Figure 3.5 displays the estimated covariance functions for ozone group and control group. Compared to the well-known kernels in Figure 1.2, none of them matches the estimated ones. This also advocates the flexibility of LMM-BPS in capturing different Gaussian processes.

Figure 3.4: Left: estimated mean function over days, and original observations displayed to show the fit of the model. Right: estimated difference function and its 95% credible interval where the difference function is the mean function of ozone group minus the mean function of control group and the grey line is the reference line at y=0. The fixed treatment effect obtained by Crainiceanu et al. (2005) is shown in blue dashed line at y=0.31.
However, comparing the growth curves may not be the best way for detecting the treatment effects, since the treatment effects are accumulated as time goes on. The analysis result that the difference between two groups is not seen at the beginning phase, say the first 300 days, may be not because there is no treatment effect, but because the treatment effects are not accumulated enough. So it is more reasonable that we compare the growth rates of the two groups to detect the treatment effects. With discussion in Section 2.3.1, we can make inference on the derivative of the Gaussian process, the growth rates in our case, with LMM-BPS. We computed the estimated mean function of growth rates using (2.4). The plot on the left panel of Figure 3.6 shows the mean functions of growth rates for both the ozone group and the control group. The growth rates of the two groups start around 8, decrease in the first year, and stabilize at around 1 through the second year. The plot on the right panel of Figure 3.6 displays the mean function of the difference between the growth rates of the two groups along with pointwise 95%
credible interval. It shows that before Day 400, the difference is negative (the growth rates of the control group is higher than that of the ozone group), while after that day the difference of the growth rates between the two groups snakes around 0. Especially between Day 250 and Day 370, the 95% credible interval is below 0, implying that the growth rates of the ozone group is significantly smaller that that of the control group. In conclusion, the baby Sitka spruce trees grown in the ozone-enriched environment are growing slower than those in the natural environment. Especially in the last quarter of the first year, the negative treatment effects are statistically significant. But the ozone treatment does not have much impact on those adult ones.

Figure 3.6: Left: estimated mean function of growth rate over days. Right: estimated difference function of growth rate and its 95% credible interval and the grey line is the reference line at y=0.
3.4.2 Analysis of Berkeley Growth Study

Another interesting data set is the Berkeley growth data which monitor the growth of children from age 1 to age 18. In the study, the heights of 39 males and 54 females were collected at irregularly spaced time points. It is accessible in R:

```r
> library(fda)
> data(growth)
```

Figure 3.7 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 adopted the same model we used in the previous real data analysis, i.e. Model (3.29). The linear transformation \( t = \frac{age}{20} \) 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 \( m = 2, \ldots, 15 \), plotted in Figure 3.8. It is clear that BPDC reach its minimum at \( m=7 \) in the female group and \( m=14 \) in the male group. The prediction based on these picked tuning parameters is pretty good demonstrated by a high The coefficient of determination,

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

Figure 3.9 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 3.10 shows the estimated growth rates as
Figure 3.7: Growth curves of females and males from age 1 to 18. Black: the female group. Red: the male group.

Figure 3.8: BPDC of models for female and male group.
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.

![Estimated Growth Curves](image1)

![Diff of Mean between Female and Male](image2)

Figure 3.9: 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.

### 3.5 Discussion

We proposed a predictive divergence based criterion BPDC for Bayesian models. This proposed criterion is demonstrated to be better than two popular Bayesian criteria, DIC and LPML, in terms of choosing the correct model for linear mixed models in a carefully designed simulation study. An supplementary simulation study demonstrates that the Bernstein polynomial approximation to a Gaussian process with nonlinear mean and
nonlinear covariance functions is very good in terms of both integrated bias and pointwise bias. In the real data analysis, we analyzed the Sitka data and Berkeley growth data with our proposed LMM-BPS and the newly constructed model selection criterion, BPDC. Model estimation is given and discussed. In analysis of Sitka data, conclusions about the growth of Sitka spruce trees are made, successfully revealing the negative effects of ozone treatment on baby Sitka spruce trees, especially in the last quarter of the first year. Similar analysis was carried out on the Berkeley growth, and interesting results that females grow significantly faster than males between age 10 to 14 is found.
Chapter 4

Longitudinal Analysis Subject to Data Irregularity

4.1 Introduction

In previous chapters, we have proposed a class of linear mixed effects models based on Bernstein polynomial sieve for approximating Gaussian processes and a new Bayesian model selection criterion, BPDC, to determine the size of the sieve. All proposed methodologies are based on the fact that we obtain the complete longitudinal data. However, it is common that the longitudinal data may have missing or censored values. These data irregularities must be taken into account when conducting appropriate statistical analysis. Otherwise, ignorance or improper analysis will lead to biased results and unreliable inference.

Missing values in longitudinal data may be caused by missing visits, withdrawal from a study, losses to follow-up, unexpected event, and so on. When dealing with missing values, we first need to identify the probabilistic framework for missing mechanism. Little
(1995) provided a good summary of the missing mechanism in longitudinal data studies. Suppose the \( j - \)th measurement for the \( i - \)th subject is \( Y_{ij} \), missing indicator is \( r_{ij} \) (e.g. \( r_{ij} = 1 \) if \( Y_{ij} \) is observed, otherwise \( r_{ij} = 0 \)), and the vector of predictors is \( X_{ij} \). Then the entire collection of data is \((Y, R, X)\) where some components of \( Y_{ij} \)'s may have missing values. Let \( Y_{obs} \) denote the collection of all observed measurements, i.e. \( Y_{obs} = \{Y_{ij} : r_{ij} = 1, i = 1, \ldots, J; j = 1, \ldots, I_i\} \). Little (1995) described three types of missing mechanisms.

1. Missing completely at random (MCAR). The missing mechanism is not related to the observed variable \( Y \), say, \( P(r = 1 | Y, X) = P(r = 1 | X) \).

2. Missing at random (MAR). The missing mechanism only depends on the observed response, i.e. \( P(r = 1 | Y, X) = P(r = 1 | Y_{obs}, X) \).

3. Missing not at random (MNAR). The missing mechanism is allowed to depend on unobserved response. In this case, \( P(r = 1 | Y, X) \) cannot be simplified. It is also called nonignorable.

MNAR is beyond our research for now, for which no universal method is available. There exists some literature focusing on MNAR such as Little (1994), Cowles et al. (1996), Michiels et al. (2002), and Verbeke and Molenberghs (2009). For MCAR and MAR, we can employ various imputation algorithms or models to compensate for the missing information. Actually, MCAR is a special case of MAR, so it is enough to consider MAR in our study. For MAR, several methods are generally used for handling missing values: (1) complete case method which is only valid for MCAR; (2) imputation methods such as the last observation carried forward (LOCF; Mallinckrodt et al. 2001) and multiple imputation (MI; Rubin 1987); (3) likelihood-based methods such as direct maximum likelihood method (Mallinckrodt et al. 2003) and EM algorithm (Dempster et al. 1977); (4)
Bayesian methods (Daniels and Hogan 2008) where missing values are imputed using the posterior predictive distribution

\[
P(Y_{mis} | Y_{obs}) = \int P(Y_{mis} | Y_{obs}, \theta) P(\theta | Y_{obs}) d\theta.
\] (4.1)

The Bayesian methods can be implemented with WinBUGS/OpenBUGS even without changing the original model specification but simply including ‘NA’s at places where \( Y_{ij} \)’s are missing. So our proposed LMM-BPS model can handle missing data naturally by imputing the missing values from the corresponding posterior predictive distribution as given in (4.1).

