

RICE UNIVERSITY

**Functional Data Classification and Covariance Estimation**

by

**Hongxiao Zhu**

A THESIS SUBMITTED  
IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE

**Doctor of Philosophy**

APPROVED, THESIS COMMITTEE:

---

Dennis D. Cox, Chair  
Professor of Statistics, Rice University

---

Marina Vannucci  
Professor of Statistics, Rice University

---

Wotao Yin  
Assistant Professor  
Department of Computational and  
Applied Mathematics, Rice University

---

Jong Soo Lee  
Visiting Assistant Professor, Department  
of Statistics, Carnegie Mellon University

HOUSTON, TEXAS

DECEMBER, 2008

## **Abstract**

# **Functional Data Classification and Covariance Estimation**

by

Hongxiao Zhu

Focusing on the analysis of functional data, the first part of this dissertation proposes three statistical models for functional data classification and applies them to a real problem of cervical pre-cancer diagnosis; the second part of the dissertation discusses covariance estimation of functional data.

The functional data classification problem is motivated by the analysis of fluorescence spectroscopy, a type of clinical data used to quantitatively detect early-stage cervical cancer. Three statistical models are proposed for different purposes of the data analysis. The first one is a *Bayesian probit model with variable selection*, which extracts features from the fluorescence spectroscopy and selects a subset from these features for more accurate classification. The second model, designed for the practical purpose of building a more cost-effective device, is a *functional generalized linear model with selection of functional predictors*. This model selects a subset from

the multiple functional predictors through a logistic regression with a grouped Lasso penalty. The first two models are appropriate for functional data that are not contaminated by random effects. However, in our real data, random effects caused by devices artifacts are too significant to be ignored. We therefore introduce the third model, the *Bayesian hierarchical model with functional predictor selection*, which extends the first two models for this more complex data. Besides retaining high classification accuracy, this model is able to select effective functional predictors while adjusting for the random effects.

The second problem focused on by this dissertation is the covariance estimation of functional data. We discuss the properties of the covariance operator associated with Gaussian measure defined on a separable Hilbert Space and propose a suitable prior for Bayesian estimation. The limit of Inverse Wishart distribution as the dimension approaches infinity is also discussed. This research provides a new perspective for covariance estimation in functional data analysis.

## Acknowledgements

I would like to first express my sincere gratitude to my advisor, Dr. Dennis D. Cox, for his invaluable support, encouragement and guidance throughout this research work. As a knowledgeable and insightful statistician, he has provided most useful suggestions that help to make crucial breakthroughs during this research trip, and moreover, he has enlightened me how to do good academic research, which, I believe, will benefit my whole career.

I am highly grateful to Dr. Marina Vannucci for her help and suggestions for improving the quality of this work. I am also indebted to Dr. Jong Soo Lee and Dr. Wotao Yin, who have served in my committee and have contributed many useful comments for my dissertation.

I would like to thank many other prestigious scholars who have taught me statistics and mathematics during my graduate study at Rice: Dr. Mckay Hyde, Dr. Valen Johnson, Dr. Peter Müller, Dr. Rudolf Riedi, Dr. Raúl Rojas, Dr. Javier Rojo, Dr. David Scott, etc. Special thanks goes to Dr. Michelle Follen. She has provided me the opportunity to work on real biomedical problems and collaborate with researchers from many other disciplines.

Finally, I wish to thank my family for their love and support throughout my life. I am especially thankful to my husband, Xiaowei Wu, who has helped to proofread and revise this work.

This dissertation was partially supported by National Cancer Institute grant PO1-CA82710 and by the National Science Foundation grant DMS0505584.

# Contents

<b>Abstract</b>	ii
<b>Acknowledgements</b>	iv
<b>List of Figures</b>	xi
<b>List of Tables</b>	xiv
<b>1 Introduction and Literature Review</b>	1
1.1 Introduction . . . . .	1
1.2 A Functional Data Example. . . . .	4
1.3 Literature Review . . . . .	8
1.3.1 Functional Principal Component Analysis . . . . .	9
1.3.2 Functional Data Regression . . . . .	12
1.3.3 Functional Data Covariance Estimation . . . . .	15
1.4 Outline of the Dissertation . . . . .	16
<b>2 Background</b>	17

2.1	Convergence of Markov Chain Monte Carlo . . . . .	17
2.1.1	General Definitions and Results . . . . .	18
2.1.2	Gibbs Sampling . . . . .	22
2.1.3	Metropolis Sampling . . . . .	25
2.2	Bayesian Variable Selection . . . . .	27
<b>3</b>	<b>A Bayesian Probit Model with Variable Selection for Functional Data Classification</b>	<b>34</b>
3.1	Introduction . . . . .	34
3.2	The Proposed Model . . . . .	35
3.3	Priors . . . . .	37
3.4	Posteriors . . . . .	38
3.5	Parameter Settings . . . . .	41
3.6	Markov Chain Monte Carlo . . . . .	42
3.7	Simulation Study . . . . .	43
3.8	Fluorescence Spectroscopy Data Classification . . . . .	50
3.9	Conclusion . . . . .	54
<b>4</b>	<b>A Functional Generalized Linear Model with Functional Predictor Selection</b>	<b>57</b>
4.1	Introduction . . . . .	57
4.2	The Proposed Model . . . . .	58

4.3	Simulation Study . . . . .	62
4.4	Real Data Application . . . . .	68
4.5	Discussion . . . . .	72
<b>5</b>	<b>A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors</b>	<b>75</b>
5.1	Motivation . . . . .	76
5.2	Bayesian Hierarchical Model with Selection of Functional Predictors .	78
5.2.1	The Proposed Model . . . . .	78
5.2.2	The Posterior Inference . . . . .	81
5.3	Markov Chain Monte Carlo . . . . .	84
5.3.1	Algorithm 1 . . . . .	84
5.3.2	Algorithm 2 (EMC) . . . . .	87
5.4	Setting Parameters . . . . .	90
5.5	Simulation Results . . . . .	91
5.5.1	Simulation 1 . . . . .	92
5.5.2	Simulation 2 . . . . .	96
5.6	Fluorescence Spectroscopy Data Application . . . . .	98
5.7	Discussion . . . . .	105
<b>6</b>	<b>Priors for Covariance Operators in Functional Data Analysis</b>	<b>108</b>
6.1	Grid Refinement Invariance Principle . . . . .	108

6.2 Gaussian Measures . . . . .	109
6.2.1 Gaussian Measures Defined on Finite-dimensional Hilbert Space	110
6.2.2 Gaussian Measures Defined on Infinite-dimensional Hilbert Space	112
6.3 A Possible Prior for Covariance Operators . . . . .	116
6.4 A Markov Chain Monte Carlo . . . . .	126
6.4.1 Derivations of the Posterior Distribution . . . . .	127
6.4.2 Notes on Some Computational Tricks . . . . .	132
6.4.3 Simulation Results . . . . .	134
6.5 Inverse-Wishart Prior and its limiting Behavior . . . . .	142
6.5.1 Definition and Some Facts about Wishart and Inverse Wishart distribution . . . . .	143
6.5.2 Conjugate Inverse Wishart Priors for the Covariance in Multivariate Normal Model . . . . .	144
6.5.3 A Simulation Study using the Bayesian Model with Conjugate Inverse-Wishart Prior . . . . .	145
6.5.4 Limiting Behavior of the First Two Moments of the Inverse-Wishart Distribution . . . . .	148
<b>7 Conclusion and Discussion</b>	<b>152</b>
<b>A Integrating <math>b_l</math>'s, <math>b_0</math> and <math>\alpha</math> Out Sequentially from the Conditional Posterior (5.10).</b>	<b>156</b>

<b>B Proof of Proposition 4.2.1</b>	<b>159</b>
<b>C Verification for Convergence of the MCMC Algorithm 1 in Chapter 5.</b>	<b>162</b>
C.1 The Verification of Algorithm 1 . . . . .	162
C.2 Reversible Condition of Metropolis-Hastings . . . . .	167
<b>D Some Details on EMC Algorithms</b>	<b>170</b>
<b>Bibliography</b>	<b>175</b>

# List of Figures

1.1	A multivariate data example compared with a functional data example	3
1.2	Fluorescence spectroscopy for <i>in vivo</i> detection of cervical pre-cancer	5
1.3	Fluorescence spectroscopy EEM plot	6
1.4	Normal EEM vs. disease EEM	8
2.1	The plot of normal densities with relatively large (1) and small (0.1) variances.	29
3.1	BVS model simulation 1: posterior estimation of $\beta(t)$ and the corresponding simultaneous 95% credibility band.	46
3.2	BVS model simulation 2: marginal posterior estimate of $\tau$ .	48
3.3	BVS model simulation 2: prediction performance comparison using empirical ROC curves.	49
3.4	BVS model real data application: marginal posterior probability of $\tau$ .	53
3.5	BVS model real data application: prediction performance comparison using empirical ROC curves.	55

3.6 BVS model convergence diagnostic: marginal posterior of $\tau$ from different chains. . . . .	56
4.1 FGLM: the plot of simulated data. . . . .	63
4.2 FGLM simulation: estimated paths of coefficients at different $\lambda$ values. . . . .	64
4.3 FGLM simulation: estimated coefficient $\hat{\beta}_2(t)$ at 6 selected $\lambda$ values. . . . .	65
4.4 FGLM simulation: prediction results at different $\lambda$ values. . . . .	68
4.5 FGLM real data application: the selected functional predictors at different $\lambda$ values. . . . .	70
4.6 FGLM real data application: prediction results at different $\lambda$ values. . . . .	71
4.7 FGLM real data application: ROC curves obtained when training 4 different classifiers. . . . .	73
5.1 BHFPS: box-plot of the first functional principle component scores of one spectral curve. . . . .	78
5.2 BHFPS simulation 1: the autocorrelation plot for posterior samples of $\sigma_b^2$ . . . . .	94
5.3 BHFPS simulation 1: the posterior estimation of the non-zero coefficient functions. . . . .	95
5.4 BHFPS simulation 2: the marginal posterior probabilities $P\{\tau_j = 1, j = 1, \dots, J\}$ . . . . .	98
5.5 BHFPS real data application: the marginal posterior probabilities $P\{\tau_j = 1, j = 1, \dots, 16\}$ . . . . .	101

5.6 BHFPS real data application: the top 10 most frequently visited models.	102
5.7 BHFPS real data application : ROC curves obtained by test set prediction.	104
6.1 Plot of Brownian Motion sample paths	135
6.2 True covariance function of Brownian Motion	135
6.3 The plot of prior parameter $\{w_j\}_j$	136
6.4 Plot of the prior covariance function for $\vec{Z}_j$ 's.	137
6.5 Plot of the averaged posterior covariance.	138
6.6 Trace plot of the posterior samples of $c$ .	139
6.7 Posterior average of the mean function $\mu(t)$ .	140
6.8 Plot of the component-wise estimation error for the Bayes estimate.	141
6.9 Plot of the component-wise estimation error for the sample estimate.	142
6.10 The posterior average of covariance using Inverse-Wishart prior.	146
6.11 One Brownian Motion path sampled at three grid levels.	147
7.1 A comparison of the functional predictor selection results of FGLM and BHFPS.	154
D.1 EMC algorithm: the temperature effect.	171

# List of Tables

1.1 Diagnosis description . . . . .	7
3.1 BVS model simulation 1: estimation of coefficient $\beta$ compared with MLE. . . . .	45
3.2 BVS model simulation 2: prediction results compared with 3 other classification methods. . . . .	50
3.3 BVS model real data application: prediction results compared with 3 other classification methods. . . . .	54
4.1 FGLM simulation: the estimated coefficient values compared with the true values. . . . .	67
4.2 FGLM real data application: the classification results using 4 different methods. . . . .	72
5.1 BHFPS real data application: acceptance rates of parameters in EMC algorithm. . . . .	100
5.2 BHFPS real data application : prediction on test set results . . . . .	103

6.1	Compare the estimation errors of the Bayes estimate with the sample estimate. . . . .	141
6.2	The estimation error of the Bayes model with Inverse-Wishart prior .	146
6.3	The estimation error comparison (IW prior) under different grid levels.	147

# Chapter 1

## Introduction and Literature

### Review

#### 1.1 Introduction

Statistical theories generally fall into two categories: univariate and multivariate, according to the dimensionality of the underlying random variables. For univariate theory, the object of interest is a one-dimensional random variable (denoted by  $X$ ) which maps the sample space  $\Omega$  to the real line  $\mathbb{R}$ , i.e.,

$$X : (\Omega, \mathcal{B}(\Omega)) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

Here, for a given set  $A$ ,  $\mathcal{B}(A)$  represents the  $\sigma$ -field generated by subsets of  $A$ . The pair  $(A, \mathcal{B}(A))$  is a measurable space, and the map  $X$  is measurable by the definition of random variable. If the random element of interest is more than one dimensional,

we use a random vector (denoted by  $\vec{X}$ ) instead and the measurable map becomes

$$\vec{X} : (\Omega, \mathcal{B}(\Omega)) \mapsto (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)),$$

where  $\mathbb{R}^m$  is a  $m$ -dimensional Euclidean space. Statistical analysis for finite dimensional random vectors (or random matrices) is called multivariate data analysis (see, for example, Muirhead [50]). When  $m$  approaches infinity, the random vector becomes a random sequence. A more general extension is to treat  $m$  as an index variable taking values from some index sets  $T$  (which can be uncountable). Then the measurable map can be treated as a random function with argument in  $T$ . Under this setting, we call the observed data, usually in forms of curves and images, “functional data”. The statistical methods for analyzing functional data are named “functional data analysis” (FDA), coined by Ramsay and Dalzell [59]. In many cases, the index set  $T$  is a dense set such as a temporal or spatial domain, therefore ideally functional data can have as high resolution as possible. In this dissertation, we let  $X(t)$  be the random function indexed by  $t, t \in T$  and  $x(t)$  be its data realization. Alternatively, Ferraty and Vieu [19] call  $X(t)$  a functional variable, defined as follows:

**Definition 1.1.1.** *A random variable is called functional variable if it takes values in an infinite dimensional space (or functional space). An observation of the functional variable is called a functional data.*(Ferraty and Vieu [19])

Many real data, such as most images and signals, can be treated as functional data. Figure 1.1 shows an example of multivariate data and functional data. Another practical example in medical research is shown in Section 1.2.

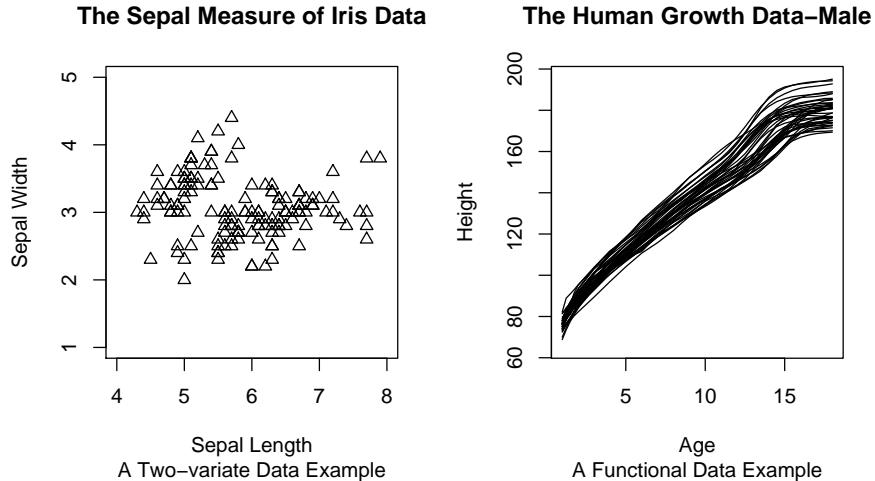


Figure 1.1: The left panel is the data plot for the sepal length and width (in centimeters) for 150 iris flowers, which is an example of multivariate data. There are two measurements, length and width. The right panel is the plot of 39 boys' heights measured through age 1 to 18, which is an example of functional data.

The research in FDA started in the 1980s. As time goes on, FDA becomes one of the most important new statistical methodologies with diverse applications in many areas. As a relatively new field, FDA borrows many ideas from non-parametric statistics and multivariate data analysis, and adopts techniques from signal/image processing, longitudinal data analysis and data mining. Generally speaking, we can categorize current statistical methods in FDA literature as follows:

1. **Smoothing and Registration.** As preprocessing steps, smoothing and registration techniques help filter out noise (or observation errors) of the original data and align them appropriately on their domain. Nonparametric regression methods, such as smoothing spline and penalized methods, are usually used for smoothing functional data. Registration is usually done by setting up a

registration criterion, or using landmarks or warping functions.

2. **Functional Principal Component Analysis (FPCA).** As an important dimension reduction technique in multivariate analysis, Principal Component Analysis (PCA) finds the dominate modes of variation in the data. By changing summations to integrations, this technique can also be extended to the functional case.
3. **Regression.** Many works concerning regression problems in functional data have been done, from both frequentist and Bayesian perspectives. It turns out that most classical regression models in multivariate analysis, such as multivariate ANOVA, mixture effects model, generalized linear regression, have their analogous version in FDA.
4. **Hypothesis Testing.** The topic of hypothesis testing in functional data is not as well developed as other FDA methods. The main difficulty lies in the assumption of infinite-dimensionality of the functional space. Recently, some new methods are proposed on testing whether one group of functional data has zero mean, or whether two groups have the same mean function.

## 1.2 A Functional Data Example.

The work in this dissertation is motivated by a series of fluorescence spectroscopy data in cancer research. As a special type of functional data, spectroscopy data

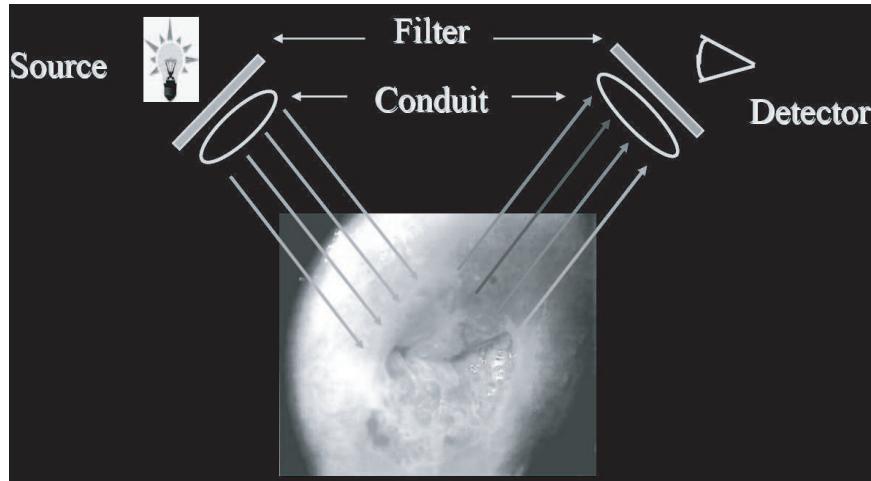


Figure 1.2: Using fluorescence spectroscopy to detect cervical pre-cancer *in vivo*. This picture is obtained from <http://www.eng.uct.ac.za/biaolab/Education/tutorials> [65].

contain the spectra of particular lights emitted (or absorbed) by a given material. This section gives a brief introduction to the fluorescence spectroscopy data used in cervical pre-cancer diagnosis.