Censoring in longitudinal data usually arises when assay measurements are subject to quantification limits. With the lower and/or the upper detection limits of an assay, data could be left, right, or interval censored. Examples can be found in Barletta et al. (2004), Moulton and Halsey (1995), Singh and Nocerino (2002), and David M. Vock (2011). In existing literature, one simple way to handle the censoring is to use the proportion of the censored measurements as the predictor (Hammer et al. 2002), which leads to unbiased analysis but inefficient estimates. More importantly, the detection limits are only related to the assay, and may not be clinically meaningful. Another strategy is to substitute the censored data with imputed values. Ad hoc procedures use the whole or half quantification limit as a substitution, but produce biased estimates. Paxton et al. (1997) suggested an iterative imputation procedure, but ignored the correlated structure of the data and failed to adjust the variance of the parameter estimates. Hughes (1999) modified the usual EM estimation procedure to obtain maximum likelihood estimates for mixed effects model to account for censoring, which is known as Monte Carlo Expectation Maximization (MCEM) algorithm. Jacqumin-Gadda et al. (2000) proposed a full likelihood
approach for mixed model on left-censored Gaussian data, with a combination of the Simplex algorithm and the Marquardt algorithm. They showed that the likelihood of the left-censored data has the form that

\[ L(\theta) = \prod_{i=1}^{J} f(Y_{\text{obs},i}|\theta)P(Y_{\text{cen},i} < c_i 1|Y_{\text{obs},i}, \theta), \]  

(4.2)

where \( Y_{\text{obs},i} \) denotes the vector of observed data for subject \( i \), \( Y_{\text{cen},i} \) denotes the vector of potential observed values when data are censored for subject \( i \), \( c_i \) is the censored value (e.g. detection limit) for subject \( i \), and \( \theta \) is a vector of unknown parameters in the model.

The way that Bayesian methods handle censored data is very similar to this full likelihood approach. In the Bayesian model, we impute \( Y_{\text{cen},i} \) from a truncated distribution

\[ P_{c_i}(Y_{\text{cen},i}|Y_{\text{obs},i}, \theta) = \frac{P(Y_{\text{cen},i}|Y_{\text{obs},i}, \theta)I(Y_{\text{cen},i} < c_i 1)}{\int_{Y_{\text{cen},i} < c_i 1} P(Y_{\text{cen},i}|Y_{\text{obs},i}, \theta) dY_{\text{cen},i}}. \]  

(4.3)

In WinBUGS/OpenBUGS, we can use a truncation operator “I( ,c_i)” to sample from a left truncated distribution with cutoff value \( c_i \). Then, WinBUGS/OpenBUGS will automatically sample \( Y_{\text{cen},i} \) from the distribution with density function (4.3).

With data irregularities in longitudinal data, not only we need to accommodate the missing/censored mechanism into our statistical model and related estimating methods, but also the model selection criteria need to be modified. Celeux et al. (2006) presented a few possible extensions of DIC for missing data models and compared their performances. Daniels et al. (2012) proposed a criterion using the posterior predictive distribution for incomplete longitudinal data, which is shown to have computational benefits over DIC. However, the performance of their proposed criterion is not as good as DIC that is based on the observed data likelihood. Gelfand and Ghosh (1998) proposed a criterion
using a minimum posterior predictive loss approach based on independent data, and they extended the criterion to handle censored observations. We are going to extend BPDC for longitudinal analysis with data irregularities.

This chapter is organized as follows. In Section 4.2, we propose a modified BPDC to handle missing and censored values. Then, the modified BPDC is compared to modified DIC and LPML in a variety of scenarios within a simulation study in Section 4.3. Results obtained from our simulation study demonstrate the superiority of the modified BPDC to the other two. In Section 4.4, we applied our proposed methodologies to two real data sets, one of which contains missing values and the other of which contains left-censored values. Finally, Section 4.5 summaries the chapter with some discussions and conclusions.

4.2 Modified Bayesian Model Selection Criteria for Missing and Censored Data

As we have introduced in Section 3.2, DIC and LPML are two popular Bayesian model selection criteria used in practice. We also proposed a new Bayesian model selection criteria based on predictive divergence, BPDC, which is shown to be superior to DIC and LPML in choosing the size of the LMM-BPS within the assumed framework of our simulation study. In this section, we extend the definition of BPDC to select the model in the presence of missing and/or censored values, and propose the modified BPDC. Also, we present versions of modified DIC and LPML for selecting models in the presence of missing/censored data models.

Suppose missing data and censored data are handled using Bayesian methods that are discussed in Section 4.1. We discuss modified DIC, modified LPML, and modified
4.2.1 Modified DIC

For model with missing data, Celeux et al. (2006) proposed eight different DICs which depend on whether or not the likelihood includes imputed values. Generally, these eight DICs can fall into three types, observed DICs, complete DICs, and conditional DICs.

1. The observed DICs are related to the observed likelihood function
   \[ f(Y_{\text{obs}}|\theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}}. \]

2. The complete DICs are defined on the complete distribution \( f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \) where \( Y_{\text{mis}} \) is sampled from (4.1).

3. The conditional DICs are defined on the conditional distribution \( f(Y_{\text{obs}}|Y_{\text{mis}}, \theta) \), where \( Y_{\text{mis}} \) is again sampled from (4.1).

Celeux et al. (2006) compared these eight DICs and suggested an observed DIC (DIC\(_3\) in their paper) and a complete DIC (DIC\(_4\) in their paper) were the best two in terms of correctly selecting the true model. For computational convenience,

\[
\text{DIC}_3 = -4E_\theta[\log f(Y_{\text{obs}}|\theta)|Y_{\text{obs}}] + 2 \log \hat{f}(Y_{\text{obs}}) \tag{4.4}
\]

is chosen as our modified DIC for missing data models, denoted as \( DIC_m \). It can be estimated with

\[-4 \frac{1}{L} \sum_{l=1}^{L} \log f(Y_{\text{obs}}|\theta^{(l)}) + 2 \log \frac{1}{L} \sum_{l=1}^{L} f(Y_{\text{obs}}|\theta^{(l)}) ,\]

where \( \theta^{(l)} \sim p(\theta|Y_{\text{obs}}) \) for \( l = 1, \ldots, L \), and \( L \) is chosen large enough for MCMC convergence.
In case of left-censored data, we can modify the definition of DIC (Equation 3.3) by replacing \( \log f(Y|\theta) \) with the log-likelihood function suggested in Jacqmin-Gadda et al. (2000) such that \( \sum_{i=1}^{I} \log \{ f(Y_{obs,i}|\theta)P(Y_{cen,i} < c_i|Y_{obs,i}, \theta) \} \). Then, the modified DIC for left-censored data models is

\[
\text{DIC}_c = -4 \sum_{i=1}^{I} E_{\hat{\theta}}[\log \{ f(Y_{obs,i}|\theta)P(Y_{cen,i} < c_i|Y_{obs,i}, \theta) \}|\theta]|Y_{obs}] + 2 I \sum_{i=1}^{I} \log \{ f(Y_{obs,i}|\hat{\theta})P(Y_{cen,i} < c_i|Y_{obs,i}, \hat{\theta}) \},
\]

where \( \hat{\theta} = E[\theta|Y_{obs}, c] \) and \( c = (c_1, \ldots, c_I)^T \). With MCMC samples \( \theta^{(l)} \sim p(\theta|Y_{obs}, c) \), we can estimate \( \text{DIC}_c \) with

\[
-4 \frac{1}{L} \sum_{i=1}^{I} \sum_{l=1}^{L} \log \{ f(Y_{obs,i}|\theta^{(l)})P(Y_{cen,i} < c_i|Y_{obs,i}, \theta^{(l)}) \} + 2 \sum_{i=1}^{I} \log \{ f(Y_{obs,i}|\tilde{\theta})P(Y_{cen,i} < c_i|Y_{obs,i}, \tilde{\theta}) \},
\]

where \( \tilde{\theta} \) denotes \( \frac{1}{L} \sum_{l=1}^{L} \theta^{(l)} \). If values are right-censored or interval-censored, we can extend the definition (4.5) by replacing \( P(Y_{cen,i} < c_i|Y_{obs,i}, \theta) \) with \( P(Y_{cen,i} > c_i|Y_{obs,i}, \theta) \) or \( P(c_i,1 < Y_{cen,i} < c_{i,2}|Y_{obs,i}, \theta) \), respectively.

### 4.2.2 Modified LPML

To modify LPML in the presence of missing and censored data, we can borrow the ideas that Celeux et al. (2006) used in modifying DIC. For data missing at random, the
modified LPML defined on observed likelihood is

\[
LPML_m = -2 \sum_{i=1}^{I} \log \int f(Y_{obs,i}|\theta, Y_{obs,-i}) p(\theta|Y_{obs,-i}) d\theta,
\]

(4.6)

where \( Y_{obs,-i} \) denotes the observed leave-one-out data. Use the results similar to those in Section 3.2.2, LPML\(_m\) can be estimated by

\[
\hat{LPML}_m = 2 \sum_{i=1}^{I} \log \left[ \frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(Y_{obs,i}|\theta^{(l)})} \right],
\]

where \( \theta^{(l)} \) are samples obtained from \( p(\theta|Y_{obs}) \).

If data are left-censored, the modified LPML is then defined as

\[
LPML_c = -2 \sum_{i=1}^{I} \log \int f(Y_{obs,i}|\theta) P(Y_{cen,i} < c_{i-1} | Y_{obs,i}, \theta) p(\theta|Y_{obs,-i}, c_{-i}) d\theta,
\]

(4.7)

where \( c_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_I)^T \). We can again estimate LPML\(_c\) with

\[
\hat{LPML}_c = 2 \sum_{i=1}^{I} \log \left[ \frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(Y_{obs,i}|\theta^{(l)})} P(Y_{cen,i} < c_{i-1} | Y_{obs,i}, \theta^{(l)})} \right],
\]

where \( \theta^{(l)} \) are samples from \( p(\theta|Y_{obs}, c) \). Similar extensions to right-censored data and interval-censored data can be accomplished easily.

### 4.2.3 Modified BPDC

For data missing at random, we have tried to extend BPDC in three ways that was introduced in Section 4.2.1. In a small simulation study, the observed BPDC was found superior to the conditional BPDC and complete BPDC. Thus, we define BPDC for miss-
ing data models as

$$\text{BPDC}_m = \sum_{i=1}^{I} \int -2 \log f(Y_{obs,i} | \theta)p(\theta | Y_{obs,-i}) d\theta. \quad (4.8)$$

where $Y_{obs,-i}$ denotes the observed leave-ith-out data. Applying importance sampling method, we can estimate BPDC with

$$\overline{\text{BPDC}}_m = -2 \sum_{i=1}^{I} \sum_{l=1}^{L} \log f(Y_{obs,i} | \theta^{(l)}) \bar{w}_m^l (p_i, p),$$

where $\theta^{(l)}$ are samples from $p(\theta | Y_{obs})$, and

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

Notice that since we have missing values in the data set, $f(Y_{obs,i} | \theta)$ is the marginal density function of $Y_{obs,i}$ after integrating out the $Y_{mis,i}$.

For left-censored data, BPDC is defined as

$$\text{BPDC}_c = \sum_{i=1}^{I} \int -2 \log f(Y_{obs,i} | \theta)P(Y_{cen,i} \leq c_i 1 | Y_{obs,i}, \theta)p(\theta | Y_{obs,-i}, c_{-i}) d\theta. \quad (4.9)$$

It can be estimated by

$$\overline{\text{BPDC}}_c = -2 \sum_{i=1}^{I} \sum_{l=1}^{L} \log \{f(Y_{obs,i} | \theta^{(l)})P(Y_{cen,i} \leq c_i 1 | Y_{obs,i}, \theta^{(l)})\} \bar{w}_c^l (p_i, p),$$

83
where $\theta^{(l)}$ are samples from $p(\theta|Y_{obs}, c)$, and

$$
\bar{w}_l(p_t, p) = \left( \sum_{l=1}^L \frac{f(Y_{obs,i}|\theta^{(l)}) P(Y_{cen,i} \leq c_i 1|Y_{obs,i}, \theta^{(l)})}{f(Y_{obs,i}|\theta^{(h)}) P(Y_{cen,i} \leq c_i 1|Y_{obs,i}, \theta^{(h)})} \right)^{-1}
$$

Next we compare the modified BPDC, LPML, and DIC in selecting the size of the LMM-BPS.

### 4.3 Simulation Study

In this section, we compare the performance of modified BPDC with other two popular model selection criteria, DIC and LPML in the situation where we have missing or censored values in the data. Still we consider the linear mixed model as our true model since our target is to select the size of the LMM-BPS.

#### 4.3.1 Data Generation: Design and Setup

Our data is generated from a class of linear mixed models that is similar to the setting in the simulation study in Chapter 3, but now some values are subject to missing or
censored values.

\[ Y_{ij} = \sum_{k=1}^{m} b_{km}(t_{ij}) \beta_{ik} + \epsilon_{ij}, \quad i = 1, \ldots, I; j = 1, \ldots, J \quad (4.10) \]

\[ \beta_i \overset{iid}{\sim} N(\beta_0, D), \]

\[ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2 I), \]

with priors,

\[ \beta_0 \sim N(\tilde{\beta}_0, \sigma_0^2 I), \]

\[ D \sim IWish(R, v), \]

\[ \sigma^2 \sim IG(a, b), \]

We only consider the model with homogeneous errors this time. Time points \((t_{i1}, \ldots, t_{ij})\) are still determined as an arithmetic sequence of length \(J\) between 0.01 and 0.98 for all subjects. Priors share the values that we used in the simulation study in Chapter 3.

In the study, we have four factors under concern: (1) proportion of missing/censoring, (2) sample size, (3) number of repeated measurements for each subject, and (4) the magnitude of random errors. For the proportion of missing/censoring, we picked two levels: 20% and 40%. Missing mechanism in our simulation study is missing at random (MAR) following the rule such that \(P(r = 1|t) = 6pt(1 - t)\) where \(r\) is the indicator of missing and \(p\) is the missing probability parameter. Censoring mechanism covered in our study is left censoring, and right censoring and interval censoring should be similar. Two levels of sample size, 50 and 200, are used. For the number of repeated measurements for each subject, we also chose two levels, 10 and 30. Small and large magnitude of random errors are considered, \(\sigma^2 = 0.01\) or \(\sigma^2 = 0.25\). In summary, we have three cases in our study, and each setting is repeated at different levels of proportion of missing/censoring.
• Case 1: I=50, J = 10, $\sigma^2 = 0.25$, and

$$\beta_0 = (1, -0.7, 2)^T,$$

$$D = \begin{pmatrix} 1 & 0 & 0.2 \\ 0 & 0.7 & 0 \\ 0.2 & 0 & 1.2 \end{pmatrix}.$$

• Case 2: I=200, J = 10, $\sigma^2 = 0.25$, and

$$\beta_0 = (1, -0.7, 2)^T,$$

$$D = \begin{pmatrix} 1 & 0 & 0.2 \\ 0 & 0.7 & 0 \\ 0.2 & 0 & 1.2 \end{pmatrix}.$$

• Case 3: I=50, J = 30, $\sigma^2 = 0.01$, and

$$\beta_0 = (1, 0.5, 0.8, -0.7)^T,$$

$$D = \begin{pmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix}.$$

The data were generated from (4.10) with the above $\beta_0$, $D$, and $\sigma^2$. We performed 200 simulations for every specific case. WinBUGS14 was used to perform the MCMC methods, where the first 3000 samples were dropped as burn-in samples and 3000 samples were kept for posterior inference. Three model selection criteria that we suggested in
the previous section were recorded: (1) modified BPDC, (2) modified LPML, and (3) modified DIC. At every value of \( m \), we computed the Monte Carlo means and Monte Carlo standard errors of criteria. Also, we recorded the model choices based on different criteria for every trial.