Cervical cancer is known to be one of the leading causes of cancer deaths in women. Early-stage diagnosis using automatic, low cost screening devices plays an important role in the prevention of cervical cancer. Among the existing diagnosis tools, fluorescence spectroscopy is a promising technology to quantitatively detect cervical pre-cancer in a non-invasive way [57]. Figure 1.2 illustrates the mechanism of measuring fluorescence spectroscopy *in vivo*. This technology works as follows: First, an excitation light at a fixed wavelength illuminates the cervical tissue. During illumination, the endogenous fluorescent molecules in tissue absorb the excitation light and emit fluorescent light. The emitted light is then captured by an optical detector

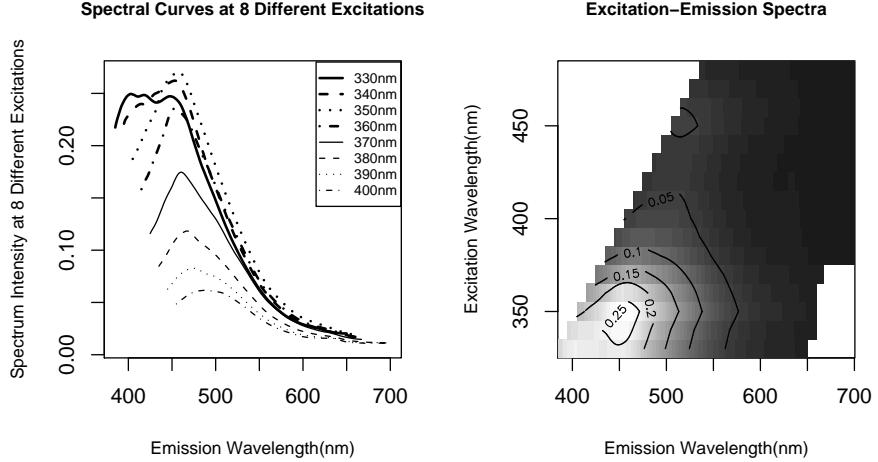


Figure 1.3: Left panel: spectral curves at 8 different excitation wavelengths ranging from 330nm to 400nm. Right panel: heat plot of an excitation-emission matrix (EEM).

which produces the corresponding spectrum as a smooth curve. By adjusting the wavelength of the excitation light, the detector records multiple spectral curves. In each measurement, the excitation light is varied at 16 different excitation wavelengths, ranging from 330 nm to 480 nm with increments of 10 nm. This produces 16 spectral curves for each measurement. In each curve, the fluorescence intensities are recorded at emission wavelengths ranging between 385 nm and 700 nm. Through data preprocessing, the curves are truncated so that some intensity points at the smallest and largest emission wavelengths are removed.

Figure 1.3 illustrates one observation. The left panel shows the first 8 of the total 16 spectral curves in this observation. The right panel shows a heat plot of the spectral intensities, by stacking up all the 16 spectral curves in the order of their excitation wavelength. We call such a set of fluorescence spectroscopy curves an

<i>Disease Level</i>	<i>Description</i>	<i>Diagnosis</i>
Cancer	Evidence of cancer	Diseased
CIS	Carcinoma in situ	
CIN III	Severe cervical intraepithelia neoplasia	
CIN II	Moderate cervical intraepithelia neoplasia	
CIN I	Mild cervical intraepithelia neoplasia	Normal
HPV	HPV associatied changes	
Atpia	Atpia	
Normal	No evidence of disease	

Table 1.1: The diagnosis levels and the discription

excitation-emission matrix (EEM).

The data considered in this dissertation contain 2414 measurements taken from 1006 patients. Each patient has 1 or more (up to 6) sites measured and there exists repeated measurements (although not for every patient). All measurements come from two devices (called Fast EEM2 and Fast EEM3), four probes and three clinics (MDACC, LBJ and BCCA). The colposcopic tissue type of the measurements can be either squamous or columnar. The menopausal status of the patients can be pre- peri- and post-menopausal. After pre-processing such as background correction and smoothing, the data were carefully split into training set and test set by balancing various factors. The proportion of diseased cases in the training and test sets are 10% and 9%, respectively.

The goal of our study is to discriminate normal from diseased measurements based on the EEM. Table 1.1 lists the detailed disease categories provided by pathologists in a progressive order. In our study, we consider all cases from CIN II or worse as diseased, and cases from CIN I or better as normal.

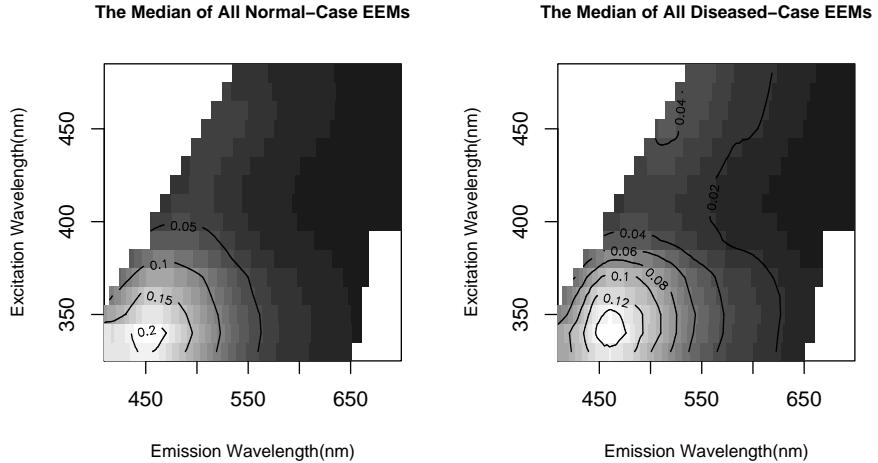


Figure 1.4: The heat plots for the median values of all normal-case EEMs versus the median values of all disease-case EEMs.

Figure 1.4 shows the heat plots of the median values of all normal-case EEMs versus those of all disease-case EEMs. Differences between the two plots are hard to be detected by naked eyes, although the normal-case EEM seems to have higher peak than the diseased-case EEM.

### 1.3 Literature Review

Much attention has been given to FDA since the 1980s. Early works include Ramsay [58], Ramsay and Dalzell [59] and Rice and Silverman [64]. More recently, Ramsay and Silverman ([62],[60]) did a systematic survey and addressed some applications issues [61]. As summarized in Section 1.1, there are mainly four areas of FDA that have received considerable attentions. Since this dissertation focuses on classification and covariance estimation, we will only review the literature related to such topics, which

include functional principal component analysis, regression and covariance estimation. Other topics, like smoothing and registration of functional data, are well presented in Chapter 3 – 5 and Chapter 7 of Ramsay and Silverman [62]; and one can find a detailed review of hypothesis testing of FDA in Chapter 4 of Lee [36].

### 1.3.1 Functional Principal Component Analysis

As one of the basic and widely used techniques proposed for FDA, Functional principal component analysis (FPCA) is a direct extension of multivariate principal component analysis(PCA). FPCA was first introduced by Ramsay ([58], [59]), Rice and Silverman [64], and was studied in detail by Ramsay and Silverman ([62],[60]). We briefly summarize these works in this section. Later chapters will use compatible notations.

In multivariate data analysis, principal components are computed by eigenvalue decomposition of the covariance matrix. Let  $X$  be a multivariate data matrix of size  $n \times p$ , its sample covariance  $V$  can be computed by  $V = \frac{1}{n}X^T X$ . The first eigenvector of  $X$  (denote  $\phi_1$ ) can be obtained by

$$\phi_1 = \underset{\|\phi\|=1}{\operatorname{argmax}} \phi^T V \phi,$$

which is equivalent to solving for the largest eigenvalue  $\lambda$  and the corresponding eigenvector  $\phi$  from

$$V\phi = \lambda\phi. \quad (1.1)$$

The first principal component scores can thus be obtained by  $X^T \phi_1$ . Solving Equation (1.1) subject to the condition  $\phi_2^T \phi_1 = 0$  gives the second eigenvector. Similarly,

one can find out all eigenvectors.

In the functional data case, one can define the covariance operator  $V$  by

$$V\phi(s) = \int_T v(s, t)\phi(t)dt,$$

where  $v(s, t) = 1/n \sum_i x_i(s)x_i(t)$  is the sample covariance function and  $\phi(\cdot)$  is the eigenfunction. The largest eigenvalue  $\rho$  and the corresponding eigenfunction  $\phi(\cdot)$  can be solved from

$$V\phi = \rho\phi, \quad (1.2)$$

which is of the same form as Equation (1.1) except that  $V$  and  $\phi$  are defined differently. The first principal component score for  $x_i(t)$  can be computed from  $\langle x_i(t), \phi_1(t) \rangle$ . Similar to the multivariate case, the second and later eigenfunctions can be obtained by adding the orthogonal constraint to Equation (1.2). To solve Equation (1.2), one can either discretize the  $x_i(t)$ 's on a finite grid, or expand them on another set of orthonormal basis.

In order to obtain eigenfunctions with sufficient smoothness, Rice and Silverman introduces a smoothed PCA method by adding a roughness penalty [64]. In their paper, the first eigenfunction is obtained by

$$\phi_1 = \underset{\|\phi\|=1}{\operatorname{argmax}} \langle \phi, (V - \lambda D)\phi \rangle,$$

where  $D$  is a roughening operator taking form of  $F^T F$ , where  $F$  is a second-order differencing operator. The subsequent eigenfunctions are obtained by adding additional

orthogonal conditions. The estimation of smoothed eigenfunctions is obtained by finding the eigenfunctions of  $V - \lambda D$ , where  $\lambda$  is chosen by cross-validation. Later on, this method was improved in Silverman [68], where the first eigenfunction is solved by

$$\phi_1 = \operatorname{argmax} \frac{\langle \phi, V\phi \rangle}{\|\phi\|^2 + \lambda[\phi, \phi]},$$

and  $[\phi, \phi] = \int (\phi''(t))^2 dt$ .

Following Silverman's smoothed FPCA, more theoretical results of FPCA have been investigated. Ocaña, Aguilera and Valderrama [54] assume Hilbert valued random variables and established equivalences between FPCA with a proposed inner product in the data space and certain FPCA with a given well-suited inner product. They also extended Silverman's method to a more general framework based on Hilbert valued random variables. Cardot [12] proposed a non-parametric conditional FPCA method and provided some consistency properties. Hall and Vial [29] studied the extrema of empirical principal component functions and compared them with those of the true principal component functions. They found that the empirical principal component functions can hardly distinguish a “shoulder” in a curve from a small bump. So they suggest a bootstrap method to assess the strength of the extrema. More properties of FPCA were discussed by Hall and Hosseini-Nasab [27], where they studied properties of FPCA through stochastic expansions. Their work demonstrated the fact that the properties of eigenfunction estimations are affected by the spacing among eigenvalues. They also propose bootstrap methods to construct simultaneous

confidence regions for eigenvalues and eigenvectors.

The sparsity of functional data has also caught much attention. James, Hastie and Sugar[33] introduce a reduced rank mixed effect model to estimate the principal component functions when data are irregular and sparse. Hall, Müller and Wang [28] focus on the effect of the sampling plan to the estimation of principal component functions. They indicate that the sparsity of the functional data can affect the convergence rates for the estimated eigenfunctions, but not for the estimated eigenvalues. Yao and Lee [80] propose penalized spline models for sparse functional data or longitudinal data. They developed an iterative procedure to reduce the dependence between the measurements within each subject (the dependence between the discrete points measured on the same curve).

Besides these theoretical works, many others aim at applying FPCA to solve a broad range of functional data problems, such as Grambsch et al. [25], James [32], Chiou, Müller and Wang [15], Park [55].

### 1.3.2 Functional Data Regression

To extend multivariate regression to the functional case, the most straightforward way is by using the point-wise models, which is similar to the varying coefficient model or the contemporary model (see Hastie and Tibshirani [31] and Staniswalis and Lee [70]). Let  $Y_i(t)$  be the functional responses and  $x_i(t)$  be the covariates,  $i = 1, \dots, n$ .

Suppose that the point-wise model takes form

$$y_i(t) = \alpha(t) + x_i(t)\beta(t) + \epsilon_i(t).$$

Cardot, Ferraty and Sarda [13], James [32] and Malfait et al. [41] considered the case where the response values at time  $t$  are explained by the predictor curves  $x_i(s)$  through:

$$y_i(t) = \alpha(t) + \int_{T_i} x_i(s)\beta(s, t)ds + \epsilon_i(t),$$

where  $T_i = [0, t]$  or  $[t - \delta, t]$ .

In many cases, regression with functional predictors and scalar responses is of particular interest. James [32] extended the generalized linear model (GLM) using spline basis to include functional predictors. Müller and Stadtmüller [51] proposed a similar method based on truncated Karhunen-Loéve expansion and proved some asymptotic properties of the estimation. To summarize the basic structure, let us assume that the functional generalized linear model takes form

$$Y = g(\alpha + \int \beta(t)X(t)dt) + \epsilon,$$

where  $Y$  is a univariate response variable,  $X(t)$  is the functional predictor, and  $g(\cdot)$  is an appropriately defined link function. Cardot and Sarda [14] analyzed the link between a scalar response and a functional predictor in a regression setting by means of a functional GLM. Besse et al.[6] also discussed several estimation methods under functional GLM setting. Li and Hsing [38] investigated the convergence rate of the estimation of the regression weight function in a functional linear regression model.

Another interesting model is the functional analysis of variance (FANOVA), in which functional responses are assumed. The predictors are usually real or dummy variables. The FANOVA model can be written as

$$Y_{li}(t) = \mu(t) + \alpha_l(t) + \epsilon_{li}(t),$$

where  $Y_{li}$  is the  $i$ th observation in group  $l$ ,  $\mu(t)$  is the grand mean and  $\alpha_l(t)$  is the effect of group  $l$  such that  $\sum_l \alpha_l(t) = 0$  for all  $t$ . This model can be written in a more general form as

$$\mathbf{y}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t),$$

where  $\mathbf{Z}$  is a design matrix and  $\boldsymbol{\beta}(t)$  is a vector of regression functions. Here both  $\mathbf{y}(t)$  and  $\boldsymbol{\epsilon}(t)$  can be vector of functions. Detailed fitting procedures can be found in Ramsay and Silverman [60]. Cardot[11] proposed a nonparametric estimator of regression function when the predictor is real but the response is functional.

From Bayesian perspectives, Morris et al. applied discrete wavelet transform (DWT) to the modeling of hierarchical functional data [49]. Morris and Carroll [48] extended linear mixed model to functional mixed model, which is given by

$$\mathbf{Y}(t) = X\mathbf{B}(t) + Z\mathbf{U}(t) + \mathbf{E}(t),$$

where  $\mathbf{Y}(t)$  is a vector of  $N$  functional responses and  $\mathbf{B}(t)$  is a  $p$ -vector of fixed effect functions associated with the  $N \times p$  design matrix  $X$ .  $\mathbf{U}(t)$  is a  $m$ -vector of random-effect functions associated with the  $N \times m$  design matrix  $Z$ .  $\mathbf{E}(t)$  is a vector of error process. The above model is transformed to wavelet domain through DWT,

where Bayesian methods are used to estimate the regression parameters. A similar model was applied to the accelerometer data in Morris et al. [46], [47]. McKeague [42] used Bayesian nonparametric regression and time warping to solve the signature verification problem. Behseta et al. [5] discussed some methods to account for estimation variation using Bayesian hierarchical models. More recent works on Bayesian functional data regression can be found in [10], [71], etc.

### 1.3.3 Functional Data Covariance Estimation

The most popular way of estimating the covariance of functional data is through orthogonal expansions, that is, write the covariance function as a weighted linear combination of eigenvalues and eigenfunctions:

$$\gamma(s, t) = \sum_k \lambda_k \phi_k(s) \phi_k(t),$$

and the estimation methods are the same as in FPCA in Section 1.3.1. Smoothing steps are usually introduced when estimating the eigenfunctions, such as the penalized method in Rice and Silverman [64] and the scatter-plot smoothing in Yao et al. [79]. Alternatively, Lee [37] estimated the covariance matrix through sample estimates on a finite grid. They then smoothed the eigenvectors of the covariance matrix to obtain the eigenfunctions. A summary of these works can be found in the dissertation of Lee [36].

Yao [78] applied kernel method in Longitudinal data analysis to estimate the mean and covariance function of functional data, based on the Nadaraya-Waston

estimator or local linear estimator. He also derived the asymptotic distribution of such nonparametric estimator for functional data contaminated with measurement error.

## 1.4 Outline of the Dissertation

We introduce some background knowledge in Chapter 2. In Chapter 3, a Bayesian probit model with variable selection is proposed for functional data classification and applied to the fluorescence spectroscopy data. To select a subset of the multiple functional predictors for more cost-effective classification, we propose a functional generalized linear model with a grouped-lasso penalty in Chapter 4, from a frequentist point of view. Chapter 5 extends the Bayesian probit model in Chapter 3 to account for random effects and to select functional predictors. Chapter 6 discusses covariance estimation of functional data. Further conclusions and discussions are put in Chapter 7.

# Chapter 2

## Background

### 2.1 Convergence of Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) originated in statistical physics, marked by a paper of Metropolis et al. [44] in 1953. Since then, MCMC has become increasingly popular in Bayesian modeling. In this section, we review some theoretical background of MCMC, especially on the convergence of Gibbs and Metropolis algorithms. The review is based mainly on Tierney's work [73], and partly on Professor Dennis D. Cox's class notes for Stochastic Process (taught in Spring, 2008). We only consider Markov Chains with continuous state space.

### 2.1.1 General Definitions and Results

Let  $\pi$  be the posterior distribution of interest. Suppose  $\pi$  is supported on  $E \subset \mathbb{R}^k$  and is absolutely continuous with respect to a  $\sigma$ -finite measure  $\mu$ , i.e.,  $\pi(dx) = \pi(x)\mu(dx)$ .

The main purpose of MCMC algorithms is to generate dependent samples (Markov chain)  $X_n, n = 1, 2, \dots$  with equilibrium distribution  $\pi$ . In other words, we want  $X_n$  to converge in distribution to  $\pi$  as  $n$  increases.

Assume that a time-homogeneous Markov chain with invariant distribution  $\pi$  has transition kernel defined by

$$P(X_n, A) = Pr\{X_{n+1} \in A | X_0, \dots, X_n\} = Pr\{X_{n+1} \in A | X_n\} = Pr\{X_1 \in A | X_0\}$$

for all measurable sets  $A \in \mathcal{E}$ , where  $\mathcal{E}$  is the  $\sigma$ -field generated by  $E$ .  $\pi$  is called an **invariant distribution** with respect to  $P(\cdot, A)$  if  $\pi(A) = \int P(x, A)\pi(dx)$ . The conditional distribution of  $X_n$  given  $X_0$  is written as

$$P^n(X_0, A) = Pr\{X_n \in A | X_0\},$$

where  $P^n$  denotes the  $n$ th iterate of the kernel  $P$ . A formal definition of the transition kernel is stated in Definition 2.1.1.