4.3.2 Results and Discussion

Results for missing data models are presented in Table 4.1, 4.2, and 4.3. In all three cases, the modified BPDC has the highest percentage in choosing the correct model compared to LPML and DIC no matter missing rate is 20% or 40%. Especially for Case 3, where we have a lot of observations for every subject and the random error variance is small, the percentage of choosing the correct model with BPDC is as high as 97% and 94.5% for missing rate of 20% and 40% respectively. Comparing the cases with missing rate 20% and 40%, we do not find much difference in the percentage of model selection decisions. It is probably that because the true model is one of the candidate models. As long as we still have some observations for every subject excluding the missing values, the ability for detecting the correct model does not change much. By looking at Table 4.2 and 4.3, we can see that the standard errors of the three criteria are very close, which shows that the stability of our modified BPDC is about the same as the other two.

Results for censored data models are present in Table 4.4, 4.5 and 4.6. Similar to the missing data cases, the modified BPDC outperforms LPML and DIC in terms of choosing the correct model in all the three cases. But the proportion of censored data affects the performance of model selection criteria a lot, which is clearly shown in Case 2. In Case 2, the percentages of choosing the correct model are 88.5%, 64.5%, and 66.0% based on BPDC, LPML, and DIC respectively for data containing 20% censored values.
The percentages decrease to 56.0%, 34.5%, and 34.0% when the proportion of censored data increases to 40%. Table 4.5 and 4.6 show that the modified BPDC for censored data models is as stable as the modified LPML and DIC.

By comparing Table 4.1 and 4.4 with Table 3.1 and 3.2 in Chapter 3 where complete data is assumed, it appears that most of the proposed modified model selection criteria perform as good as the case where there is no data irregularities. Only for Case 3 in the missing data settings, the modified DIC does not perform that well as the modified BPDC and LPML.

In conclusion,

1. The modified BPDC for missing/censored data models outperforms the modified LPML and DIC disregarding the proportion of missing information, the magnitude of random error variance, and the sample size.

2. When data is missing at random, the modified BPDC performs very well in choosing the correct model even though the proportion of missing data is as high as 40%.

3. For the censored data model, large proportion of censored data results in misleading the model selection decision, especially when more subjects are enrolled in the data.

### 4.4 Real Data Study

We have discussed how to handle data subject to irregularities within our frame of LMM-BPS and the model selection criteria. In this section, we applied our methods to a National Institute of Mental Health (NIMH) Schizophrenia collaborative study on treatment related changes in overall severity, where the data have missing values. Another application
Table 4.1: Percentage of model selection decisions for missing data models (Case 1-3)

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th></th>
<th>40%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPDC</td>
<td>LPML</td>
<td>DIC</td>
<td>BPDC</td>
</tr>
<tr>
<td>m</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>72.0</td>
<td>56.5</td>
<td>40.0</td>
<td>73.0</td>
</tr>
<tr>
<td>4</td>
<td>19.0</td>
<td>22.0</td>
<td>25.0</td>
<td>17.5</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
<td>10.0</td>
<td>14.0</td>
<td>8.0</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>11.5</td>
<td>21.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Case 2 (m=3)

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th></th>
<th>40%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPDC</td>
<td>LPML</td>
<td>DIC</td>
<td>BPDC</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>80.0</td>
<td>61.0</td>
<td>38.0</td>
<td>74.0</td>
</tr>
<tr>
<td>4</td>
<td>12.5</td>
<td>18.5</td>
<td>19.5</td>
<td>14.0</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>8.5</td>
<td>18.5</td>
<td>6.5</td>
</tr>
<tr>
<td>6</td>
<td>2.5</td>
<td>12.0</td>
<td>24.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Case 3 (m=4)

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th></th>
<th>40%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPDC</td>
<td>LPML</td>
<td>DIC</td>
<td>BPDC</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>97.0</td>
<td>83.0</td>
<td>46.5</td>
<td>94.5</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>8.0</td>
<td>10.5</td>
<td>4.5</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>3.5</td>
<td>6.0</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1.5</td>
<td>5.5</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1.0</td>
<td>4.5</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2.0</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.5</td>
<td>5.5</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.5</td>
<td>5.5</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>7.0</td>
<td>0</td>
</tr>
</tbody>
</table>

a BPDC: modified Bayesian predictive divergence criterion; LPML: modified log pseudo marginal likelihood; DIC: modified marginal deviance information criterion.
b The true values of m are specified in parentheses besides the case number.
c 20% and 40% indicate the percentages of missing data.
<table>
<thead>
<tr>
<th>m</th>
<th>BPDC 20% (se)</th>
<th>LPML 20% (se)</th>
<th>DIC 20% (se)</th>
<th>BPDC 40% (se)</th>
<th>LPML 40% (se)</th>
<th>DIC 40% (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1049.68 (2.60)</td>
<td>1041.78 (2.59)</td>
<td>1036.51 (2.58)</td>
<td>844.50 (2.61)</td>
<td>836.33 (2.59)</td>
<td>831.16 (2.58)</td>
</tr>
<tr>
<td>3</td>
<td>897.00 (2.39)</td>
<td>884.42 (2.36)</td>
<td>876.54 (2.37)</td>
<td>729.73 (2.39)</td>
<td>718.10 (2.35)</td>
<td>710.28 (2.34)</td>
</tr>
<tr>
<td>4</td>
<td>900.87 (2.41)</td>
<td>886.23 (2.37)</td>
<td>877.93 (2.38)</td>
<td>733.11 (2.43)</td>
<td>719.67 (2.39)</td>
<td>710.90 (2.38)</td>
</tr>
<tr>
<td>5</td>
<td>903.91 (2.41)</td>
<td>887.40 (2.37)</td>
<td>877.98 (2.38)</td>
<td>736.00 (2.43)</td>
<td>720.98 (2.39)</td>
<td>711.53 (2.38)</td>
</tr>
<tr>
<td>6</td>
<td>905.87 (2.44)</td>
<td>888.28 (2.41)</td>
<td>878.19 (2.43)</td>
<td>738.37 (2.39)</td>
<td>721.89 (2.37)</td>
<td>711.37 (2.35)</td>
</tr>
</tbody>
</table>

Case 1 (m=3)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC 20% (se)</th>
<th>LPML 20% (se)</th>
<th>DIC 20% (se)</th>
<th>BPDC 40% (se)</th>
<th>LPML 40% (se)</th>
<th>DIC 40% (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4119.27 (5.20)</td>
<td>4112.65 (5.19)</td>
<td>4108.00 (5.19)</td>
<td>3304.55 (4.61)</td>
<td>3298.00 (4.60)</td>
<td>3293.45 (4.60)</td>
</tr>
<tr>
<td>3</td>
<td>3456.94 (4.42)</td>
<td>3445.93 (4.42)</td>
<td>3438.77 (4.43)</td>
<td>2824.64 (4.00)</td>
<td>2813.52 (3.99)</td>
<td>2806.89 (4.01)</td>
</tr>
<tr>
<td>4</td>
<td>3460.47 (4.44)</td>
<td>3447.55 (4.44)</td>
<td>3439.12 (4.45)</td>
<td>2827.86 (4.00)</td>
<td>2814.76 (3.98)</td>
<td>2807.35 (4.01)</td>
</tr>
<tr>
<td>5</td>
<td>3463.81 (4.44)</td>
<td>3448.92 (4.45)</td>
<td>3439.29 (4.45)</td>
<td>2832.09 (4.12)</td>
<td>2817.18 (4.12)</td>
<td>2809.05 (4.15)</td>
</tr>
<tr>
<td>6</td>
<td>3466.76 (4.40)</td>
<td>3450.28 (4.41)</td>
<td>3439.71 (4.41)</td>
<td>2833.30 (3.98)</td>
<td>2817.06 (3.99)</td>
<td>2807.55 (3.99)</td>
</tr>
</tbody>
</table>

Case 2 (m=3)

*The smallest value of each criterion among all candidate models is in bold font.*
Table 4.3: Average of model selection criteria and standard errors for missing data models (Case 3)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>DIC (se)</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>DIC (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-839.53</td>
<td>-845.30</td>
<td>-849.12</td>
<td>-594.78</td>
<td>-600.13</td>
<td>-603.62</td>
</tr>
<tr>
<td></td>
<td>(3.90)</td>
<td>(3.90)</td>
<td>(3.90)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.15)</td>
</tr>
<tr>
<td>3</td>
<td>-1615.38</td>
<td>-1626.55</td>
<td>-1633.99</td>
<td>-1175.62</td>
<td>-1186.00</td>
<td>-1192.93</td>
</tr>
<tr>
<td></td>
<td>(4.07)</td>
<td>(4.07)</td>
<td>(4.07)</td>
<td>(3.19)</td>
<td>(3.19)</td>
<td>(3.18)</td>
</tr>
<tr>
<td>4</td>
<td><strong>-1871.72</strong></td>
<td><strong>-1886.64</strong></td>
<td><strong>-1896.49</strong></td>
<td><strong>-1372.19</strong></td>
<td><strong>-1386.51</strong></td>
<td><strong>-1395.92</strong></td>
</tr>
<tr>
<td></td>
<td>(3.96)</td>
<td>(3.96)</td>
<td>(3.97)</td>
<td>(3.30)</td>
<td>(3.30)</td>
<td>(3.30)</td>
</tr>
<tr>
<td>5</td>
<td>-1867.50</td>
<td>-1884.64</td>
<td>-1895.83</td>
<td>-1368.52</td>
<td>-1384.83</td>
<td>-1395.43</td>
</tr>
<tr>
<td></td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.30)</td>
<td>(3.29)</td>
<td>(3.29)</td>
</tr>
<tr>
<td>6</td>
<td>-1863.90</td>
<td>-1883.00</td>
<td>-1895.32</td>
<td>-1364.94</td>
<td>-1383.10</td>
<td>-1394.73</td>
</tr>
<tr>
<td></td>
<td>(3.95)</td>
<td>(3.95)</td>
<td>(3.95)</td>
<td>(3.30)</td>
<td>(3.30)</td>
<td>(3.30)</td>
</tr>
<tr>
<td>7</td>
<td>-1860.66</td>
<td>-1881.51</td>
<td>-1894.90</td>
<td>-1361.48</td>
<td>-1381.54</td>
<td>-1394.23</td>
</tr>
<tr>
<td></td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.32)</td>
<td>(3.31)</td>
<td>(3.31)</td>
</tr>
<tr>
<td>8</td>
<td>-1856.89</td>
<td>-1879.58</td>
<td>-1893.98</td>
<td>-1358.67</td>
<td>-1380.16</td>
<td>-1393.64</td>
</tr>
<tr>
<td></td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.98)</td>
<td>(3.31)</td>
<td>(3.31)</td>
<td>(3.31)</td>
</tr>
<tr>
<td>9</td>
<td>-1854.10</td>
<td>-1878.24</td>
<td>-1893.48</td>
<td>-1355.67</td>
<td>-1378.61</td>
<td>-1393.03</td>
</tr>
<tr>
<td></td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.33)</td>
<td>(3.33)</td>
<td>(3.33)</td>
</tr>
<tr>
<td>10</td>
<td>-1851.86</td>
<td>-1877.36</td>
<td>-1893.36</td>
<td>-1353.21</td>
<td>-1377.51</td>
<td>-1392.70</td>
</tr>
<tr>
<td></td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.34)</td>
<td>(3.33)</td>
<td>(3.33)</td>
</tr>
<tr>
<td>11</td>
<td>-1849.44</td>
<td>-1876.27</td>
<td>-1893.14</td>
<td>-1350.99</td>
<td>-1376.54</td>
<td>-1392.41</td>
</tr>
<tr>
<td></td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.97)</td>
<td>(3.38)</td>
<td>(3.36)</td>
<td>(3.36)</td>
</tr>
<tr>
<td>12</td>
<td>-1847.12</td>
<td>-1875.19</td>
<td>-1892.78</td>
<td>-1349.13</td>
<td>-1375.67</td>
<td>-1392.09</td>
</tr>
<tr>
<td></td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.98)</td>
<td>(3.35)</td>
<td>(3.34)</td>
<td>(3.34)</td>
</tr>
<tr>
<td>13</td>
<td>-1845.22</td>
<td>-1874.25</td>
<td>-1892.35</td>
<td>-1347.07</td>
<td>-1374.70</td>
<td>-1391.72</td>
</tr>
<tr>
<td></td>
<td>(4.01)</td>
<td>(4.00)</td>
<td>(4.01)</td>
<td>(3.36)</td>
<td>(3.35)</td>
<td>(3.34)</td>
</tr>
</tbody>
</table>

*The smallest value of each criterion among all candidate models is in bold font.*
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>20%</th>
<th></th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>BPDC</td>
<td>LPML</td>
<td>DIC</td>
<td>BPDC</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>73.0</td>
<td>62.5</td>
<td>64.5</td>
<td>73.0</td>
</tr>
<tr>
<td>4</td>
<td>18.5</td>
<td>18.5</td>
<td>19.0</td>
<td>19.0</td>
</tr>
<tr>
<td>5</td>
<td>6.0</td>
<td>10.0</td>
<td>7.5</td>
<td>5.0</td>
</tr>
<tr>
<td>6</td>
<td>2.5</td>
<td>9.0</td>
<td>9.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Case 1 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>88.5</td>
<td>64.5</td>
<td>66.0</td>
<td>56.0</td>
</tr>
<tr>
<td>4</td>
<td>4.5</td>
<td>15.0</td>
<td>14.5</td>
<td>18.0</td>
</tr>
<tr>
<td>5</td>
<td>3.0</td>
<td>8.0</td>
<td>8.0</td>
<td>16.5</td>
</tr>
<tr>
<td>6</td>
<td>4.0</td>
<td>12.5</td>
<td>11.5</td>
<td>9.5</td>
</tr>
<tr>
<td>Case 2 (m=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>89.5</td>
<td>69.5</td>
<td>72.0</td>
<td>70.0</td>
</tr>
<tr>
<td>5</td>
<td>7.5</td>
<td>14.5</td>
<td>13.5</td>
<td>23.3</td>
</tr>
<tr>
<td>6</td>
<td>2.5</td>
<td>8.0</td>
<td>7.0</td>
<td>2.2</td>
</tr>
<tr>
<td>7</td>
<td>0.5</td>
<td>3.0</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1.0</td>
<td>2.0</td>
<td>0.6</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1.5</td>
<td>1.5</td>
<td>0.6</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1.0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Case 3 (m=4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* BPDC: modified Bayesian predictive divergence criterion; LPML: modified log pseudo marginal likelihood; DIC: modified marginal deviance information criterion.

*b* The true values of $m$ are specified in parentheses besides the case number.