**Definition 2.1.1. (Transition Kernel)** *Let  $\mathcal{E}$  be a countably generated  $\sigma$ -algebra on  $E$ . A (Markov) transition kernel on  $(E, \mathcal{E})$  is a map  $P : E \times \mathcal{E} \rightarrow [0, 1]$  such that:*

(1)  $\forall A \in \mathcal{E}$ , the function  $P(\cdot, A)$  is measurable;

(2)  $\forall x \in E$ , the function  $P(x, \cdot)$  is a probability measure on  $(E, \mathcal{E})$ .

For a probability measure  $\nu$ , a transition kernel  $P$  on  $(E, \mathcal{E})$  and a real-valued  $\mathcal{E}$ -measurable function  $h$ , define  $\nu P$ ,  $Ph$  and  $\nu h$  by

$$(\nu P)(A) = \int P(x, A)\nu(dx), \quad (Ph)(x) = \int h(y)P(x, dy), \quad \nu h = \int h(y)\nu(dy),$$

$\forall x \in E$  and  $A \in \mathcal{E}$ . In other words,  $P(\cdot, \cdot)$  is an operator that plays two roles. For a probability measure  $\nu$  on  $(E, \mathcal{E})$ ,  $\nu P$  is a probability measure.  $\nu P$  can be thought of as the distribution of  $X_{n+1}$  when  $X_n \sim \nu$ . For a bounded function  $h : E \rightarrow R$ ,  $Ph$  can be thought of as a conditional expectation:  $(Ph)(x) = E[h(X_{n+1})|X_n = x]$ . A non-negative real-valued function  $h$  is called harmonic for  $P$  if  $h = Ph$ .

**Definition 2.1.2. (Irreducible)** A transition kernel  $P$  on  $(E, \mathcal{E})$  is  $\pi$ -irreducible if  $\pi(E) > 0$  and for each  $x \in E$  and each  $A \in \mathcal{E}$  with  $\pi(A) > 0$ , there exists an integer  $n = n(x, A) \geq 1$  such that  $P^n(x, A) > 0$ .

A Markov chain with invariant distribution  $\pi$  is irreducible if, for any initial state, it has positive probability of entering any set to which  $\pi$  assigns positive probability.

**Definition 2.1.3. (Periodic)** A  $\pi$ -irreducible transition kernel  $P$  is periodic if there exists an integer  $d \geq 2$  and a sequence  $\{E_0, E_1, \dots, E_{d-1}\}$  of  $d$  nonempty disjoint sets in  $\mathcal{E}$  such that for all  $i = 1, \dots, d-1$  and all  $x \in E_i$ ,

$$P(x, E_j) = 1 \text{ for } j = i + 1 \pmod{d}.$$

In this case, we call  $C = \bigcup_{i=0}^{d-1} E_i$  a  $d$ -cycle. If  $P$  is not periodic, we call it **aperiodic**.

In other words, a chain is periodic if there are portions of the state space it can only visit at certain regularly spaced times.

**Definition 2.1.4. (Recurrence)** A  $\pi$ -irreducible chain  $\{X_n\}$  with invariant distribution  $\pi$  is recurrent if for each  $B$  with  $\pi(B) > 0$ ,

$$P_x\{X_n \in B \text{ i.o.}\} > 0 \text{ for all } x,$$

$$P_x\{X_n \in B \text{ i.o.}\} = 1 \text{ for } \pi\text{-almost all } x.$$

The chain is **Harris recurrent** if  $P_x\{X_n \in B \text{ i.o.}\} = 1$  for all  $x$ .

Here  $P_x\{A\}$  denotes the probability that event  $A$  happens when a Markov chain with transition kernel  $P$  starts at  $x$ . The notation  $\{A_n \text{ i.o.}\}$  means that sequence  $A_n$  occurs infinitely often, i.e.,  $\sum \mathbf{1}_{A_n} = \infty$ . The chain is called **positive recurrent** if the total mass of its invariant measure is finite; otherwise it is **null recurrent** (Note here we assume the chain is  $\pi$ -irreducible and  $\pi$ -invariant).

Theorem 2.1.5 summarizes the condition for the convergence of a Markov Chain.

The total variation norm used there is defined by

$$\|\mu\| = \sup_{A \in \mathcal{E}} \mu(A) - \inf_{A \in \mathcal{E}} \mu(A)$$

for a bounded signed measure  $\mu$  on  $(E, \mathcal{E})$ .

**Theorem 2.1.5.** Suppose  $P$  is  $\pi$ -irreducible and  $\pi P = \pi$ . Then  $P$  is positive recurrent and  $\pi$  is the unique invariant distribution of  $P$ . If  $P$  is also aperiodic, then for  $\pi$ -almost all  $x$ ,

$$\|P^n(x, \cdot) - \pi\| \rightarrow 0,$$

with  $\|\cdot\|$  denoting the total variation distance. If  $P$  is Harris recurrent, then the convergence occurs at all  $x$ . [73]

In fact, the assumptions in Theorem 2.1.5 are essentially necessary and sufficient: if  $\|P^n(x, \cdot) - \pi\| \rightarrow 0$ , for all  $x$ , then the chain is  $\pi$ -irreducible, aperiodic, positive Harris recurrent and has invariant distribution  $\pi$ .

In practice, given a Markov chain, we need to check the following rules to guarantee the convergence:

Rule 1. Check that  $\pi$  is a proper probability measure.

Rule 2. Check  $\pi P = \pi$ .

Rule 3. Check that  $P(\cdot, \cdot)$  is irreducible.

Rule 4. Check that  $P(\cdot, \cdot)$  is aperiodic.

Rule 5. Check Harris recurrence (optional).

Rule 6. Convergence diagnostics.

For Rule 6, several methods can be used to test the convergence of a Markov Chain (see, for example, Gamerman and Lopes [20]). Rule 5 is usually optional, but in many situations, it can be verified by the following results stated in Theorem 2.1.6 and Corollary 2.1.7.

**Theorem 2.1.6.** *If  $P$  is recurrent, then it is Harris recurrent if and only if every bounded harmonic function is a constant. [73]*

**Corollary 2.1.7.** *Suppose  $P$  is irreducible and  $\pi P = \pi$ . If  $P(x, \cdot)$  is absolutely continuous with respect to  $\pi$  for all  $x$ , then  $P$  is Harris recurrent. [73]*

### 2.1.2 Gibbs Sampling

Gibbs sampler constructs a Markov chain with invariant distribution  $\pi$  using conditioning. We give a simple definition for Gibbs sampler as in Gamerman and Lopes [20]. Let  $x = (x_1, \dots, x_d)^T$  and  $x \sim \pi$ . Each component of  $x$  can be a scalar or a vector. Assume that all full conditional distributions  $\pi_i(x_i|x_{-i}), i = 1, \dots, d$  are available, i.e., samples can be drawn from the conditional distributions. Here  $x_{-i}$  denotes the vector formed by knocking out  $x_i$  from  $x$ , i.e.,  $x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$ . A Gibbs sampler includes the following steps:

Step 1. Set initial value  $x^{(0)}$ .

Step 2. Based on current sample  $x$ , obtain a new sample  $\tilde{x}$  through successive generations of values:

$$\tilde{x}_1 \sim \pi_1(x_1|x_2, \dots, x_d),$$

$$\tilde{x}_2 \sim \pi_2(x_2|\tilde{x}_1, x_3, \dots, x_d),$$

⋮

$$\tilde{x}_d \sim \pi_d(x_d|\tilde{x}_1, \dots, \tilde{x}_{d-1});$$

Step 3. Repeat step 2 until convergence is reached.

**Example 2.1.8.** Consider  $E = \mathbb{R}^2$ .  $x \in E$  can be written as  $x = (x_1, x_2)^T$ , where  $x_1$  and  $x_2$  represents the two coordinates of  $x$ . Assume  $x \sim \pi$ , with

$$\pi(x) \propto C \exp \left\{ -\frac{1}{2}(x_1^2 + x_1^2 x_2^2 + x_2^2) \right\},$$

where  $C$  is a constant. From here we can easily find that the conditional density  $\pi_1(x_1|x_2) \propto c(x_2) \exp\{-\frac{1}{2}x_1^2(x_2^2 + 1)\}$  and  $\pi_2(x_2|x_1) \propto c(x_1) \exp\{-\frac{1}{2}x_2^2(x_1^2 + 1)\}$ , for  $c(x_1)$  and  $c(x_2)$  functions of  $x_1$  and  $x_2$ , respectively. This indicates that  $x_1|x_2 \sim N(0, \frac{1}{1+x_2^2})$ , and  $x_2|x_1 \sim N(0, \frac{1}{1+x_1^2})$ . A Gibbs sampler can thus be constructed as follows:

*Step 1.* Initialize  $x$ .

*Step 2.* For current value of  $x$ , obtain a new sample  $\tilde{x}$  through successive generations of values

$$\begin{aligned}\tilde{x}_1|x_2 &\sim N(0, \frac{1}{1+x_2^2}), \\ \tilde{x}_2|\tilde{x}_1 &\sim N(0, \frac{1}{1+\tilde{x}_1^2}).\end{aligned}$$

*Step 3.* Repeat step 2 until convergence is reached.

We now check Rule 1 to Rule 5 for the convergence of this Gibbs sampler. We first find the transition kernel  $P(x, A) = \Pr\{\tilde{x} \in A|x\}$  with the corresponding transition density  $\pi(\tilde{x}|x) = \pi((\tilde{x}_1, \tilde{x}_2)|(x_1, x_2)) = \pi_2(\tilde{x}_2|\tilde{x}_1)\pi_1(\tilde{x}_1|x_2)$ .

**Rule 1** Since  $\pi(x) \propto C \exp\{-\frac{1}{2}(x_1^2 + x_1^2 x_2^2 + x_2^2)\} \leq C \exp\{-\frac{1}{2}(x_1^2 + x_2^2)\}$ , and  $C \exp\{-\frac{1}{2}(x_1^2 + x_2^2)\}$  is integrable, hence  $\pi$  is a proper probability measure.

**Rule 2** Check  $\pi P = \pi$ .

$$\begin{aligned}
\pi P(A) &= \int_E P(x, A) \pi(dx) = \int_E \int_A \pi(\tilde{x}|x) \pi(x) d\tilde{x} dx \\
&= \iint_E \int_A \pi_2(\tilde{x}_2|\tilde{x}_1) \pi_1(\tilde{x}_1|x_2) \pi(x_1, x_2) d\tilde{x}_1 d\tilde{x}_2 dx_1 dx_2 \\
&= \iint_A \left[ \int_{\mathbb{R}} \pi_2(\tilde{x}_2|\tilde{x}_1) \pi_1(\tilde{x}_1|x_2) \left( \int_{\mathbb{R}} \pi(x_1, x_2) dx_1 \right) dx_2 \right] d\tilde{x}_1 d\tilde{x}_2 \\
&= \iint_A \left[ \int_{\mathbb{R}} \pi_2(\tilde{x}_2|\tilde{x}_1) \pi_1(\tilde{x}_1|x_2) \pi(x_2) dx_2 \right] d\tilde{x}_1 d\tilde{x}_2 \\
&= \iint_A \left[ \pi_2(\tilde{x}_2|\tilde{x}_1) \left( \int_{\mathbb{R}} \pi_1(\tilde{x}_1|x_2) \pi(x_2) dx_2 \right) \right] d\tilde{x}_1 d\tilde{x}_2 \\
&= \iint_A [\pi_2(\tilde{x}_2|\tilde{x}_1) \pi_1(\tilde{x}_1)] d\tilde{x}_1 d\tilde{x}_2 \\
&= \int_A \pi(\tilde{x}) d\tilde{x} = \pi(A), \forall A \in \mathcal{E}.
\end{aligned}$$

**Rule 3** Check that  $P(\cdot, \cdot)$  is irreducible. It is easy to see that that  $\pi(x)$  is fully supported on  $\mathbb{R}^2$ , thus  $E = \mathbb{R}^2$ . We then have  $\forall x \in E, \forall A \in \mathcal{E}$  with  $\pi(A) > 0$ ,

$$P(x, A) = \Pr\{\tilde{x} \in A|x\} > 0,$$

hence  $P(\cdot, \cdot)$  is irreducible by definition.

**Rule 4** Check that  $P(\cdot, \cdot)$  is aperiodic. From Rule 3, the chain can get anywhere starting from any  $x$  in one-step. Therefore  $P(\cdot, \cdot)$  is aperiodic.

**Rule 5** Check Harris recurrent. Since  $\pi(\tilde{x}|x) = \pi_2(\tilde{x}_2|\tilde{x}_1) \pi_1(\tilde{x}_1|x_2)$  and  $P(x, \cdot)$  is absolutely continuous with respect to  $\pi$ , Harris recurrent follows from Corollary 2.1.7. Therefore by Theorem 2.1.5, the Gibbs sampler constructed here converges to an equilibrium distribution  $\pi$  in total variation, and the convergence occurs for any starting values  $x \in \mathbb{R}^2$ .

### 2.1.3 Metropolis Sampling

Assume that  $\pi$  is absolute continuous with respect to  $\mu$  and let  $Q$  be a transition kernel of the form

$$Q(x, dy) = q(x, y)\mu(dy).$$

Let  $E^+ = \{x : \pi(x) > 0\}$  and assume that  $Q(x, E^+) = 1$  for  $x \notin E^+$ . Also assume that  $\pi$  is not concentrated on a single point. For a given  $X_n = x$ , we propose a candidate value  $Y = y$  for the next point  $X_{n+1}$  from the distribution  $Q(x, \cdot)$ , and accept it with probability

$$\alpha(x, y) = \min \left\{ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right\}.$$

Otherwise, the candidate is rejected and the chain remains at  $X_{n+1} = x$ .

If we define the off-diagonal density of a Metropolis kernel as

$$p(x, y) = q(x, y)\alpha(x, y)\mathbf{1}_{\{x \neq y\}},$$

and set  $r(x) = 1 - \int p(x, y)dy$ , then the Metropolis kernel  $P$  can be written as

$$P(x, dy) = p(x, y)\mu(dy) + r(x)\delta_x(dy), \quad (2.1)$$

where  $\delta_x$  denotes a point mass at  $x$ . The value  $r(x)$  is the probability that the algorithm remains at  $x$ .

**Proposition 2.1.9.** *For the Metropolis kernel defined above, we have*

$$\pi(x)p(x, y) = \pi(y)p(y, x), \quad (2.2)$$

*which is called reversibility condition.*

*Proof.* If  $x = y$ , then  $p(x, y) = 0$ , both sides equal 0. If  $x \neq y$  and  $\pi(y)q(y, x) \geq \pi(x)q(x, y)$ , we have  $\alpha(x, y) = 1$ . Therefore the left hand side(LHS) of Equation (2.2) is

$$\text{LHS} = \pi(x)p(x, y) = \pi(x)q(x, y)\alpha(x, y) = \pi(x)q(x, y).$$

The right hand side(RHS) of Equation (2.2) is

$$\text{RHS} = \pi(y)p(y, x) = \pi(y)q(y, x)\alpha(y, x) = \pi(y)q(y, x)\frac{\pi(x)q(x, y)}{\pi(y)q(y, x)} = \pi(x)q(x, y).$$

Therefore LHS=RHS, the equality holds. By symmetry, the case of  $\pi(y)q(y, x) < \pi(x)q(x, y)$  is obvious.  $\square$

**Proposition 2.1.10.** *For the Metropolis kernel defined above, we have  $\pi P = \pi$ , hence  $\pi$  is an invariant distribution for  $P$ .*

*Proof.* For all  $A \in \mathcal{E}$ , we have  $P(x, A) = \int_A p(x, y)\mu(dy) + r(x)\delta_x(A)$  by (2.1) and

$$\begin{aligned}\pi P(A) &= \int P(x, A)\mu(dx) \\ &= \int \left[ \int_A p(x, y)\mu(dy) \right] \pi(x)\mu(dx) + \int r(x)\delta_x(A)\pi(x)\mu(dx) \\ &= \int_A \left[ \int p(x, y)\pi(x)\mu(dx) \right] \mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\ &= \int_A \left[ \int p(y, x)\pi(y)\mu(dx) \right] \mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\ &= \int_A (1 - r(y))\pi(y)\mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\ &= \int_A \pi(y)\mu(dy) = \pi(A).\end{aligned}$$

$\square$

For the Metropolis kernel  $P$  to be irreducible, it is necessary that  $Q$  is irreducible. But this is not a sufficient condition because irreducibility of  $P$  depends on both  $Q$  and  $\pi$ . If  $P$  is irreducible and  $\pi(\{x : r(x) > 0\}) > 0$ , then the Metropolis kernel is aperiodic. [73]

**Corollary 2.1.11.** *Suppose  $P$  is a  $\pi$ -irreducible Metropolis kernel. Then  $P$  is Harris recurrent.* [73]

The Metropolis sampler is very general in the sense that there exists different choices for the “proposal” distribution  $q(x, y)$ . Tierney introduced four types of chains: random walk chains, independence chains, rejection sampling chains and grid based chains [73]. One can also combine different sampling algorithms to form a hybrid algorithm. More advanced algorithms can be found in Liu [40].

## 2.2 Bayesian Variable Selection

As a type of model selection method, Bayesian variable selection (BVS) has received much attention in recent years (see, for example, Chipman, George and McCulloch [16], Clyde and George [17] for literature reviews on this topic). In this section, we summarize the basic scheme of Bayesian variable selection for normal linear models based on the work of George and McCulloch ([21],[22]).

Given a dependent variable  $Y$  and  $p$  predictor variables  $\{X_1, \dots, X_p\}$ , a multiple

linear regression model takes the form

$$Y = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p + \epsilon, \quad (2.3)$$

where  $\epsilon \sim N(0, \sigma^2)$ . Here  $p$  can be large (e.g., larger than the number of observations).

The purpose of variable selection is to find a subset of the  $p$  predictors which can “best” explain the response  $Y$ . This often happens in the case when some predictors in  $\{X_1, \dots, X_p\}$  are redundant and a parsimonious model is sought. There are totally  $2^p$  choices for such a subset. When  $p$  is moderate (e.g., less than 20), one can go through all the possible choices and determine the best subset based on some selection criteria such as SSE, adjusted  $R^2$ ,  $C_p$ , AIC, BIC, etc. (see, for example, Kutner et al. [35], Page 353-360). When  $p$  is large, however, it becomes unrealistic to compute the criteria for all possible models. Therefore it becomes necessary to develop some efficient computational algorithms to search for the best subset. There are some traditional searching methods such as forward or backward selection (details can be found in Miller ([45], Page 42-46)). From a Bayesian point of view, this problem can be solved by formulating a hierarchical mixture prior to the regression coefficients, which is called Bayesian variable selection (BVS).