*c* 20% and 40% indicate the percentages of left censoring data.
Table 4.5: Average of model selection criteria and standard errors for censored data models (Case 1-2)

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th></th>
<th>40%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPDC (se)</td>
<td>LPML (se)</td>
<td>DIC (se)</td>
<td>BPDC (se)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1192.76 (2.60)</td>
<td>1184.81 (2.59)</td>
<td>1183.49 (2.58)</td>
<td>1058.29 (3.13)</td>
</tr>
<tr>
<td>3</td>
<td><strong>1015.84</strong> (2.32)</td>
<td><strong>1003.11</strong> (2.29)</td>
<td><strong>1001.70</strong> (2.28)</td>
<td><strong>899.64</strong> (2.50)</td>
</tr>
<tr>
<td>4</td>
<td>1019.76 (2.32)</td>
<td>1004.90 (2.29)</td>
<td>1003.42 (2.28)</td>
<td>903.13 (2.48)</td>
</tr>
<tr>
<td>5</td>
<td>1022.58 (2.35)</td>
<td>1006.15 (2.31)</td>
<td>1003.79 (2.30)</td>
<td>906.24 (2.51)</td>
</tr>
<tr>
<td>6</td>
<td>1025.37 (2.29)</td>
<td>1007.39 (2.26)</td>
<td>1005.99 (2.24)</td>
<td>908.94 (2.53)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4706.58 (5.38)</td>
<td>4699.64 (5.38)</td>
<td>4698.79 (5.38)</td>
<td>4209.59 (6.33)</td>
</tr>
<tr>
<td>3</td>
<td><strong>3941.86</strong> (4.22)</td>
<td><strong>3931.13</strong> (4.22)</td>
<td><strong>3930.80</strong> (4.22)</td>
<td><strong>3393.24</strong> (6.31)</td>
</tr>
<tr>
<td>4</td>
<td>3946.32 (4.23)</td>
<td>3933.41 (4.24)</td>
<td>3933.04 (4.24)</td>
<td>3394.34 (6.41)</td>
</tr>
<tr>
<td>5</td>
<td>3949.76 (4.25)</td>
<td>3934.96 (4.25)</td>
<td>3934.60 (4.25)</td>
<td>3396.07 (6.42)</td>
</tr>
</tbody>
</table>

*a The smallest value of each criterion among all candidate models is in bold font.*
Table 4.6: Average of model selection criteria and standard errors for censored data models (Case 3)

<table>
<thead>
<tr>
<th>m</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>DIC (se)</th>
<th>BPDC (se)</th>
<th>LPML (se)</th>
<th>DIC (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-1304.19</td>
<td>-1310.19</td>
<td>-1310.63</td>
<td>-1157.93</td>
<td>-1164.20</td>
<td>-1164.55</td>
</tr>
<tr>
<td></td>
<td>(4.49)</td>
<td>(4.47)</td>
<td>(4.46)</td>
<td>(3.74)</td>
<td>(3.73)</td>
<td>(3.73)</td>
</tr>
<tr>
<td>3</td>
<td>-1684.62</td>
<td>-1694.10</td>
<td>-1694.63</td>
<td>-1193.30</td>
<td>-1201.90</td>
<td>-1202.24</td>
</tr>
<tr>
<td></td>
<td>(4.03)</td>
<td>(4.02)</td>
<td>(4.02)</td>
<td>(3.76)</td>
<td>(3.76)</td>
<td>(3.76)</td>
</tr>
<tr>
<td>4</td>
<td><strong>-1798.44</strong></td>
<td><strong>-1811.66</strong></td>
<td><strong>-1812.54</strong></td>
<td><strong>-1224.52</strong></td>
<td><strong>-1235.53</strong></td>
<td><strong>-1236.15</strong></td>
</tr>
<tr>
<td></td>
<td>(3.91)</td>
<td>(3.91)</td>
<td>(3.90)</td>
<td>(3.82)</td>
<td>(3.81)</td>
<td>(3.81)</td>
</tr>
<tr>
<td>5</td>
<td>-1795.40</td>
<td>-1810.37</td>
<td>-1811.27</td>
<td>-1222.74</td>
<td>-1234.84</td>
<td>-1235.44</td>
</tr>
<tr>
<td></td>
<td>(3.91)</td>
<td>(3.90)</td>
<td>(3.90)</td>
<td>(3.82)</td>
<td>(3.82)</td>
<td>(3.81)</td>
</tr>
<tr>
<td>6</td>
<td>-1792.91</td>
<td>-1809.28</td>
<td>-1810.12</td>
<td>-1220.46</td>
<td>-1233.73</td>
<td>-1234.32</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(3.92)</td>
<td>(3.92)</td>
<td>(3.84)</td>
<td>(3.83)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>7</td>
<td>-1789.92</td>
<td>-1807.81</td>
<td>-1808.66</td>
<td>-1218.25</td>
<td>-1232.63</td>
<td>-1233.21</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.84)</td>
<td>(3.83)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>8</td>
<td>-1787.30</td>
<td>-1806.48</td>
<td>-1807.37</td>
<td>-1216.18</td>
<td>-1231.59</td>
<td>-1232.17</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.85)</td>
<td>(3.84)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>9</td>
<td>-1785.05</td>
<td>-1805.36</td>
<td>-1806.24</td>
<td>-1214.13</td>
<td>-1230.51</td>
<td>-1231.18</td>
</tr>
<tr>
<td></td>
<td>(3.95)</td>
<td>(3.94)</td>
<td>(3.93)</td>
<td>(3.86)</td>
<td>(3.84)</td>
<td>(3.84)</td>
</tr>
<tr>
<td>10</td>
<td>-1782.67</td>
<td>-1804.32</td>
<td>-1805.36</td>
<td>-1212.57</td>
<td>-1229.83</td>
<td>-1230.54</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.93)</td>
<td>(3.85)</td>
<td>(3.83)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>11</td>
<td>-1780.55</td>
<td>-1803.21</td>
<td>-1804.25</td>
<td>-1211.41</td>
<td>-1229.31</td>
<td>-1230.02</td>
</tr>
<tr>
<td></td>
<td>(3.91)</td>
<td>(3.91)</td>
<td>(3.91)</td>
<td>(3.85)</td>
<td>(3.83)</td>
<td>(3.82)</td>
</tr>
<tr>
<td>12</td>
<td>-1778.80</td>
<td>-1802.41</td>
<td>-1803.52</td>
<td>-1209.96</td>
<td>-1228.69</td>
<td>-1229.46</td>
</tr>
<tr>
<td></td>
<td>(3.94)</td>
<td>(3.93)</td>
<td>(3.92)</td>
<td>(3.85)</td>
<td>(3.83)</td>
<td>(3.82)</td>
</tr>
<tr>
<td>13</td>
<td>-1777.23</td>
<td>-1801.63</td>
<td>-1802.69</td>
<td>-1208.30</td>
<td>-1227.78</td>
<td>-1228.66</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(3.92)</td>
<td>(3.92)</td>
<td>(3.80)</td>
<td>(3.80)</td>
<td>(3.81)</td>
</tr>
</tbody>
</table>

*The smallest value of each criterion among all candidate models is in bold font.*
is on the AIDS Clinical Trials Group (ACTG) 398 longitudinal viral load data, where the data have left-censored values.

4.4.1 The Schizophrenia Data with Missing Values

The Schizophrenia data with missing values was first introduced in Gibbons et al. (1988). Hedeker and Gibbons (1994) and Hedeker and Gibbons (1997) explored different statistical models afterwards. In the study, 437 patients were randomly given placebo or drugs (chlorpromazine, fluphenazine, or thioridazine) and were followed for 6 weeks. Question raised from the study is whether the drugs relieve the symptoms of schizophrenia or not. Gibbons et al. (1988) picked Item 79 of the Inpatient Multidimensional Psychiatric Scale (IMPS) as the response of interests to examine the curative effect, which ranges between 1 to 7 with 1 indicating normal and a larger number indicating more severely ill.\(^1\) Data analyzed in this paper were accessed from Donald Hedeker’s website http://tigger.uic.edu/~hedeker/long.html. Figure 4.1 shows the severity score curves of the first 8 patients from the placebo group and the drug group respectively. In Table 4.7, sample size available at every week is listed. Ideally, we would get one record of the severity score weekly for every patient in the study. But clearly lots of observations are missing, especially at weeks 2, 3, and 5.