The BVS method introduces a hyper-parameter  $\tau$  to the priors of  $\beta_i$ ,  $i = 1, \dots, p$ , where  $\tau = (\tau_1, \dots, \tau_p)^T$ . Each component of  $\tau$  takes values either 1 or 0, indicating whether the corresponding regression coefficient is included in the subset. Posterior inferences of  $\tau$  then help to decide the best subset of the predictor variables. The prior distribution of  $\beta_i$  is usually set to be a mixture normal distribution controlled

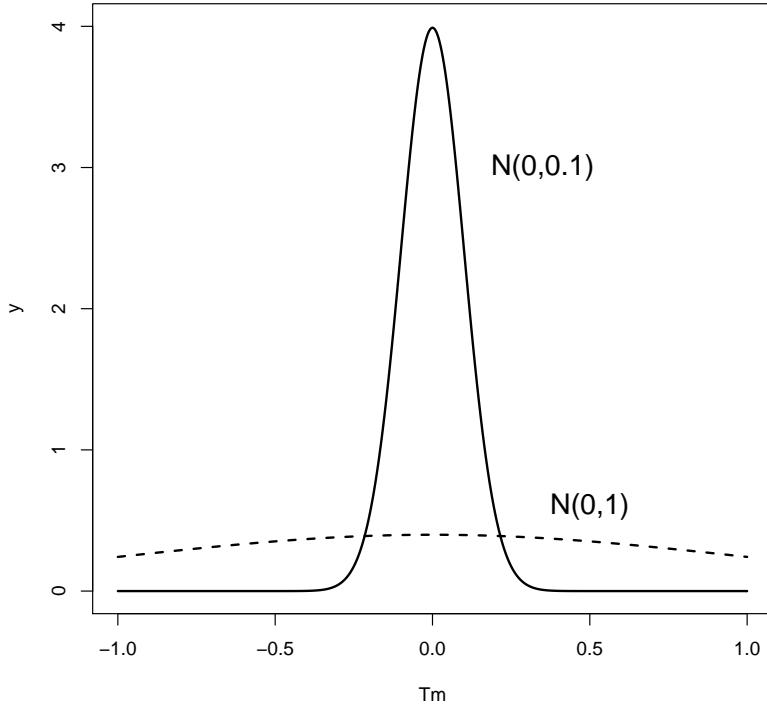


Figure 2.1: The plot of normal densities with relatively large (1) and small (0.1) variances.

by  $\tau$ . For example, the mixture normal prior can be

$$\beta_i | \tau_i \sim \tau_i N(0, v_{1i}^2) + (1 - \tau_i) N(0, v_{0i}^2), \quad (2.4)$$

where  $v_{1i}$  and  $v_{0i}$  are nonnegative parameters, and  $v_{1i}$  is far from zero but  $v_{0i}$  is close to zero, i.e.,  $v_{1i} \gg v_{0i} > 0$ . Usually we set  $v_{1i}$ 's and  $v_{0i}$ 's to be constant for all index  $i$ . The prior (2.4) is actually a normal distribution with variance either large or close to zero depending on the value of  $\tau_i$ . When  $\tau_i = 0$ ,  $\beta_i$  has a normal prior with small variance  $v_{0i}$ , and since  $v_{0i}$  is close to zero,  $\beta_i$  can be *a priori* excluded from the subset. Figure 2.1 shows the plot of two normal densities, one with relatively large (1) variance and the other with small (0.1) variance. One could also introduce

correlations between  $\beta_i$ 's by letting  $\beta = (\beta_1, \dots, \beta_p)^T$  and write priors for  $\beta$  as

$$\beta|\tau \sim N(0, D_\tau R_\tau D_\tau), \quad (2.5)$$

where  $D_\tau = \text{diag}(u_1, \dots, u_p)$  with  $u_i = \tau_i v_{1i} + (1 - \tau_i) v_{0i}$ , and  $R_\tau$  is the prior correlation matrix.  $\tau_i$  is usually set to have a hyper-prior of independent Bernoulli( $\omega$ ). The prior for  $\beta_0$  can be normal or non-informative (i.e.,  $\pi(\beta_0) \propto 1$ ). The prior for  $\sigma^2$  is often chosen to be the conjugate prior of the normal likelihood, i.e., Inverse-Gamma( $d_1, d_2$ ). Using Bayes theorem, the posterior distribution corresponding to the above prior settings can be determined as:

$$\pi(\tau, \beta_0, \beta, \sigma^2 | y) \propto \pi(y | \tau, \beta_0, \beta, \sigma^2) \pi(\beta_0) \pi(\beta | \tau) \pi(\tau) \pi(\sigma^2). \quad (2.6)$$

It is always possible to integrate out  $\beta_0$ ,  $\beta$  and  $\sigma^2$  from (2.6) to obtain the marginal posterior  $\pi(\tau | y)$ . MCMC algorithms can thus be designed to obtain the posterior samples of  $\tau$  based on  $\pi(\tau | y)$  or  $\pi(\tau, \beta_0, \beta, \sigma^2 | y)$ , which will be discussed later in this section.

As a modification of the mixture normal prior in (2.4), we can let  $v_{0i} \equiv 0$  so the prior for  $\beta_i$  becomes

$$\beta_i | \tau_i \sim \tau_i N(0, v_{1i}^2) + (1 - \tau_i) \delta_0, \quad (2.7)$$

where  $\delta_0$  is a point mass at zero. This prior is different from (2.4) in that when  $\tau_i = 0$ ,  $\beta_i$  follows a degenerate distribution (constant), hence the joint prior  $\pi(\beta | \tau)$  in (2.5) has singular covariance. In such a setting, we usually replace  $\beta$  by  $\beta_\tau$ , where  $\beta_\tau$  is a

sub-vector of  $\beta$  formed by removing the zero components of  $\beta$ . The prior in (2.5) is then reduced to

$$\beta_\tau | \tau \sim N(0, D_{1\tau} R_{1\tau} D_{1\tau}), \quad (2.8)$$

With this prior, the posterior distribution can be derived similarly as in (2.6).

The prior correlation  $R_\tau$  in (2.5) can be chosen to be an identity matrix or a so called g-prior  $R_\tau \propto (X^T X)^{-1}$ , where  $X$  is a  $n \times p$  design matrix when there are  $n$  observations. The  $i$ th row of  $X$  is  $(X_{i1}, \dots, X_{ip})$ . In case of the  $\beta_\tau$  prior in (2.8), the g-prior for  $R_{1\tau}$  takes the form  $R_{1\tau} \propto (X_\tau^T X_\tau)^{-1}$ , where  $X_\tau$  is formulated by removing the columns of  $X$  with zero coefficients (i.e., columns that the corresponding  $\tau$  components are 0).

The MCMC algorithm plays an important role in posterior inference. In case that one can integrate out  $\beta_0$ ,  $\beta$  and  $\sigma^2$  from the joint posterior to obtain the marginal posterior  $\pi(\tau|y)$ , several algorithms are available to sample  $\tau$  from  $\pi(\tau|y)$ , including:

1. **Gibbs Sampling.** A Gibbs sampling can be used to update  $\tau$  component-wisely. For each component  $\tau_i$ , compute the posterior odds

$$\theta_i = \frac{\pi(\tau_i = 1, \tau_{(i)}|y)}{\pi(\tau_i = 0, \tau_{(i)}|y)}, \quad (2.9)$$

where  $\tau_{(i)} = (\tau_1, \dots, \tau_{i-1}, \tau_{i+1}, \dots, \tau_p)$ . Using this ratio, we can compute the posterior probability of  $\tau_i = 1$  (i.e.,  $\theta_i/(1 + \theta_i)$ ) and sample  $\tau_i$  based on this probability.  $\tau_i$  can be updated in either a fixed or random order. It is also feasible to update components of  $\tau$  in groups rather than one by one.

## 2. Metropolis-Hastings.

Metropolis-Hastings is another choice to update  $\tau$ . We first generate a candidate sample  $\tilde{\tau}$  from a transition kernel (a proposal distribution)  $f(\tilde{\tau}|\tau)$ , then update  $\tau$  by  $\tilde{\tau}$  with probability

$$\min\left\{\frac{\pi(\tilde{\tau}|y)}{\pi(\tau|y)} \frac{f(\tau|\tilde{\tau})}{f(\tilde{\tau}|\tau)}, 1\right\}. \quad (2.10)$$

For convenience, the transition kernel can be chosen to be symmetric so that the  $f(\tau|\tilde{\tau})$  term and  $f(\tilde{\tau}|\tau)$  term in the proposal ratio in (2.10) are canceled. For example, the candidate sample  $\tilde{\tau}$  can be generated by one of the following operations to form a symmetric transition kernel:

- (a) Randomly change one component of  $\tau$ .
- (b) Randomly change  $d$  components of  $\tau$  with a pre-specified probability  $q_d$ .
- (c) With probability  $\phi$ , randomly change one component of  $\tau$ ; with probability  $1 - \phi$ , randomly choose two components with value 0 and 1 and swap them ([9], Page 524).

More adaptive sampling schemes can be found in Nott and Kohn [52]. Note that the MCMC algorithm will be different if using priors in (2.7) rather than that in (2.4). When using the point mass prior (2.7) to compute the posterior density  $\pi(\tau|y)$ , the dimension of the design matrix  $X$  need to be adjusted in each MCMC iteration according the value of  $\tau$ , i.e., for each proposed value  $\tilde{\tau}$ , the marginal posterior  $\pi(\tilde{\tau}|y)$  need to be computed by plugging in  $X_{\tilde{\tau}}$  rather than  $X$ . This may speed up the computation since only part of the data are

used in most iterations. When using the mixture normal priors in forms of (2.4), we do not have to adjust for the size of  $X$ .

When the parameters  $\beta_0$ ,  $\beta$  or  $\sigma^2$  can not be integrated out from the joint posterior, such as in the case of generalized linear models (see, for example, Nott [53]), we need to adopt more complex MCMC algorithm for posterior sampling. In such a case, if using point mass prior, the dimension of  $\beta$  varies when the number of “1” components in  $\tau$  changes. More advanced algorithms such as reversible jump MCMC can be applied for better mixing of the posterior samples.

# **Chapter 3**

## **A Bayesian Probit Model with Variable Selection for Functional Data Classification**

### **3.1 Introduction**

In this chapter, we propose a Bayesian variable selection (BVS) model to perform binary classification based on multiple functional predictors. We use a latent variable to connect the functional predictors with the binary response. Priors for the coefficient functions are set to be Gaussian processes which depend on a hyper-parameter that enables variable selection. An orthonormal basis is used to decompose the covariance function of the Gaussian process priors and to represent the functional predictors and

the coefficient functions by their basis coefficients. Posterior inference is implemented by function approximation with truncated orthonormal basis expansion. For posterior sampling, we suggest a Hybrid Gibbs/Metropolis-Hastings sampler. Simulations show that this model produces accurate variable selection and good classification results. Application to the EEM measurements of fluorescence spectroscopy data gives improved classification as compared to several other classification methods.

### 3.2 The Proposed Model

Suppose we observe  $n$  i.i.d. observations, each contains  $J$  functions. For  $i = 1, \dots, n$  and  $j = 1, \dots, J$ , denote  $x_{ij}(t)$  as the  $j$ th function observed from the  $i$ th observation. We assume  $x_{ij}(t) \in L^2(T_j)$  for a compact domain  $T_j$ . Let the response  $y_i$  be a binary class that the  $i$ th observation belongs to. Here  $y_i$ 's are assumed to be conditionally independent given the functional predictors  $x_{ij}(t), j = 1, \dots, J$ . Similar to the method used in James [32] as well as Müller and Stadtmüller [51], a generalized functional linear regression model for multiple functional predictors can be constructed by associating a univariate latent variable  $z_i$  with  $y_i$  through

$$y_i = \begin{cases} 1 & \text{if } z_i < 0, \\ 0 & \text{if } z_i \geq 0. \end{cases}$$

where

$$z_i = \beta_0 + \sum_{j=1}^J \int_{T_j} x_{ij}(s) \beta_j(s) ds + \epsilon_i, \quad (3.1)$$

and  $\epsilon_i \sim N(0, 1)$  determines a probit link between  $y_i$  and  $z_i$ . We assume  $\beta_j(t) \in L^2(T_j)$  for  $j = 1, \dots, J$ . Based on the above model setting, standard functional regression estimation paradigms, such as the EM algorithm in James [32], or the estimating equation method in Müller & Stadtmüller [51], can be performed to estimate the intercept  $\beta_0$  and the coefficient functions  $\beta_j(t)$ 's. However, these standard estimating paradigms are designed for cases with  $J = 1$ . It is not clear whether they can be extended to models with multiple functional predictors. Also, when the  $x_{ij}(t)$ 's contain redundant information, the efficiency of the model will be reduced. This motivates us to consider the variable selection method. Due to the infinite dimensionality of functional data, point-wise selection from the predictors  $x_{ij}(t)$  is not a practical choice. A simple method is to discretize  $x_{ij}(t)$  on a finite grid and transform the problem to a multivariate model, but this ignores the correlation between contiguous points on the grid. In this paper, we consider variable selection in the orthogonally transformed domain.

### 3.3 Priors

Based on the model proposed in Section 3.2, we construct priors to the regression coefficients from a functional data perspective. The priors are set to be

$$\begin{aligned} \beta_0 &\sim N(0, h^2), \\ \beta_j(t) | \tau^j &\sim GP(0, \gamma_{\tau^j}), \\ \tau_k^j &\sim \text{Bernoulli}(\omega_k^j), \quad k \in \mathbb{N}, j = 1, \dots, J. \end{aligned} \tag{3.2}$$

Here  $\tau^j = \{\tau_k^j\}_{k=1}^\infty$  is a binary sequence of 1's and 0's. Components of  $\tau^j$  are assumed to be independent across index  $k$  and  $j$ .  $GP(0, \gamma_{\tau^j})$  represents a Gaussian process with zero mean and covariance function  $\gamma_{\tau^j}$ . The covariance function  $\gamma_{\tau^j}$  can be decomposed as

$$\gamma_{\tau^j}(s, t) = \sum_{k=1}^{\infty} w_k^j [\tau_k^j \nu_1^2 + (1 - \tau_k^j) \nu_0^2] \phi_k^j(s) \phi_k^j(t), \tag{3.3}$$

where  $\{\phi_k^j\}_{k=1}^\infty$  is a complete orthonormal basis of  $L^2(T_j)$ , and  $\{w_k^j\}_{k=1}^\infty$  is a sequence of weights such that  $\sum_{k=1}^{\infty} w_k^j < \infty$ . We let  $\nu_1 \gg \nu_0 > 0$ , and let  $\nu_0$  to be close to zero so that the factor  $[\tau_k^j \nu_1^2 + (1 - \tau_k^j) \nu_0^2]$  is either  $\nu_1$  or  $\nu_0$  according to the binary value of  $\tau_k^j$ . Note that we treat  $\{w_k^j\}_k$  and  $\{\phi_k^j\}_k$  as prior parameters and will make specific choice of them. The values for  $h$ ,  $\nu_1$ ,  $\nu_0$  and  $\omega_k^j$ 's are also pre-specified. For simplicity, we assume the priors for  $\beta_j(t)$  are independent across index  $j$ .

### 3.4 Posteriors

Based on the model in Section 3.2 and prior settings in Section 3.3, posterior inference can be conducted by finite dimensional approximation. Since  $\{\phi_k^j\}_{k=1}^\infty$  is an orthonormal basis on  $L^2(T_j)$ , we can expand  $x_{ij}(t)$ 's and  $\beta_j(t)$ 's by

$$x_{ij}(t) = \sum_{k=1}^{\infty} c_{ijk} \phi_k^j(t), \quad \beta_j(t) = \sum_{k=1}^{\infty} b_{jk} \phi_k^j(t). \quad (3.4)$$

The truncated version of (3.4) can be used to approximate  $x_{ij}(t)$  and  $\beta_j(t)$  since  $\sum_{k=1}^{\infty} c_{ijk}^2 < \infty$  and  $\sum_{k=1}^{\infty} b_{jk}^2 < \infty$ . Note that the orthonormal basis  $\{\phi_k^j\}_{k=1}^\infty$  can be chosen to be a known basis such as a Fourier or wavelet basis. If we assume in addition that  $x_{ij}(t)$ 's have zero mean and  $\int_{T_j} E[x_{ij}(t)^2] dt < \infty$ , Mercer's theorem and Karhunen-Loève theorem (Ash and Gardner [3]) suggest to take the orthonormal basis to be the eigenfunctions of the covariance operator  $K$  defined by

$$Kx(t) = \int x(s) k(s, t) ds, \quad k(s, t) = Cov(x(s), x(t)). \quad (3.5)$$

In this case, the coefficients  $\{c_{ijk}\}_{k=1}^\infty$  are called functional principal component (FPC) scores of  $x_{ij}(t)$ . The FPC method is different with orthonormal expansion using known basis in that the eigenfunctions need to be estimated. Various methods for estimating the eigenfunctions can be found in Ramsay and Silverman [60], Hall, Müller and Wang [28].

Once the orthonormal basis has been chosen or estimated, we can approximate Equation (3.1) by

$$z_i = \beta_0 + \sum_{j=1}^J \sum_{k=1}^{p_j} c_{ijk} b_{jk} + \epsilon_i, \quad (3.6)$$

where  $p_j$  is the truncation parameter for the  $j$ th functional predictor. We thus transfer the functional regression to a multiple linear regression. For convenience, denote

$$C_i = (1, c_{i11}, \dots, c_{i1p_1}, \dots, c_{iJ1}, \dots, c_{iJp_J})^T,$$

$$\beta = (\beta_0, b_{11}, \dots, b_{1p_1}, \dots, b_{J1}, \dots, b_{Jp_J})^T,$$

Equation (3.6) can be simplified to

$$z_i = C_i\beta + \epsilon_i. \quad (3.7)$$

Let  $Z = (z_1, \dots, z_n)^T$ ,  $Y = (y_1, \dots, y_n)^T$  and  $X = (C_1, \dots, C_n)^T$ , the conditional density  $\pi(Z|\beta, Y)$  is

$$\prod_{i=1}^n \phi(z_i - C_i\beta) [\Phi^{-1}(-C_i\beta) I_{\{z_i < 0\} \cap \{y_i = 1\}} + (1 - \Phi(-C_i\beta))^{-1} I_{\{z_i \geq 0\} \cap \{y_i = 0\}}], \quad (3.8)$$

where  $\phi(\cdot)$  represents a standard normal density with corresponding distribution function  $\Phi(\cdot)$ , and  $I_{\{\cdot\}}$  is an indicator function. Equation (3.8) shows that the conditional distribution of  $Z$  given  $\beta$  and  $Y$  is truncated normal.

Using the truncated orthonormal basis expansion, the priors for  $\beta_j(t)$ 's in Equation (3.2) become

$$\pi(\beta|\tau) = N(0, \Sigma_\tau), \quad (3.9)$$

where  $\tau = (\tau_1^1, \dots, \tau_{p_1}^1, \dots, \tau_1^J, \dots, \tau_{p_J}^J)$  and

$$\Sigma_\tau = D_\tau W^{1/2} R W^{1/2} D_\tau. \quad (3.10)$$

Here we have  $R = I$  because of the independence assumption between  $\beta_j(t)$ 's, and

$$W = \text{diag}(1, w_1^1, \dots, w_{p_1}^1, \dots, w_1^J, \dots, w_{p_J}^J). \quad (3.11)$$

Finally,  $D_\tau = \text{diag}(h, \nu_{11}, \dots, \nu_{1p_1}, \dots, \nu_{J1}, \dots, \nu_{Jp_J})$  with

$$\nu_{jk} = \tau_k^j \nu_1 + (1 - \tau_k^j) \nu_0, \quad (3.12)$$

for  $k = 1, \dots, p_j$  and  $j = 1, \dots, J$ . The diagonal form of  $\Sigma_\tau$  makes the components of  $\beta$  *a priori* independent.  $\nu_{jk}$ 's in the diagonal of  $D_\tau$  have mixture normal priors, which indicate whether the components of  $\beta$  have large or nearly zero variances. Such a prior was used in George and McCulloch ([21], [22]) for Bayesian variable selection in multiple linear regression.

The joint posterior distribution can therefore be obtained by multiplying conditional distribution in Equation (3.8) with the priors, i.e.,

$$\pi(\beta, \tau, Z|Y) = \pi(Z|\beta, \tau, Y)\pi(\beta|\tau)\pi(\tau) \quad (3.13)$$

Integrating out  $\beta$  from Equation (3.13) gives the marginal posterior density  $\pi(\tau, Z|Y)$ .

Conditional on  $Z$  and  $Y$ , we have

$$\pi(\tau|Z, Y) \propto |X^T X + \Sigma_\tau^{-1}|^{-\frac{1}{2}} |\Sigma_\tau|^{-\frac{1}{2}} \exp \left\{ \frac{1}{2} Z^T X (X^T X + \Sigma_\tau^{-1})^{-1} X^T Z \right\} \pi(\tau). \quad (3.14)$$

Based on Equation (3.8), (3.13) and (3.14), we can design a MCMC algorithm for posterior inference.

### 3.5 Parameter Settings

Note that in Section 3.3, the truncation parameters  $p_j$  are pre-determined parameters for function approximation. One could set up priors for each  $p_j$  and adopt reversible jump MCMC[26] for posterior sampling. This strategy is reasonable but causes extra complications for MCMC. Another way of determining  $p_j$  is through cross-validation, i.e., maximizing the prediction performance on test set. This method is straightforward but only applicable for  $p_j \equiv p$ . It is also computationally expensive since it requires training the model on all possible choices of  $p$ . In this study, we propose a simple practical method for determining  $p_j$ 's by setting an approximation criterion. For example, if we use FPC analysis, the criterion can be set as  $\hat{f}(p_j) = \sum_{k=1}^{p_j} \hat{\lambda}_k / \sum_{k=1}^K \hat{\lambda}_k \geq c_1$ , for  $0 < c_1 \leq 1$ ,  $1 \leq p_j \leq K$ . Here  $\hat{\lambda}_k$ 's are the estimated eigenvalues,  $K$  is the maximum number of non-zero eigenvalues. Note that  $\hat{f}(p_j)$  represents the proportion of variability explained by the first  $p_j$  FPC's. Empirically we often choose  $c_1$  between 0.99 and 1. In the case of using a known orthonormal basis, we suggest the criterion to be  $\hat{f}(p_j) = 1 - \sum_i \|x_{ij}(t) - \hat{x}_{ij}(t)\|^2 / \sum_i \|x_{ij}(t)\|^2 \geq c_2$ , where  $\hat{x}_j(t)$  is the estimated function of  $x_j(t)$  after truncating at  $p_j$ , and  $\|\cdot\|$  is the  $L^2$  norm. Similarly, the suggested value for  $c_2$  is also between 0.99 and 1.

The weights sequences  $\{w_k^j\}_{k=1}^\infty$  in Equation (3.3) determine the weight matrix  $W$  in (3.10). Here we give a brief discussion on the choices of  $\{w_k^j\}_{k=1}^\infty$ . First we know that  $w_k^j > 0$  and  $\sum_{k=1}^\infty w_k^j < \infty$ . The main effect of  $w_k^j$  is to shrink more on the higher orders of the orthonormal basis  $\{\phi_k^j(t)\}$  toward zero so that the series in

(3.3) converges. In this paper, we always set  $1 = w_1^j > w_2^j > \dots > 0$  so that all the weights are between 0 and 1. Let  $w_k^j = m_1^{(k-1)m_2}$  for all  $k = 1, \dots, \infty$  and all  $j$ , where  $0 < m_1 < 1$  and  $m_2$  is a positive integer. Clearly, smaller value of  $m_1$  or larger value of  $m_2$  makes  $\{w_k^j\}_{k=1}^\infty$  decay to zero faster. The values of  $\{w_k^j\}_{k=1}^\infty$  are truncated at  $p_j$  to form the weight matrix  $W$ . We usually take  $m_1$  between 0.7 and 1, and  $m_2$  to be 1, 2 or 3.

The prior parameters  $\nu_1$  and  $\nu_0$  must satisfy  $\nu_1 >> \nu_0 > 0$ . Usual value for  $\nu_1$  is between 10 and 1000, and for  $\nu_0$  is between 0.0001 and 0.2.

### 3.6 Markov Chain Monte Carlo

Based on the results derived in Section 3.2 through Section 3.4, we propose the following MCMC algorithm for posterior sampling:

Step 0: Set up initial values for  $\beta$ ,  $\tau$  and the prior parameters for  $h$ ,  $\nu_1, \nu_0$  and  $w_k^j$ 's.

Step 1: Conditional on  $Y$  and current values of  $\beta$ , sample  $Z$  from the truncated normal distribution with density (3.8).

Step 2: Conditional on  $Y$  and current values of  $Z$ , update  $\tau$  using Metropolis-Hastings. Based on current  $\tau$ , a candidate  $\tau^c$  is firstly generated using the “switch/swap” proposal (see Brown et al. [8]), i.e., with probability  $\varphi$ , randomly swap one 1 term with one 0 term; and with probability  $1 - \varphi$ ,

randomly pick one position and switch it. Compute the ratio

$$r_\tau = \frac{\pi(\tau^c | Z, Y)}{\pi(\tau | Z, Y)},$$

and update  $\tau = \tau^c$  with probability  $\min(1, r_\tau)$ .

Step 3: Conditional on  $Y$  and current  $Z, \tau$ , update  $\beta$  from a multivariate normal distribution:

$$\beta | Z, \tau, Y \sim N((X^T X + \Sigma_\tau^{-1})^{-1} X^T Z, (X^T X + \Sigma_\tau^{-1})^{-1})$$

Repeat Step 1 – 3 until convergence.

This MCMC algorithm is a hybrid Gibbs/Metropolis-Hastings sampling process since it performs Metropolis-Hastings updates within a large Gibbs sampling iteration. Note that although  $\tau_j = 0$  indicates that the  $j$ th covariate (among the concatenated basis coefficients of the functional predictors) is not selected, we do not remove this covariate in the MCMC iteration.

### 3.7 Simulation Study

Two simulations are conducted to evaluate the performance of the proposed BVS model on functional data classification. Simulation 1 uses only one functional predictor, i.e.,  $J = 1$  in Equation (3.1). For simplicity, the functional predictor is generated using only 5 orthonormal cosine bases on interval  $[0, 1]$ . Simulation 2 considers multiple functional predictors for each observation, i.e.,  $J = 20$  in Equation (3.1). Thus

the total number of variables to be selected is relatively large. The variable selection results are discussed and prediction results are compared with several other classifiers.

*Simulation 1:* Let the sample size  $n = 1000$ , we simulate a single functional predictor for each observation, i.e.,  $J = 1$  in Equation (3.1). Functional predictors  $x_i(t)$  are generated using the first 5 cosine bases on closed set  $[0, 1]$ , i.e.,  $\phi_0(t) = 1, \phi_k(t) = \sqrt{2} \cos(k\pi t), k = 1, \dots, 4$ . The mean curve is determined by cosine coefficients  $c = (-1.12, -1.82, 7.77, 2.15, -3.25)$ . By adding an independent random error  $N(0, 1)$  to each component of  $c$ , we generate the functional predictor for each observation. For the true coefficient function  $\beta(t)$ , we set the first 5 cosine bases scores as  $b_1 = b_3 = b_4 = 0, b_2 = 5$ , and  $b_5 = -4$ , corresponding to the true value of  $\tau = (0, 1, 0, 0, 1)^T$ . Latent variables  $z_i$  are generated using Equation (3.1) by numerical integration. Here the true  $\beta_0$  is set to be  $-3.5$ . Binary responses  $y_i$  are generated from the sign of  $z_i$ . We randomly take 800 observations as training set and the rest as test set. Note that in this simulation, the way of functional data generation is actually multivariate, in the sense that all the true parameters are pre-defined as the coefficients of a fixed number of cosine bases. This simplified simulation helps to verify our proposed model and MCMC algorithm in a straightforward way.

The proposed model is applied to the above simulated data. For convenience of comparing the estimated regression coefficients with the true on their basis coefficients, we choose to use cosine basis to approximate the functional predictors. The criterion in Section 3.5 with  $c_2 = 0.99$  gives the truncation parameter  $p = 5$ . This

<i>True</i>		<i>MLE</i>		<i>BVS</i>				
$\tau$	$\beta$	$\hat{\beta}$	<i>S.E.</i>	$\hat{\beta}$	<i>S.E.</i>	95% 2.5%	<i>C.I.</i>	$\hat{\omega}_i$
						95% 2.5%	97.5%	
-	-3.5	-2.28	1.02	-2.77	0.46	-3.67	-1.88	-
0	0	0.12	0.11	0.00	0.00	0.00	0.00	0.0
1	5	4.41	0.37	4.37	0.39	3.64	5.18	1.0
0	0	0.01	0.12	0.00	0.00	0.00	0.00	0.0
0	0	-0.25	0.12	0.00	0.00	0.00	0.00	0.0
1	-4	-3.51	0.30	-3.44	0.31	-4.09	-2.87	1.0

Table 3.1: Simulation 1: the estimation of  $\beta$  compared with maximum likelihood estimation(MLE). Note that  $\omega_i$  indicates  $P\{\tau_i = 1\}$ . BVS: The Bayesian variable selection model proposed in Section 3.2.

model is trained on the training set using the MCMC algorithm stated in Section 3.6, with  $\omega_i \equiv \omega = 0.2$ ,  $R = I$ ,  $\nu_1 = 100$ , and  $\nu_0 = 0.001$ . The weight sequence  $\{w_k\}_k^\infty$  is set by the method stated in Section 3.5 with parameters  $m_1 = 0.9$ ,  $m_2 = 1$ . The Markov chain consists of 20000 iterations in total with a 3000 burn-in period. By averaging the posterior samples of  $\tau$ , we obtain the marginal posterior probability  $P\{\tau_i = 1, i = 1, \dots, 5\}$  as  $(0, 1, 0, 0, 1)^T$ , which indicates that our algorithm has picked out the correct non-zero basis (second and fifth) scores successfully. Table 3.1 lists the estimation results for  $\beta$ , using the BVS model and the maximum likelihood estimation method (the GLM with probit-link). From Table 3.1, we see that the posterior estimation of the coefficient scores is as good as the maximum likelihood estimate. The posterior prediction of the coefficient curve  $\beta(t)$  can be easily computed by conducting inverse cosine transform to the posterior samples of  $\{b_k, k = 1, \dots, 5\}$ . Figure 3.1 shows the posterior mean of the coefficient function and the corresponding

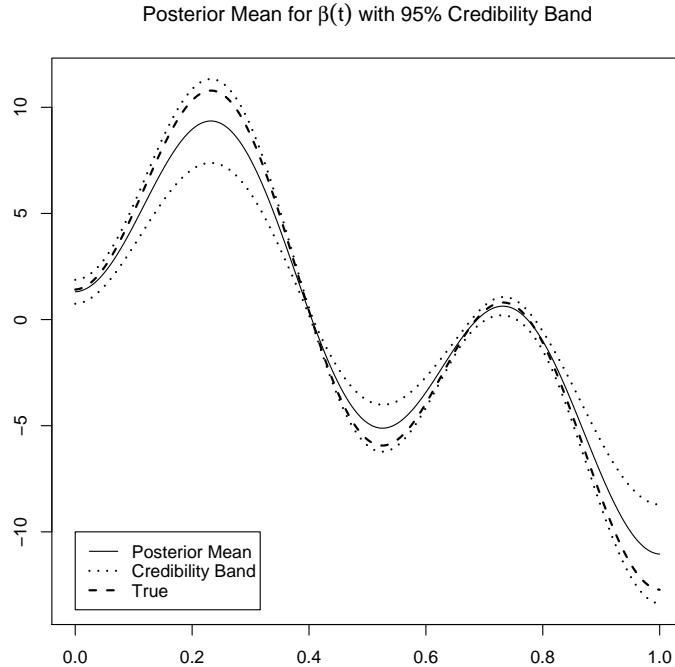


Figure 3.1: Simulation 1: the posterior estimation of  $\beta(t)$  and the corresponding simultaneous 95% credibility band compared with the true value of  $\beta(t)$ .

simultaneous 95% credibility band, as compared with the true. The simultaneous credibility band is obtained by finding a constant  $M$ , such that 95% of the simulated posterior functions fall into the interval  $\hat{\beta}(t) \pm M\hat{\sigma}(t), \forall t$ , where  $\hat{\beta}(t)$  and  $\hat{\sigma}(t)$  are the posterior mean and standard deviation of the coefficient functions. From Figure 3.1, we see that the true coefficient function lies in the 95% credibility band.

Prediction can be done by applying the posterior samples of  $\beta$  to the test set using Equation (3.6). If treating  $y_i = 1$  as diseased and  $y_i = 0$  as normal class, the out-of-sample prediction of the test set provides sensitivity 92.7% and specificity 97.1% with corresponding threshold 0.526. The resulting misclassification rate is 5% and the area under ROC curve (AUC) is 0.99. Note that the sensitivity and specificity

reported here is obtained by maximizing the sum of sensitivity and specificity. For more information about ROC curves, see Zweig and Campbell [86].

Instead of using cosine basis for dimension reduction, we also tried to use FPC for orthonormal basis expansion. We use the approximation criterion stated in Section 3.5 with  $c_1 = 0.99$ , and get  $p = 5$ . The prior parameters are set to be the same as in the cosine basis case. After 20000 MCMC iterations with a 5000 burn-in period, prediction on the test set gives sensitivity of 96.9% and specificity of 91.3% under threshold 0.282. The corresponding misclassification error is 6% and the area under ROC curve (AUC) is 0.988. These results shows that using FPC for function approximation produces as accurate prediction as using cosine basis, although the data are generated based on a different type of basis.

*Simulation 2:* In this simulation, we evaluate the performance of the model with multiple functional predictors. The functional predictors are generated similarly as in simulation 1 using the first 5 cosine bases, except that now we set  $J = 20$  in Equation (3.1). Therefore the total number of scores  $K = J \times p = 100$ . For the coefficient scores  $\beta$ , we randomly choose 24 out of 100 and set them to be nonzero, which take values from a uniform distribution with support  $[-4, 5]$  (the 0 value is excluded). We set the intercept  $\beta_0 = -1.5$ . Latent variables and binary responses are generated following the same way as in Simulation 1.

Similar to Simulation 1, we choose cosine basis to approximate the functional predictors for simplicity. The approximation criterion in Section 3.5 with  $c_2 = 0.99$

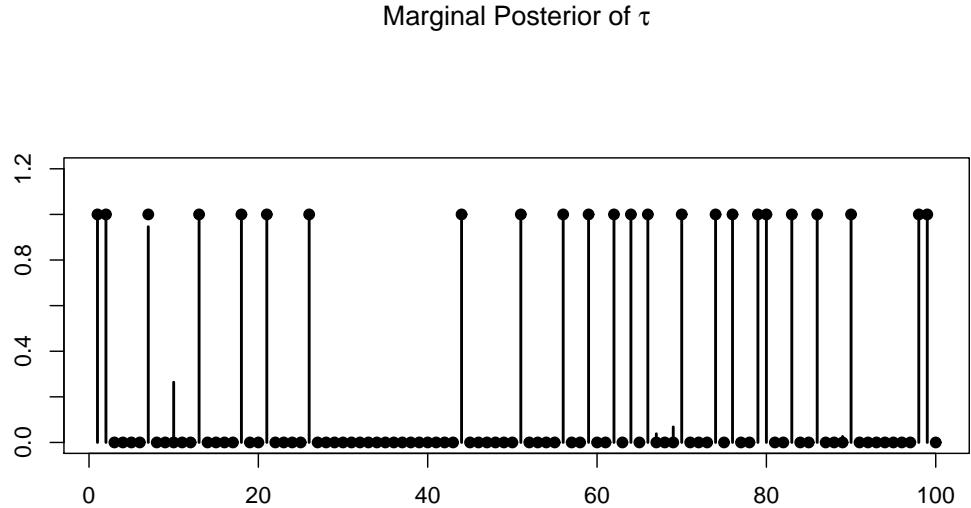


Figure 3.2: Simulation 2: marginal posterior estimate of  $\tau$  as compared with the true  $\tau$ . The solid dots represent the true values of  $\tau$ . The vertical bars indicate the frequencies of selecting the variables during all iterations (after burn-in).

gives truncation parameter  $p = 5$ . We train the proposed BVS model using the training set based on the transformed cosine basis scores. The model priors are set to be  $\omega_k^j \equiv \omega = 0.1$ ,  $R = I$ ,  $\nu_1 = 10$  and  $\nu_0 = 0.001$ . The weight sequences  $\{w_k^j\}_{k=1}^\infty$  are determined by  $m_1 = 0.9$ ,  $m_2 = 1$  for  $j = 1, \dots, J$ , as suggested in Section 3.5. The Markov chain consists of 30000 iterations in total with a burn-in period of 10000. Figure 3.2 shows that the estimated marginal posterior probability  $\Pr\{\tau_1 = 1, \dots, \tau_K = 1\}$  as compared with the true  $\tau$ . From Figure 3.2, we see that all 24 nonzero components of  $\beta$  are correctly found. The marginal posterior estimate for  $\tau$  matches perfectly to the true  $\tau$ . These results show that, even with fairly large number of functional predictors  $J = 20$ , the proposed model is still able to provide accurate estimates of  $\tau$ .