Previously, the Gibbons et al. (1988) and Hedeker and Gibbons (1994) did not take into account the missingness, but built the model as if the data is complete. Gibbons et al. (1988) used a random-effects regression model assuming the severity score as a continuous variable. Hedeker and Gibbons (1994) dichotomized the response and built a random-effects binary probit regression model. They built another random-effects ordinal

\(^1\)IMPS: 1 = normal, not at all ill; 2 = borderline mentally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; and 7 = among the most extremely ill.
Figure 4.1: Severity scores across weeks for the first 8 patients in each group. Solid segments are used to link two measurements in the neighborhood, and dashed segments are used to link two measurements with missing values in the middle.

Table 4.7: Sample size across weeks of the schizophrenia study.

<table>
<thead>
<tr>
<th>Group</th>
<th>Week</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>placebo(n=108)</td>
<td>107</td>
<td>105</td>
<td>5</td>
<td>87</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>drug(n=329)</td>
<td>327</td>
<td>321</td>
<td>9</td>
<td>287</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>
probit regression model after grouping the response into four ordered categories. In both papers, the three drugs were grouped into one since they revealed that these three drugs had similar effects.

Hedeker and Gibbons (1997) analyzed the data with a random-effects pattern-mixture models by coding the missing pattern with a new variable “dropout”. The model is

\[ Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \beta_3 \text{Drug}_i + \beta_4 \text{Drug}_it_{ij} + \beta_5 \text{Dropout}_i \quad (4.11) \]

\[ + \beta_6 \text{Dropout}_it_{ij} + \beta_7 \text{Dropout}_i \text{Drug}_i + \beta_8 \text{Dropout}_i \text{Drug}_it_{ij} + \epsilon_{ij}, \]

\[ (b_{0i}, b_{1i})^T \overset{iid}{\sim} N(0, \Sigma), \]

\[ \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \]

where \( t = \sqrt{\text{week}} \), Drug = 1 if the patient received drugs and 0 otherwise, Dropout is an indicator if the patient had measurement at the 6th week or not. Compared to Gibbons et al. (1988), Hedeker and Gibbons (1997) improved the model by adding the Dropout variable to handle the missing data. However, it is hard to tell if this is good enough or not since there might be \( 2^7 \) different missing patterns. Also, though square root transformation was taken to form a new time variable, assuming linear relationship between the response and time for both the placebo group and drug group may not be appropriate.

We fit the LMM-BPS with constant error terms for the two groups respectively. The severity score is treated as a continuous variable. Similar to Hedeker and Gibbons (1997), \( t = \sqrt{\text{week}}/2.5 \) is used to convert week to a new variable ranged between 0 and 1. We computed modified BPDC at \( m = 2, 3, 4, 5 \), since in the complete data set every subject is supposed to have 7 observations. The parameter estimates and model selection criterion were computed based on 5000 MCMC samples, after discarding the first 15000 as burn-in
Table 4.8: BPDC of models for schizophrenia study

<table>
<thead>
<tr>
<th>Group</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>1057.6</td>
<td>1061.9</td>
<td>1064.4</td>
<td>1064.2</td>
</tr>
<tr>
<td>Drug</td>
<td>3637.3</td>
<td>3566.8</td>
<td>3570.0</td>
<td>3571.3</td>
</tr>
</tbody>
</table>

samples.

Table 4.8 records values of modified BPDC, which suggests using $\hat{m} = 2$ for placebo group and $\hat{m} = 3$ for drug group. With MCMC samples, we got $\hat{\beta}_{\text{placebo}} = (5.35, 4.50)^T$ and $\hat{\beta}_{\text{drug}} = (5.36, 4.28, 2.92)^T$, revealing the decreasing trends in both group during the study. Left panel of Figure 4.2 displays the estimated mean functions over time along with the group means by week. Despite Week 2, 3, and 5, where very limited observations are accessible, the means of observed data are very closed to the fitted mean function. We computed $R^2$ based on observed data only to be 0.828, which also suggests good fit of the model. For the placebo group, the severity score drops slightly and slowly. For the drug group, there is an remarkable decrease, suggesting that the drug reduce the severity of the schizophrenia to some extent. Right panel of Figure 4.2 shows the difference of severity scores between placebo group and drug group, in which the 95% pointwise credible interval is below 0 after about the second day. It demonstrates that drugs do relieve the symptoms of schizophrenia. Estimated covariance functions are shown in Figure 4.3. We also plotted the first derivative of the mean severity score curves in Figure 4.4, which clearly supports that those drugs help ease the severity and the curative effects are getting somewhat better as time goes on.
Figure 4.2: Left: estimated mean function over weeks, and original group means by week displayed to show the fit of the model. Right: estimated difference function between placebo group and drug group along with its 95% credible interval and the reference line in grey at y=0.

Figure 4.3: Estimated covariance functions over weeks. Left: placebo group. Right: drug group
Figure 4.4: Left: estimated first derivative of mean function over weeks. Right: estimated difference of the first derivative function between placebo group and drug group along with its 95% credible interval.

4.4.2 The ACTG 398 Study with Censored Values

The ACTG 398 study was started in October 1998, where 481 human immunodeficiency virus (HIV)-infected persons with viral load above 1000 copies/ml were enrolled in the study. All patients received at least one protease inhibitor (PI), and were randomly assigned to get a second PI (324 patients) or placebo (157 patients). In addition, some patients also received nonnucleoside reverse transcriptase inhibitors (NNRTIs) previously, which was believed to affect the PIs’ treatment. So an interaction term should be included. One of the objectives is to find if the dual protease inhibitor (DPI) regimen is superior to the existing single protease inhibitor (SPI) regimen. The HIV viral load was recorded in scheduled visits at weeks 0, 2, 4, 8, 16, and 24. However, due to device limits, the HIV viral load lower than 200 copies/ml is not accurate, which produces about 22% of left-censored measurements. Figure 4.5 shows the HIV viral load of SPI and
DPI groups over the 24 weeks’ study period. The data for analysis were accessed from Dr. Hulin Wu’s website http://www.urmc.rochester.edu/biostat/people/faculty/wusite/datasets/ACTG398.cfm.

Figure 4.5: Log10 of HIV viral load (copies/ml) across weeks. The SPI group is plotted in solid black line, while DPI group is plotted in dashed red line. Horizontal reference line in blue is plotted at log_{10}(200) copies/ml where measurements below the line is inaccurate. The light blue shaded area highlights those measurements which are not reliable.

Previously, main findings concerning the study were reported in Hammer et al. (2002), concluding that DPI and NNRTI-naive were associated with better outcomes. The same data were analyzed by Sun and Wu (2005) and Huang (2010). Sun and Wu (2005) ignored the censored data and fit a semiparametric time-varying coefficients regression model such that

\[ Y(t) = \beta_0(t) + \beta_1(t)X + (\theta_0 + \theta_1t)Z + \epsilon(t), \quad (4.12) \]
where $X$ is the indicator of NNRTI treatment and $Z$ is the indicator of DPI treatment. Though the model is reasonable, ignoring the censoring aspect may lead to biased estimates. With Model (4.12), $E(Y(t)|X = 1, Z = 0) - E(Y(t)|X = 0, Z = 0) = \beta_0(t) + \beta_1(t) - \beta_0(t) = \beta_1(t)$ and $E(Y(t)|X = 1, Z = 1) - E(Y(t)|X = 0, Z = 1) = \beta_0(t) + \beta_1(t) + \theta_0 + \theta_1t - \beta_0(t) - \theta_0 - \theta_1t = \beta_1(t)$. Thus, the model does not allow the treatment effect to change with $Z$.

Huang (2010) also proposed a semiparametric mixed model, but took care of the censored data at the same time. Their model is

$$Y_{ij} = f(t_{ij}) + \beta_1 X_{it} I(t_{ij} > 0) + \beta_2 Z_i + U_{ij} b_i + \epsilon_{ij}, \quad (4.13)$$

where they added the random effects term $U_{ij} b_i$ to model the variability within subject. For the design matrix $U_{ij}$, Huang (2010) started with three choices, and picked $U_{ij} = (1, t_{ij})$ based on BIC. However, such simple structured design matrix may not be adequate enough to capture the complex covariance function. Also, BIC may not perform well for model selection among mixed models. Other than that, Huang (2010) does not include the interaction effect between DPI treatment and NNRTI treatment either.