Applying the estimated regression coefficients to the test set for prediction, we

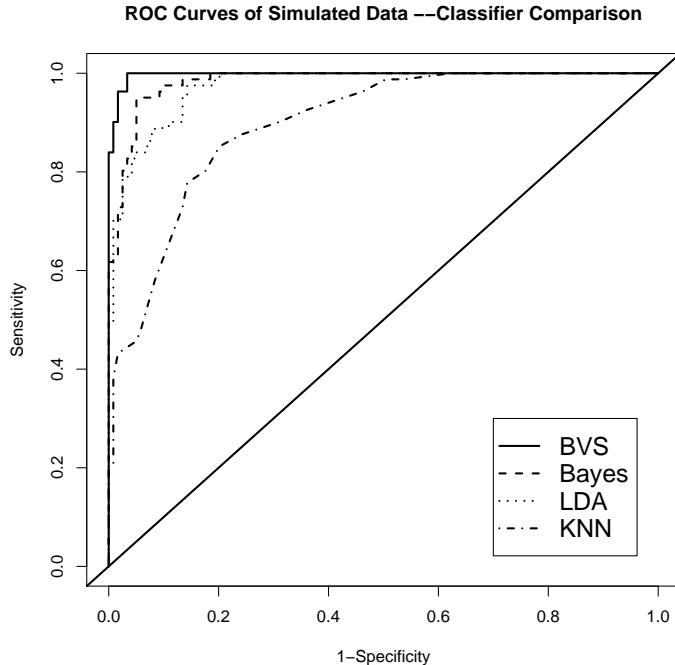


Figure 3.3: Simulation 2: the ROC curves of different classification models. BVS: the proposed Bayesian variable selection model. Bayes: the Bayesian probit model (without variable selection). LDA: Linear Discriminant Analysis. KNN: K-nearest neighbor. Note that all classifiers are based on first 5 cosine basis scores

obtain a 100% sensitivity and 96.6% specificity under the threshold 0.106. The corresponding misclassification rate is 2%. We then evaluate in Figure 3.3 the prediction performance by comparing the empirical ROC curve of the proposed model with that of three other classifiers. All the 4 methods are based on the same function approximation method, i.e., the cosine basis expansion with truncation parameter  $p_j \equiv 5$ . Among these methods, the Bayes classifier is a Bayesian probit model with latent variables. It has the similar structure as our proposed model but does not perform variable selection. The LDA classifier assumes multivariate normal distribution with common covariance matrix for both classes, and obtains the discrimination hyper-

<i>Method</i>	<i>AUC</i>	<i>Sens</i>	<i>Spec</i>	<i>Thres</i>	<i>MisR</i>
BVS	0.997	100%	96.6%	0.106	2%
Bayes	0.983	95.1%	95.0%	0.329	5%
LDA	0.974	97.5%	85.7%	0.232	9.5%
KNN	0.887	85.2%	79.8%	0.400	18%

Table 3.2: Simulation 2: the prediction results compared with 3 other classification methods. AUC: Area under the ROC curve; Sens: sensitivity; Spec: specificity; Thres: The threshold corresponding the reported sensitivity and specificity; MisR: misclassification rate. The BVS, Bayes, LDA and KNN are defined same as in Figure 3.3.

plane by equalizing the posterior densities of the two classes. Details of LDA can be found in Hastie, Tibshirani and Friedman ([30], Page 84-90). The KNN classifier is another popular classification method, which assigns category for the points in the test set by voting from their  $k$  closest points in the training set. The number of neighbors  $k$  is determined by a 20 block cross-validation using the training set. The criterion used in the cross-validation is the sum of sensitivity and specificity. Detailed prediction results are reported in Table 3.2. Note that the sensitivities and specificities listed in Table 3.2 are obtained by maximizing the sums of sensitivities and specificities on the ROC curves. Both Figure 3.3 and Table 3.2 show that the proposed variable selection model provides better prediction results.

### 3.8 Fluorescence Spectroscopy Data Classification

After evaluated by simulation, the proposed BVS model is applied to the fluorescence spectroscopy data introduced in Section 1.2. In this study, we choose part of the

clinical data measured by a fixed instrument (called FastEEM2). There are 1013 EEM measurements in this dataset obtained from 521 patients. These measurements are taken from different sites of the patient cervix and there may exist repeated measurements at the same site. To reduce possible confounding effects due to the tissue type, all normal measurements are from squamous tissue. After necessary pre-processing procedures like background correction, smoothing and registration, the EEM measurements are split randomly into a training set with 607 measurements and a test set with 406. The proportions of diseased cases within each set are 0.096 and 0.080, respectively. Both cosine basis and FPC are used to reduce the dimension of functional predictors. The truncation parameters are determined using approximation criteria suggested in Section 3.5 with  $c_1 = 0.999$  in the FPC case and  $c_2 = 0.99$  in the cosine basis case. The resulting  $p_j$ 's vary from 5 to 3 using the FPC method, and from 7 to 4 using cosine basis expansion. To reduce possible bias, the principal component scores of the test set is computed based on eigenfunctions estimated from the training set.

The proposed model is applied to the scores obtained from FPC and cosine basis expansion. For both types of scores, we set the priors as  $\omega_k^j \equiv 0.2$ ,  $\nu_1 = 100$ ,  $\nu_0 = 0.001$ ,  $R = I$  with 40000 MCMC iterations and 10000 burn-in period. The weight sequences  $\{w_k^j\}_k$  are determined as suggested in Section 3.5 with parameters  $m_1 = 0.9$ ,  $m_2 = 1$ . Figure 3.4 shows the marginal posterior probabilities of  $\tau_i = 1$  for all components of  $\tau$  in the FPC case. The x-axis represents the FPC scores from a single

excitation curve, and the y-axis represents the spectroscopy curves. Figure 3.4 shows that, in the total 60 principal component scores, only 4 have posterior probability greater than 0.4, and 3 of these scores are the third or higher principal components. One can also find the joint posterior distribution of  $\tau$  based on the frequencies of the  $\tau$  values visited during MCMC. In this real data study, there are  $2^{60}$  possible choices for  $\tau$  in total. It turns out that the frequencies for the visited models are all very small. For example, in the total of 30000 iterations(after burn-in), the most frequently visited model has a frequency of 5%. Similar to Simulation 2, we compare the posterior prediction result of the proposed model to that of three other classifiers in Table 3.3. Figure 3.5 shows the corresponding ROC curves obtained from the test set prediction. Both Table 3.3 and Figure 3.5 show that the proposed BVS model provides a better prediction than the other three classifiers in both cases of function approximation. FPC method gives 77% sensitivity and 82% specificity with area under ROC curve 0.84, whereas cosine basis expansion gives higher sensitivity but lower specificity.

To assess the convergence of the MCMC algorithm, we run multiple chains starting from different initial values of  $\tau$ . The initial values of  $\beta$  are chosen by randomly sampling its components from a normal distribution. Figure 3.6 illustrates the marginal posterior probabilities of  $\tau$  obtained from 3 different chains with different initial values. The first chain starts with a  $\tau$  with every component being assigned to be 1 or 0 randomly with probability 0.5; the second chain starts with a  $\tau$  of all 1's; the

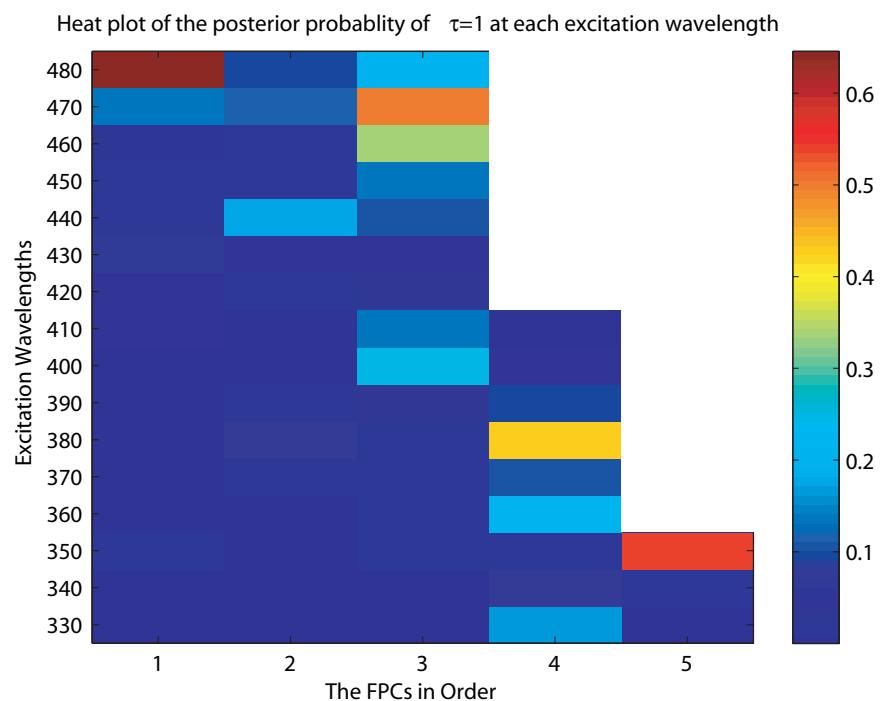


Figure 3.4: Real data application: the posterior probability of  $\tau_i = 1$  for all the scores obtained using FPC.

<i>Method</i>	<i>FPCA</i>					<i>Cosine</i>				
	<i>AUC</i>	<i>Sens</i>	<i>Spec</i>	<i>Thres</i>	<i>MisR</i>	<i>AUC</i>	<i>Sens</i>	<i>Spec</i>	<i>Thres</i>	<i>MisR</i>
BVS	0.84	77%	82%	0.13	18%	0.83	87%	72%	0.12	27%
Bayes	0.72	90%	48%	0.02	49%	0.80	90%	67%	0.09	31%
KNN	0.71	60%	84%	0.10	22%	0.73	57 %	88%	0.15	15%
LDA	0.68	77%	54%	0.03	45%	0.75	93%	54%	0.02	44%

Table 3.3: A comparison of four classification methods. FPCA: Using the functional principal components. Cosine: Using cosine basis. Sens, Spec, MisR and BVS, KNN, LDA, SVM are defined same as in Table 3.2 and Figure 3.3. The thresholds are determined by maximizing the sum of sensitivities and specificities on the empirical ROC curves.

third chains starts with a  $\tau$  of all 0's. From Figure 3.6, we see similar patterns on the marginal posterior probabilities, although there are slight differences at some components.

### 3.9 Conclusion

We have proposed a Bayesian variable selection model for binary classification, evaluated its performance by simulation and applied it to fluorescence spectroscopy data. This model uses a probit link to connect the binary responses with the functional predictors, and conducts variable selection by introducing a binary sequence to the Gaussian process prior of the coefficient function. The posterior inference is performed by function approximation using orthonormal basis. Compared with several other classifiers, the proposed model shows better prediction results in both simulation studies and real data application.

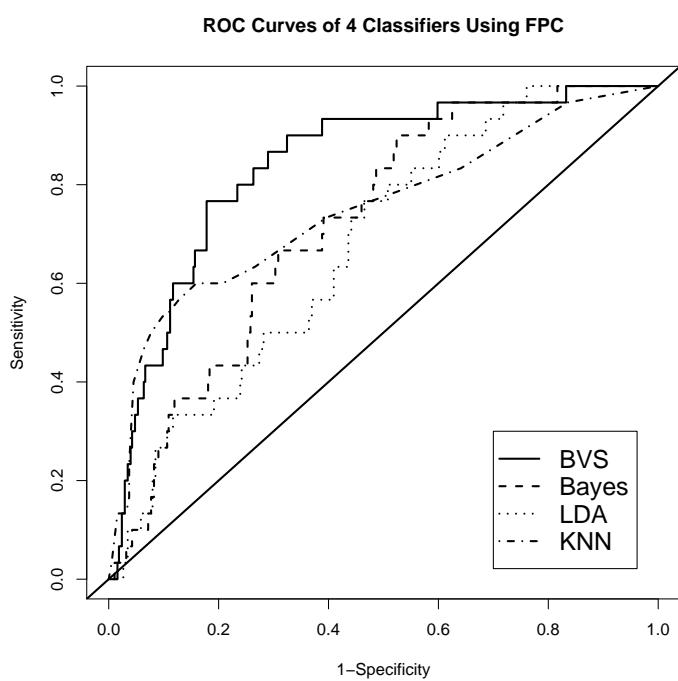


Figure 3.5: Real data application: empirical ROC curves for the test set.

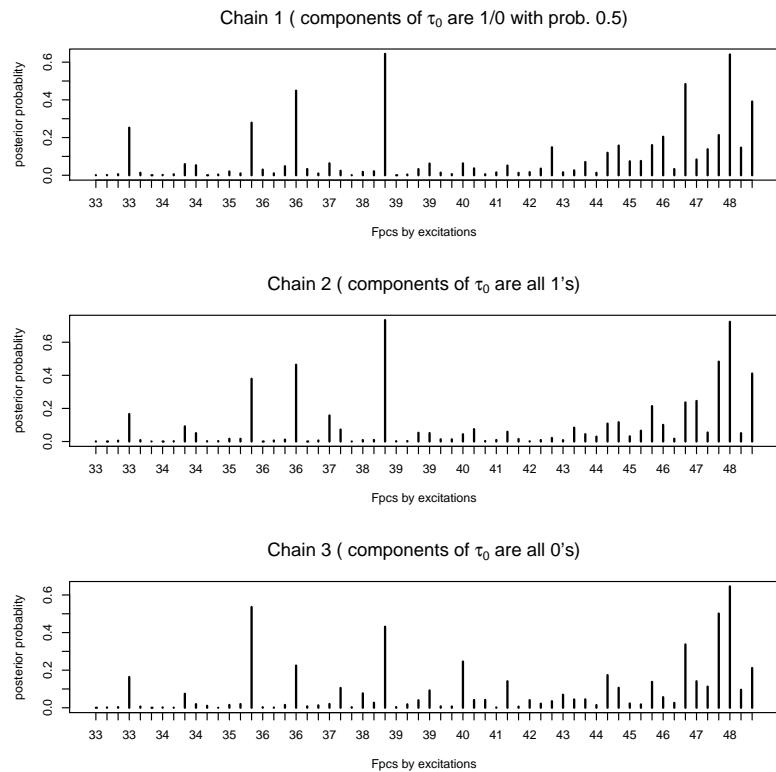


Figure 3.6: Real data application: marginal posterior of  $\tau$  obtained from different chains trained with different initial values.

# Chapter 4

## A Functional Generalized Linear Model with Functional Predictor Selection

### 4.1 Introduction

This chapter continues the study of binary classification with multiple functional predictors, with a particular emphasis on selecting functional predictors. This study is motivated by such a fact: when multiple functional predictors are involved in classification, some functions usually play more important role while others produce mainly redundant information. Selecting a subset of the functions helps to reduce the cost of data collection for future observations. For this purpose, we propose a

penalized functional generalized linear model, and reduce this model through FPC analysis to a multivariate regression with a grouped Lasso penalty. The grouped Lasso penalty makes the selection of functional predictors feasible.

## 4.2 The Proposed Model

Following the notation in Chapter 3, we consider  $n$  i.i.d. observations, each observation contains  $J$  functions. For  $i = 1, \dots, n$  and  $j = 1, \dots, J$ , let  $x_{ij}(t)$  be the  $j$ th function observed from the  $i$ th observation. Besides  $x_{ij}(t)$ , we also assume a non-functional vector  $s_i$  associated with each observation. Let binary variables  $y_i$  be the responses observed. Our functional generalized linear model is defined as

$$\rho_i = \Pr(y_i = 1 | s_i, x_{ij}(t), j = 1, \dots, J), \text{ and}$$

$$\rho_i = g^{-1}(\eta_i), \quad (4.1)$$

$$\eta_i = \alpha_0 + s_i^T \alpha + \sum_{j=1}^J \int_{T_j} x_{ij}(t) \beta_j(t) dt, \quad (4.2)$$

where  $T_j$  is the domain of  $x_{ij}(t)$ ,  $\alpha_0$  is a univariate intercept,  $\alpha$  is a vector of coefficients for the non-functional predictors, and  $\beta_j(t)$ 's are the functional regression coefficients. Here the link function  $g(\cdot)$  is a one-to-one continuous function. The selection of functional predictors is based on the following constraint on the functional regression coefficients:

$$\sum_{j=1}^J \|\beta_j\|_{L^2} < m, \quad (4.3)$$

where  $\|f\|_{L^2} = (\int f^2(t)dt)^{1/2}$ ,  $m$  is a pre-defined constant. Note that (4.3) is a combined constraint of  $L^2$  norm and  $l^1$  norm. This is an extension of the group-wise variable selection in multivariate setting proposed by Yuan and Lin [82]. Because of the properties of this combined constraint, we expect  $\beta_j \equiv 0$  for some  $j$ , depending on the shrinkage factor  $m$ .

To solve the regression coefficients from the above proposed model, we apply functional approximation using orthonormal basis expansion as done in Chapter 3. The functional predictor  $x_{ij}(t)$  is expanded by an orthonormal basis  $\{\phi_k^j\}_{k=1}^\infty$  (which can be the estimated eigenfunctions if using FPC analysis) as

$$x_{ij}(t) = \sum_{k=1}^{\infty} c_{ijk} \phi_k^j(t). \quad (4.4)$$

We then use a truncated version of (4.4) to approximate  $x_{ij}(t)$ . Note that if using FPC method, the functional predictors  $x_{ij}(t)$  should be centered at their sample mean to satisfy the zero mean assumption of the FPC analysis, and the functions from the test set should be centered using the mean estimated from the training set. The same orthonormal basis is used to expand  $\beta_j(t)$ :

$$\beta_j(t) = \sum_{k=1}^{\infty} b_{jk} \phi_k^j(t) \quad (4.5)$$

Once the coefficients for orthonormal basis or the FPC scores have been estimated, we can approximate equation (4.2) by

$$\eta_i = \alpha_0 + s_i^T \alpha + \sum_{j=1}^J \sum_{k=1}^{p_j} c_{ijk} b_{jk}, \quad (4.6)$$

where  $p_j$  is the truncation parameter for the  $j$ th functional predictor, which can be determined by approximation criterion stated in Section 3.5. The constraint condition (4.3) is then approximated by

$$\sum_{j=1}^J \|b_j\|_2 < m \quad (4.7)$$

where  $b_j = (b_{j1}, \dots, b_{jp_j})$  and  $\|\cdot\|_2$  stands for the Euclidean norm. A regression with constraint in form of (4.7) is called “grouped Lasso” by Yuan and Lin [82]. Functional predictor selection can thus be performed through selecting variables in (4.6) under this constraint, i.e., if one curve  $x_j(t)$  is selected, then the coefficients  $b_{jk}, k = 1, \dots, p_j$ , will all be non-zero.

The grouped Lasso method originates from the Lasso (Least Absolute Shrinkage and Selection Operator), which was first proposed by Tibshirani[72] for model selection in linear regression. The basic idea of Lasso is to find a subset of the predictors with non-zero coefficients by applying a  $l_1$  constraint to the regression coefficients based on the ordinary least square estimation. Yuan and Lin [82] extended the regular Lasso to the case where the predictors can be grouped, such as multi-factor ANOVA. They combine the  $l_1$  and  $l_2$  constraints so that the resulting model selects variables at the group level and is invariant under group-wise orthogonal transformation. To solve our problem based on the approximated model (4.6) and (4.7), we borrow the algorithm proposed by Meier et al. [43], where they extended the group-wise lasso regression of Yuan and Lin [82] to a logistic regression setup. Suppose the link function

in (4.1) is a logit link, i.e.,

$$\log\left(\frac{\rho_i}{1 - \rho_i}\right) = \eta_i, \quad (4.8)$$

the estimate can be obtained by minimizing the convex function

$$Q_\lambda(\theta) = -l(\theta) + \lambda \sum_{j=1}^J s(p_j) \|b_j\|_2, \quad (4.9)$$

where  $\theta = \{\alpha_0, \alpha, b_j, j = 1, \dots, J\}$ , and  $l(\cdot)$  is the log-likelihood function

$$l(\theta) = \sum_{i=1}^n \{y_i \eta_i - \log(1 + \exp(\eta_i))\}. \quad (4.10)$$

Here  $s(p_j)$  is a rescaling parameter which adjusts for the penalty according to the dimensionality of  $b_j$ , and is usually set to be  $\sqrt{p_j}$ ;  $\lambda > 0$  is a tuning parameter controlling the amount of penalty. Note that in the model of Meier et al. [43], only one term, the intercept term, is unpenalized. However, in our proposed model, in addition to the intercept  $\alpha_0$ , we also allow the coefficients of nonfunctional predictors,  $\alpha$ , to be unpenalized. Meier et al. stated the attainability of the minimum and provided a proof. Actually, the attainability holds only when some conditions are satisfied. Here we provide a general sufficient condition under which the minimum of (4.9) is attained.

**Proposition 4.2.1.** *Suppose that  $0 < \sum_{i=1}^n y_i < n$ ,  $\lambda > 0$ ,  $s(p_j) > 0, \forall j$ , and the design matrix*

$$X = \begin{pmatrix} 1 & z_1^T & c_{111} & \dots & c_{11p_1} & \dots & \dots & c_{1J1} & \dots & c_{1Jp_J} \\ \vdots & & & & & & & & & \\ 1 & z_n^T & c_{n11} & \dots & c_{n1p_1} & \dots & \dots & c_{nJ1} & \dots & c_{nJp_J} \end{pmatrix}$$

is a  $n$  by  $m$  matrix of rank  $m$ ,  $n \geq m$ . If the maximum likelihood estimator for the logistic regression (with log-likelihood in the form of Equation(4.10) exists, then (4.9) has an unique minimizer  $\theta^*$ .

The proof of Proposition 4.2.1 is in Appendix B. Meier et al. [43] proposed a Block Coordinate Gradient Descent algorithm to solve the group lasso logistic regression and provided a R package called *grplasso*. We will use this package to perform functional predictor selection based on the approximated model in Equations (4.6) and (4.7). The initialization of the algorithm is the same as in *grplasso*.

### 4.3 Simulation Study

We use simulation to verify the performance of the proposed method in classification problems with multiple functional predictors. We generate  $n = 1000$  i.i.d. observations, each contains one non-functional predictor and three functional predictors. The non-functional predictor is generated from the Uniform[0, 1] distribution, and the three functional predictors are constructed through cosine basis expansion using the first 4 bases functions  $\phi_0(t) = 1, \phi_k(t) = \sqrt{2} \cos(k\pi t), k = 1, \dots, 3$  on the domain [0, 1]. The cosine basis coefficients of each functional predictor are generated independently from a normal distribution with some fixed mean and variance 0.5. We set the coefficient functions for the first and the third functional predictors to be zero and set the coefficient function for the second to be non-zero. Figure 4.1 shows the plot of both the non-functional predictor and the functional predictors for

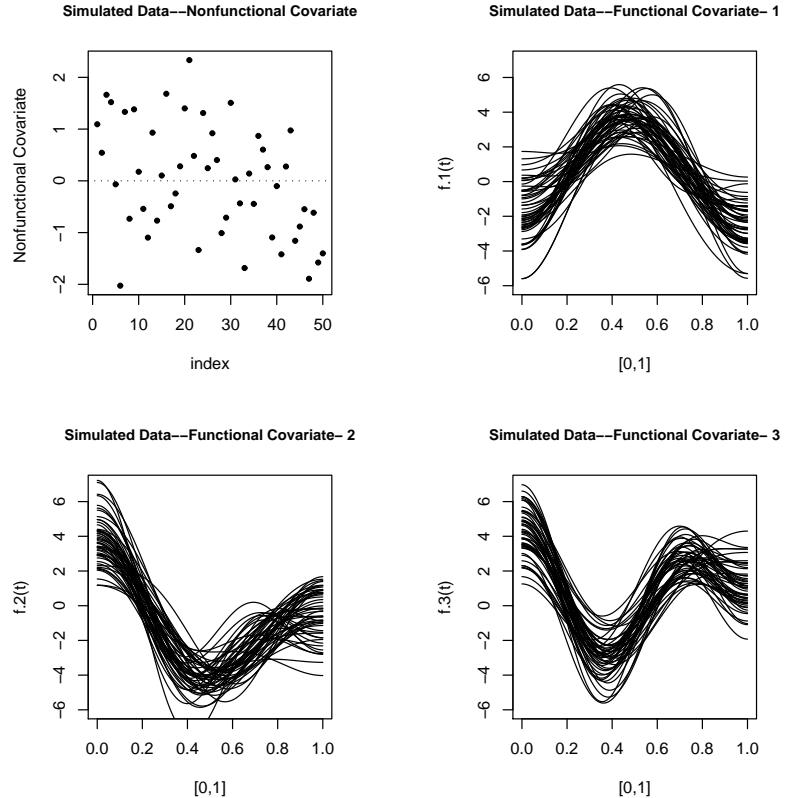


Figure 4.1: Data plot of both non-functional predictors and functional predictors for the first 50 observations used in simulation.

the first 50 observations. The binary responses  $y_i$  are generated by sampling from a Bernoulli distribution with success probability  $\rho_i = (1 + \exp(-\eta_i))^{-1}$ , where  $\eta_i$  is computed from Equation (4.2) using numerical integration. The simulated  $y_i$ 's are well balanced, with 57.3% in the 1 class. We then randomly split the data into a training set of size 800 and a test set of size 200.

Now we apply the proposed model to the simulated data for classification. In the function approximation step, one can choose an orthonormal basis different from the one in data generation. We have tried both functional principal components and cosine basis, and obtained very similar curve selection and prediction results.

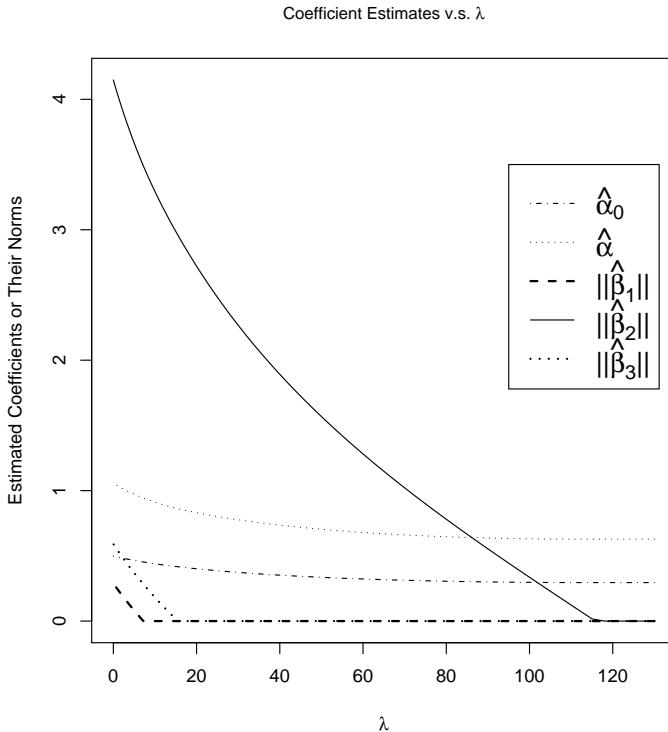


Figure 4.2: Estimated paths of coefficient vector at different  $\lambda$  values

Using function approximation with cosine basis expansion and the approximation criterion stated in Section 3.5 with  $c_2 = 0.99$ , we obtain the truncation parameter  $p_j \equiv 4$ . The group-wise Lasso regression algorithm of Meier et al.[43] is then applied to the reduced scores. Figure 4.2 shows the estimation for the regression coefficients as a function of  $\lambda$ . Note that for the estimated coefficient function  $\hat{\beta}_j$ , we plot their  $L^2$  norm, i.e.,  $||\hat{\beta}_j|| = \sqrt{\int_{T_j} \hat{\beta}_j(t)^2 dt}$ , where the function  $\hat{\beta}_j$  are obtained by the inverse transformation of the estimated coefficients  $\hat{b}_j$ . From Figure 4.2, we see that for a wide range of  $\lambda$ ,  $15.7 < \lambda < 115$ , the model correctly picks out the non-zero coefficient function  $\hat{\beta}_2$ . We also plot  $\hat{\beta}_2(t)$  under 6 selected  $\lambda$ 's in Figure 4.3 to compare with the true  $\beta_2(t)$ . Table 4.1 shows the estimated coefficients (in the form of cosine

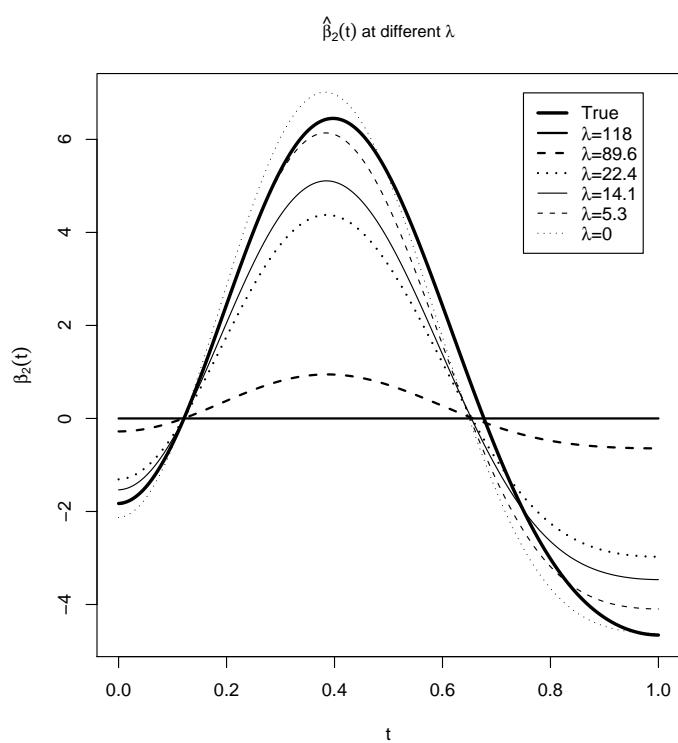


Figure 4.3: Estimated coefficient function  $\hat{\beta}_2(t)$  at 6 selected  $\lambda$  values and the true  $\beta_2(t)$ .

basis scores  $\hat{b}_j$ ) compared with the true values under the 6  $\lambda$ 's. From Table 4.1, we see that as the penalty parameter  $\lambda$  increases, the estimated coefficients shrink toward 0; when  $\lambda = 0$ , the estimates are equal to the maximum likelihood estimates, in which case all the coefficients are nonzero; when  $\lambda$  varies from 22.4 to 89.6, the coefficients of the first and the third curve are exactly 0, and the coefficient of the second is nonzero. For  $\lambda > 14.1$ , almost all the estimates are closer to 0 than their true values. We believe that these shrinkage effects are caused by the continuous-shrinkage property of Ridge and Lasso penalty (see Tibshirani [72]). As a side note, it has been suggested that there may be large bias in the estimators related to the inconsistency of the original Lasso under certain conditions, i.e., that the Lasso does not satisfy the “oracle properties” (Fan and Li[18], Zhao and Yu [83]). Some modifications have been proposed to overcome the drawbacks of Lasso and make the estimators satisfy the oracle properties(see Zou [85]). In this study, we only focus on the functional predictor selection, more research can to be done on the consistency of the grouped-Lasso regression under the functional data setup.

We plug the estimated coefficient function  $\hat{\beta}_j(t), j = 1, 2, 3$  into the test set using (4.2) to perform prediction. For each observation, the estimated success probability  $\hat{p}_i$  is computed, from which we plot a ROC curve for each  $\lambda$ . The optimal classification point is chosen from each ROC curve to maximizes the sum of sensitivity and specificity. Figure 4.4 shows the misclassification rate at the optimal point and the corresponding area under the ROC curves at different values of  $\lambda$ . From Figure 4.4,

Coef	True Values	Estimated coefficients at different $\lambda$ values					
		$\lambda=118$	$\lambda=89.6$	$\lambda=22.4$	$\lambda=14.1$	$\lambda=5.3$	$\lambda=0$
$\alpha_0$	0.5	0.3	0.3	0.39	0.42	0.46	0.5
$\alpha$	1	0.63	0.64	0.82	0.87	0.97	1.06
$b_{11}$	0	0	0	0	0	0.03	0.15
$b_{12}$	0	0	0	0	0	-0.04	-0.17
$b_{13}$	0	0	0	0	0	0.04	0.18
$b_{14}$	0	0	0	0	0	0	-0.01
$b_{21}$	1	0	0.13	0.58	0.67	0.79	0.9
$b_{22}$	2	0	0.31	1.43	1.67	2.01	2.29
$b_{23}$	-3	0	-0.42	-1.92	-2.24	-2.66	-3.02
$b_{24}$	-1	0	-0.18	-0.84	-0.99	-1.21	-1.41
$b_{31}$	0	0	0	0	0	0.02	0.03
$b_{32}$	0	0	0	0	0.01	0.07	0.13
$b_{33}$	0	0	0	0	0.04	0.34	0.56
$b_{34}$	0	0	0	0	0.01	0.09	0.14

Table 4.1: The estimated coefficient values compared with the true values at different  $\lambda$ 's

we find the “best” prediction results with sensitivity(93%), specificity(73%) and an fairly large area under ROC curve (0.88) when  $\lambda$  is around 22.4, and the resulting misclassification rate is 16%.

Since in practice the true basis is unknown, we also use FPC for dimension reduction and compare the results with those from cosine basis. For all the 3 functional predictors, the approximation criterion stated in Section 3.5 with  $c_1 = 0.99$  gives  $p_j \equiv 4$ . Actually, the first 4 principal components take into account 100% of the variability in the training data. Based on the 4 principal components for each curve, we obtain the regression coefficient estimates very similar to those in Figure 4.2, except that the scales of the coefficient norms  $\|\hat{\beta}_j\|$  are different. The prediction results are also very close to those in Figure 4.4. FPC gives the best 93% sensitivity, 73%

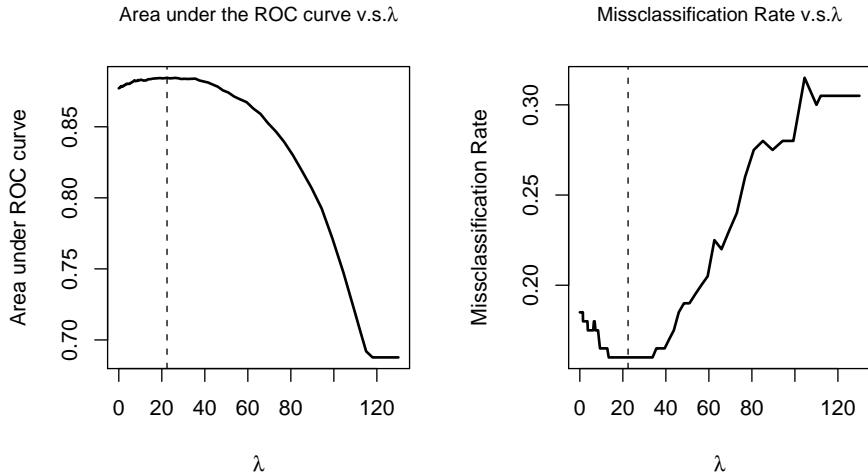


Figure 4.4: Prediction results at different  $\lambda$  values.

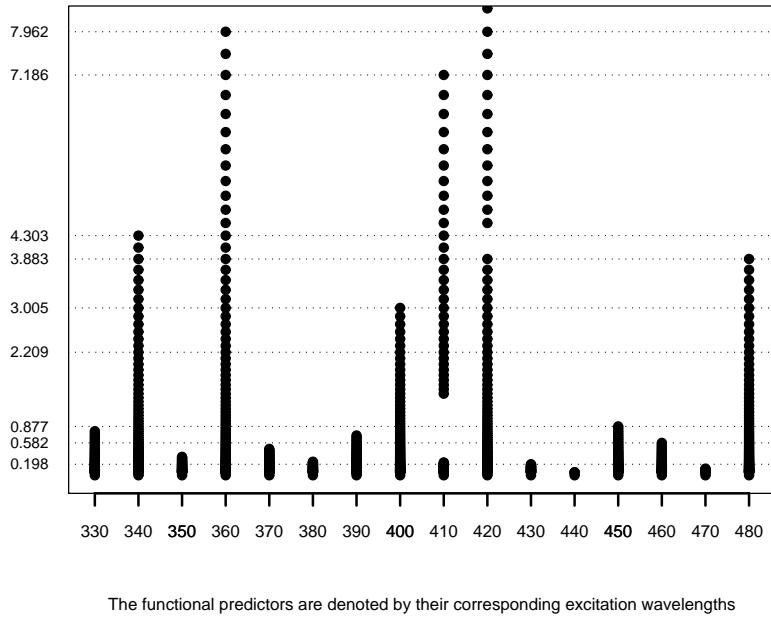
specificity and 0.88 area under ROC curve under  $\lambda = 22.4$ , with a resulting missclassification rate of 16%. Therefore, the FPC method produces exactly the same optimal prediction for the test set as the method of using cosine basis, although they perform dimension reduction in a different way.

## 4.4 Real Data Application

We apply the proposed model to part of the fluorescence data introduced in Section 1.2, which is measured using a fixed instrument (called FastEEM3) at a fixed clinic (British Columbia Cancer Agency, Vancouver, CA). There are 724 EEM measurements made on 311 patients in this dataset. Each measurement contains 16 spectral curves. The measurements are from different sites of the cervix, and there may exist repeated measurements for the same site. We split the data into a training set of size 399 and a test set of size 325, with the proportions of diseased cases 0.21

and 0.20, respectively. Two non-functional covariates are considered in this study. The first one is the colposcopic tissue type of the measurements which is obtained prior to the fluorescence spectroscopy measurements. There are two types of colposcopic tissue – squamous and columnar, which makes this covariate a binary variable. The second one is the menopausal status of patients, which can be categorized into three levels: pre-, peri- and post-menopause. We use FPC to approximate the functional predictors with the approximation criterion  $c_1 = 0.998$ . The resulting  $p_j$ 's vary between 2 and 3, with  $\sum_j p_j = 41$ . To reduce possible bias, the test set scores (the scores of orthonormal basis) are computed based on information from the training set only. For example, the eigenfunctions used for computing the FPC scores of the test set are estimated from the training set.