We fit the LMM-BPS with constant error terms for the SPI and DPI group respectively, assuming the NNRTI treatment effect is constant all through the 24 weeks but can be different in different groups.

$$Y_{ij} = \sum_{k=1}^{m(a)} b_{km(a)}(t_{ij}) \beta_{ik}^{(a)} + \beta_2^{(a)} Z_i + \sigma^{(a)} \epsilon_{ij}, \quad (4.14)$$

$\beta_i^{(a)} \overset{\text{iid}}{\sim} N(\beta^{(a)}, D^{(a)}),$

$\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, 1),$

102
Table 4.9: BPDC of models for ACTG 398 study

<table>
<thead>
<tr>
<th>Group</th>
<th>m=2</th>
<th>m=3</th>
<th>m=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPI</td>
<td>2433.2</td>
<td>2164.8</td>
<td><strong>2101.8</strong></td>
</tr>
<tr>
<td>DPI</td>
<td>4573.5</td>
<td>4196.8</td>
<td><strong>4096.2</strong></td>
</tr>
</tbody>
</table>

where \( \beta_i^{(a)} = (\beta_{i1}^{(a)}, \ldots, \beta_{im(a)})^T \), \( a = 0, 1 \) indicate SPI and DPI respectively, and \( Z_i \) is the indicator of receiving NNRTI treatment. Equivalently, Model (4.14) can be written as

\[
Y_{ij} = X_i \sum_{k=1}^{m_1} b_{km_1}(t_{ij})\beta_{ik}^{(1)} + (1 - X_i) \sum_{k=1}^{m_0} b_{km_0}(t_{ij})\beta_{ik}^{(0)} + \beta_z Z_i + \beta_{xz} X_i Z_i + \sigma(a) \epsilon_{ij},
\]

\[
\beta_i^{(a)} \overset{iid}{\sim} N(\beta(a), D(a)),
\]

\[
\epsilon_{ij} \overset{iid}{\sim} N(0, 1),
\]

Clearly in this way, we have taken into account the interaction effect of PI treatment and NNRTI treatments.

Instead of using “week” directly, we transformed it with \( t = \sqrt{\text{week}}/5 \), and built our model in terms of “t”. Response variable is log_{10}(HIV viral load). We then computed the modified BPDC at \( m = 2, 3, 4 \), since each subject has 6 observations. The parameter estimates and model selection criteria were computed based on 5000 MCMC samples, with the first 15000 samples dropped as burn-in samples.

Table 4.9 shows the model selection criteria, where smaller value is favored. It suggested to pick \( \hat{m} = 4 \) for both groups. With MCMC samples, we got \( \hat{\beta}_{\text{nntri,spi}} = 0.20 \) and \( \hat{\beta}_{\text{nntri,dpi}} = 0.41 \). This implies that NNTRI that NNTRI has interaction effect with PIs. We also computed \( R^2 \) based on accurate observations to be 0.786, which shows that
our model fit the data well. We also fit the same model pretending all measurements are
accurate, ending up with $\hat{m} = 4$ for both groups as well. The $R^2$ in that case is 0.810,
which is a little higher than 0.786. This implies that those measurements below the de-
tection limit may not be accurate but close to the true measurement, and therefore carry
useful information for fitting the model. This, on the other hand, shows that our model
is capable of handling the censored data very well. Without using those “trustable” mea-
surements directly, we still get high $R^2$. In practice, those inaccurate measurements may
be noninformative or even misleading. However, the use of our suggested model is always
a good choice.

The estimated mean functions are shown in Figure 4.6. It is obvious that the HIV
viral load drop sharply during the first 3 weeks. After that, there is a small rebound,
but still remains at a comparably low level. The curve that represents the DPI group is
below the curve that represents the SPI group for most of the time. Figure 4.7 shows the
difference between mean functions of SPI group and DPI group along with 95% credible
interval. On the left panel, it is the case where patients do not experience NNTRI before
the study. The estimated difference function is always above 0. Especially after Week 7,
the 95% credible interval is above 0, which indicates significant improvement in reducing
the HIV viral load with a second PI taken by those NNTRI-native patients. On the right
panel, it shows the case where patients had NNTRI before the study. For those patients,
the 95% credible interval contains 0 all the time through the study period (24 weeks).
It is hard to tell if DPI is better than SPI at this time or not. Figure 4.8 displays the
estimated covariance functions for the SPI group and DPI group. Compared to the DPI
group, the covariance function of the SPI group is shorter, which supports our model
assumption that the underlying Gaussian processes related to time effects have different
mean functions and covaraince functions.
Figure 4.6: Estimated mean functions over weeks. Left: NNTRI-naive group; Right: NNTRI-experienced group.

Figure 4.7: Estimated differences between mean functions of SPI and DPI groups over days. Left: NNTRI-naive group; Right: NNTRI-experienced group.
Figure 4.8: Estimated covariance functions. Left: SPI group; Right: DPI group.

4.5 Discussion

We extended our LMM-BPS to the cases where data have missing values or censored values, which can be estimated with WinBUGS easily. Also, modified BPDC as well as modified DIC and LPML for missing data and censored data models are introduced. A simulation study is illustrated to compare the performance of modified BPDC to LPML and DIC. The results show that modified BPDC outperforms the other two in all cases. Two real data sets are analyzed employing LMM-BPS and modified BPDC. Analysis of the schizophrenia data with missing values suggested that the drug has significant effects in mitigating the symptoms. Analysis of ACTG 398 study with left-censored values reveals the DPI treatment effect and its interaction with NNTRI treatment.
Chapter 5

Discussions and Future Work

In this dissertation, we proposed nonparametric models using Bernstein polynomial sieve, so-called LMM-BPS, to approximate nonlinear Gaussian process. Also natural extension of LMM-BPS to longitudinal data that are subject to data irregularities like missing and censoring is introduced. To select the tuning parameter, we proposed a new Bayesian model selection criterion based on predictive divergence, with extensions to missing data and censored data models. Simulation studies show the new criteria work better than the traditional ones, LPML and DIC. Convergence properties of LMM-BPS are established and accompanied with a simulation study. Applications to complete data, data containing missing values, and left-censored values are illustrated with interesting findings. In addition, attractive facts related to the proposed nonparametric model include that the derivative process can also be analyzed in the frame work and that the shapes of the response curve and the derivative of the response curve are maintained with the proposed model.

For the future work, incorporating time-varying predictors into the model is an interesting topic. In the dissertation, only baseline predictors are considered. With time-
varying predictor, we are introducing another stochastic process into the model, which adds difficulties to both establishing the model and making statistical inference. One suggested solution to that problem is described as follows. Suppose that $Z_i(t)$ denotes a time-varying covariate process and that $Z_i(t) \in [0, 1]$ for any $i$ and $t \in [0, T]$. Then we may extend our LMM-BPS to

$$\begin{align*}
Y_i(t) &= \sum_{k=1}^{m_1} b_{km_1}(Z_i(t)) \beta_{ik} + \epsilon_i(t), \\
Z_i(t) &= \sum_{g=1}^{m_2} b_{gm_2}(t) \alpha_{ig} + \epsilon_i(t),
\end{align*}$$

(5.1)

where $\beta_i = (\beta_{i1}, \ldots, \beta_{im_1})^T \sim \text{iid } N(\beta_0, D_\beta)$ and $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{im_2})^T \sim \text{iid } N(\alpha_0, D_\alpha)$. Also, incorporating our nonparametric models to jointly model the longitudinal data and survival data is worth a try.
REFERENCES


Christensen, R. (2010). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. CRC.