The group lasso logistic regression algorithm is used to estimate the regression coefficients as  $\lambda$  decreases from 8.5 to 0. Due to the large number of functional predictors, the plot of coefficient estimates is hard to visualize. In Figure 4.5, we summarize the excitation curves (functional predictors) selected at different  $\lambda$  values. The x-axis represents the functional predictors indexed by excitation wavelengths. The y-axis represents the  $\lambda$  values. The black spot indicates that the estimated regression coefficient at the given excitation wavelength is non-zero for the given  $\lambda$  value, therefore the corresponding functional predictor is selected. For example, we find in Figure 4.5 that when  $\lambda = 7.186$ , the curves at excitation wavelengths 360, 410 and 420 are selected. When  $\lambda = 0$ , there is no penalty, hence all the curves are

The selected functional predictors at different  $\lambda$  values, EEM3 DataFigure 4.5: The selected functional predictors (fluorescence spectral curves denoted by excitation wavelengths) at different  $\lambda$  values.

selected. As  $\lambda$  gets larger, this model puts more penalty on the functional regression coefficients, therefore selects fewer curves. At each given  $\lambda$  value, we can get a set of estimated coefficients, which can be used to do prediction on the test set. We thus determine  $\lambda$  by comparing their prediction performance on the test set. Due to the fact that the total proportion of diseased cases is small, the misclassification rate is not a good criterion for evaluating the prediction performance (see [84], page 22 for details). In order to reduce the risk of false negatives, we wish to keep a high sensitivity. It turns out that in such rare-disease diagnosis problems, using the criterion that the sum of sensitivity and specificity is maximized will help to remain a high enough sensitivity. Hence for each fixed  $\lambda$ , we pick a point from the

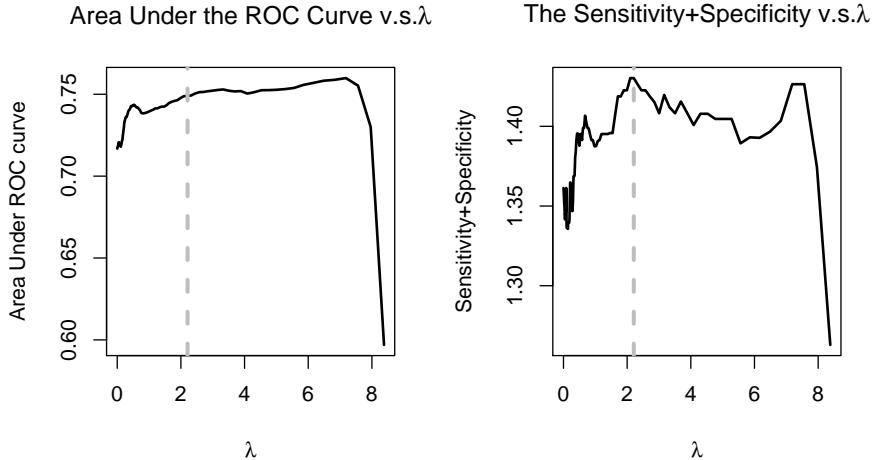


Figure 4.6: Prediction results at different  $\lambda$  values.

empirical ROC curve such that the sum of the sensitivity and specificity is maximized.

Figure 4.6 shows the area under ROC curve and the optimal sum of the sensitivity and specificity at different values of  $\lambda$ . When  $\lambda = 2.209$ , the sum reaches its maximum 1.43, with sensitivity 86% and specificity 57%. The corresponding area under ROC curve is 0.75, and the misclassification rate is 37%. As shown in Figure 4.5, when  $\lambda = 2.209$ , there are six functional predictors selected at excitations 340, 360, 400, 410, 420 and 480 nm. These selected excitation wavelengths can be used in the future for building more cost-effective devices. In Table 4.2, we compare the prediction results using the proposed model at  $\lambda = 2.209$  with the results from 3 other classification methods. The corresponding empirical ROC curves are plotted in Figure 4.7. Note that the parameter  $k$  used in the k-nearest neighbor method is determined by a 15-fold cross validation based on the training set. Both Table 4.2 and Figure 4.7 show that the 4 classification methods provide similar prediction results on the test set,

<i>Method</i>	Auc	MisR	Sens	Speci	Thresh	Sum
FGLM( $\lambda = 2.209$ )	0.75	37%	86%	57%	0.16	1.43
Logistic	0.72	43%	88%	50%	0.12	1.37
KNN	0.73	33%	78%	64%	0.23	1.42
LDA	0.74	40%	84%	54%	0.19	1.38

Table 4.2: The classification results using 4 different methods. Auc: Area under ROC curve. MisR: Misclassification rate. Sens: Sensitivity. Speci: Specificity. Thresh: The threshold used for sensitivity and specificity. Sum: The sum of sensitivity and specificity. FGLM: The proposed model at  $\lambda = 2.209$ . Logistic: logistic regression. KNN: k-nearest neighbor. LDA: linear discriminant analysis.

in the sense that their AUC's are all at the 0.70 level. Comparing with the other 3 methods, our proposed model (denoted as (FGLM)) does not improve the AUC too much. However, since the main purpose of this model is functional predictor selection rather than classification, we have gained benefits by doing inferences on functional predictor selection without losing classification power.

## 4.5 Discussion

We have proposed a functional logistic regression model to perform classification and functional predictor selection. Using the grouped Lasso penalty, the proposed model gives information on which functional predictor will be selected if we are willing to use a subset of the functional predictors for classification. For example, under penalty  $\lambda = 2.209$ , the best six functional predictors selected in our real data application are curves at excitation wavelengths 340, 360, 400, 410, 420 and 480nm. The selected functional predictors can be further used by different classifiers for new measurements.

In our proposed model, the tuning parameter  $\lambda$  is important for prediction. In

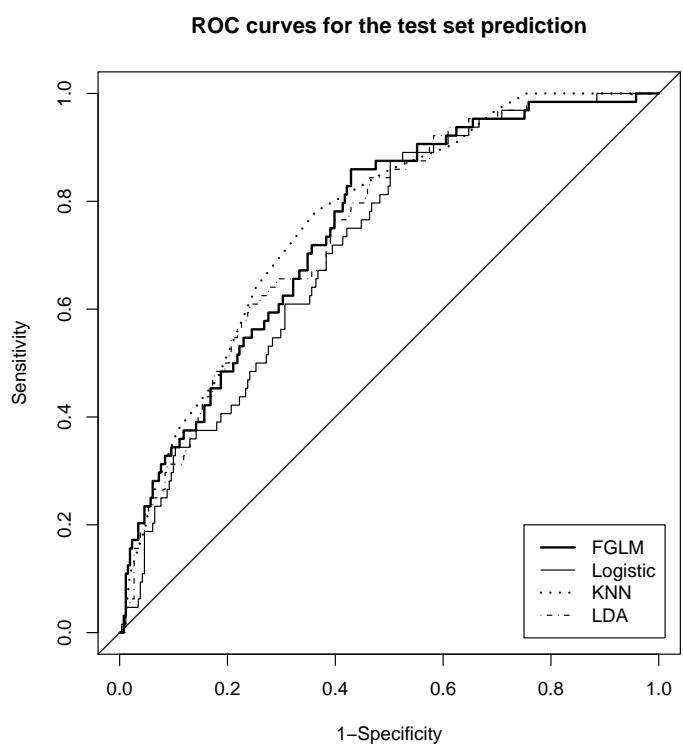


Figure 4.7: ROC curves obtained when training using 4 different classifiers and predicting on the test set.

Meier et al. [43] and in our study in this chapter, a test set is used to choose  $\lambda$  with the best prediction performance. However, in some cases there are only a small number of observations available and splitting out a test set is not possible. In such cases, we can adopt some model selection criteria such as AIC, BIC or practical  $C_p$ . AIC tends to select a model with optimal prediction, whereas BIC tends to identify the true sparse model if the true model is included in the candidate set (see Yang[77]). In the grouped Lasso linear regression model, Yuan and Lin [82] propose an approximation to the degree of freedom and use a  $C_p$  criterion to select the tuning parameter  $\lambda$ . It remains an open question whether this criterion can be extended to the logistic regression case for selecting  $\lambda$ .

There are several aspects need to be studied in the future. First, it is necessary to investigate the consistency properties of the estimated coefficient function  $\beta_j(t)$ , such as the oracle property. Second, in the group Lasso algorithm, Meier et al. [43] propose a way to find the range of the tuning parameter  $\lambda$ , and  $\lambda$  can only vary on this pre-specified grids within this range. This method, although fast, makes it difficult to find the precise  $\lambda$  value that is optimal for prediction purpose. Efficient algorithms for searching for  $\lambda$  are necessary especially when functional data are involved.

# Chapter 5

## A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors

The penalized functional generalized linear model proposed in Chapter 4 provides inferences on selecting functional predictors. However, in our real data application, there is another issue that is not considered by this model, the random batch effects. In order to perform functional predictor selection and take the random batch effects into consideration, in this chapter we extend the Bayesian Probit Model in Chapter 3 to a Bayesian hierarchical model with functional predictor selection (BHFPS). The Bayesian hierarchical structure takes into account the random batch effects, and the functional predictor selection is implemented through a block-wise variable selection

method. Fixed effects or predictors in non-functional form are also included in this model. As we have done in previous chapters, the dimension of the functional data is reduced through functional principal component analysis or orthonormal basis expansion. We use a hybrid Metropolis-Hastings/Gibbs sampler for posterior sampling and apply an Evolutionary Monte Carlo (EMC) algorithm to improve the mixing. Simulation and real data application show that the proposed BHFPS model provides accurate selection of functional predictors as well as good classification.

## 5.1 Motivation

In practical problems of functional data classification, there are often practical issues that are handled by the models proposed in Chapter 3 and Chapter 4. One of them is the presence of systematic effects which may be significant enough to bias classification, such as the artificial differences caused by measuring with different devices.

In Example 5.1.1, we use a toy example to show how the device difference misleads the classification in an unbalanced design. A similar issue is addressed in Baggerly et al. (2004).

**Example 5.1.1.** *The following table lists the counts of the objects measured by two devices for a binary classification problem. If we use the device difference to do prediction, for example, we classify all the objects measured by device one to class one, the misclassification rate is  $(5 + 50)/365 = 15\%$ , which seems quite good but is obviously useless since the device difference is purely artificial. Unfortunately, most*

*classification algorithms can hardly recognize the sources of variation and may end up with discriminating the objects based on the device difference. We call the variations caused by device or other experimental difference as “batch effects”.*

True class	Device one	Device two
Class one	300	50
Class two	5	10

In our application of fluorescence spectroscopy data introduced in Section 1.2, several factors that are brought in by the experimental design need to be considered. First, the data are obtained using two instruments with four optical probes located at three clinics. A preliminary study shows that there exists significant differences among the data from different device-clinic combinations, which puts the classification at risk since the diseased cases are rare and distributed inhomogeneously across these combinations, like the example shown in Example 5.1.1. Second, in addition to device-clinic differences, it is believed that other factors, such as the tissue type of the measurement site and the patients’ menopausal status, may confound with the fluorescence spectroscopy information in the diagnosis. These factor effects are shown by box-plots in Figure 5.1.

This motivates us to propose a Bayesian hierarchical model with selection of functional predictors for complex functional data classification problems, where multiple functional predictors are influenced by random batch effects and fixed effects.

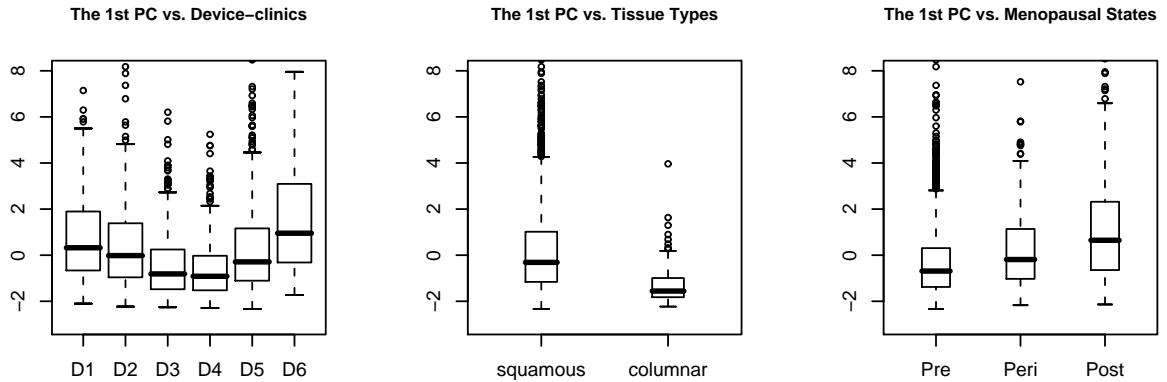


Figure 5.1: The box-plot of the first functional principle component scores of one spectral curve (measured at excitation 340 nm) versus six device-clinic combinations (left), two tissue types (middle) and three menopausal states (right). Systematic differences across different levels of these factors can be seen obviously. Note that here we only used observations from the normal class, which excludes the possibility that the differences are caused by unbalanced proportions of diseased cases in each level of the factors.

## 5.2 Bayesian Hierarchical Model with Selection of Functional Predictors

### 5.2.1 The Proposed Model

Suppose that we obtain functional observations from  $L$  exchangeable batches, in which the  $l$ th batch contains  $n_l$  observations and each observation contains  $J$  functions. For  $l = 1, \dots, L$ ,  $i = 1, \dots, n_l$  and  $j = 1, \dots, J$ , let  $x_{ij}^l(t)$  be the  $j$ th function observed from the  $i$ th observation in batch  $l$ , which takes values in  $L^2[T_j]$ , with  $T_j$  the compact domain of  $x_{ij}^l(t)$ . In addition to the functional observations, there are also non-functional observations  $s_i^l$ , which is assumed to be a vector of length  $q$ . We treat the

observations  $\{s_i^l, x_{ij}^l(t), j = 1, \dots, J\}$  as predictors and assume the binary responses  $y_i^l$  to be conditionally independent given the predictors. Similarly as in model (3.1) in Chapter 3, we introduce univariate latent variables  $z_i^l$  which link the responses  $y_i^l$  to the predictors as follows:

$$y_i^l = \begin{cases} 1 & \text{if } z_i^l < 0, \\ 0 & \text{if } z_i^l \geq 0. \end{cases}$$

$$z_i^l = (s_i^l)^T \alpha + \sum_{j=1}^J \int_{T_j} x_{ij}^l(t) \beta_j^l(t) dt + \epsilon_i^l. \quad (5.1)$$

Here we set the first component of  $s_i^l$  to be 1 to include the intercept term. For all  $i$  and  $l$ , we assume  $\epsilon_i^l$  to be i.i.d. with distribution  $N(0, 1)$ , and assume that  $\beta_j^l(t) \in L^2[T_j]$  for all  $j$ . See Albert and Chib [2] for the use of latent variables in the analysis of binary response data.

In many cases, some functional predictors do not contribute to the classification, and selecting a subset of them may actually improve the classification accuracy. In our application of fluorescence spectroscopy data, there are also economic reasons for using a subset of the  $J$  functional predictors. To this end, we introduce a hyper-parameter  $\tau$  to the priors of  $\beta_j^l(t)$ , where  $\tau = (\tau_1, \dots, \tau_J)$  and each component takes value either 1 or 0, indicating whether or not the corresponding functional predictor is selected. Note that this  $\tau$  parameter is different from the  $\tau$  used in the model of Chapter 3 in that each component determines whether the whole functional predictor is selected or not, as we will show in the following text. The proposed priors for  $\alpha$

and  $\beta_j^l(t)$  are:

$$\begin{aligned} \alpha &\sim N(0, \sigma_1^2 I_q), \\ \beta_j^l(t) \mid \beta_j^0(t), \tau_j, \sigma_b^2 &\sim GP(\beta_j^0, \sigma_b^2 \gamma_{\tau_j}), \\ \beta_j^0(t) \mid \tau_j &\sim GP(0, \sigma_0^2 \gamma_{\tau_j}), \\ \tau_j \mid \omega_j &\sim \text{Bernoulli}(\omega_j), \\ \sigma_b^2 \mid d_1, d_2 &\sim \text{Inv-gamma}(d_1, d_2), \end{aligned} \tag{5.2}$$

where  $\sigma_1^2$ ,  $\sigma_0^2$ ,  $d_1$ ,  $d_2$ ,  $\omega_j$  are pre-specified prior parameters.  $GP(\mu, \gamma)$  represents a Gaussian process with mean  $\mu(t)$  and covariance function  $\gamma(s, t)$ . We let  $\gamma_{\tau_j}$  depend on  $\tau_j$  by

$$\gamma_{\tau_j}(s, t) = [\nu_1^2 \tau_j + \nu_0^2 (1 - \tau_j)] \sum_{k=1}^{\infty} w_k^j \phi_k^j(s) \phi_k^j(t), \tag{5.3}$$

where  $\{\phi_k^j\}_{k=1}^{\infty}$  is a complete orthonormal basis of  $L^2[T_j]$ . Note that the infinite sum in Equation (5.3) is a perfectly general form for a covariance function; it is simply the spectral representation of a covariance function (Ash and Gardner [3]). We will treat  $\{\phi_k^j\}_{k=1}^{\infty}$  and  $\{w_k^j\}_{k=1}^{\infty}$  as prior parameters and make specific choices of them. In Equation (5.3), we let  $\nu_1 \gg \nu_0 > 0$  and set  $\nu_0$  to be close to 0. Under this setting, both  $\beta_j^l(t)$  and  $\beta_j^0(t)$  have covariance functions close to 0 when  $\tau_j = 0$  (i.e., the  $j$ th functional predictor is not selected), and have relatively large variances when  $\tau_j = 1$  (i.e., the  $j$ th functional predictor is selected). This type of prior is motivated by George and McCulloch ([21], [22]) where they use mixture-normal priors for variable selection. The  $w_k^j$ 's in Equation (5.3) are pre-specified positive weight parameters subject to  $\sum_{k=1}^{\infty} w_k^j < \infty$  for all  $j$ 's. We determine  $w_k^j$  using the way suggested

in Section 3.5 in Chapter 3. For simplicity, we assume that the priors of  $\beta_j^l(t)$  are independent for all  $j$  and  $l$ , and priors of  $\tau_j$  are independent for all  $j$ . In order to do practical posterior inference, we construct finite dimensional approximations to the functional predictors and coefficients. This is described in detail in Section 5.2.2 below.

### 5.2.2 The Posterior Inference

From Equation (5.1) and the standard normal assumption of  $\epsilon_i^l$ , it is easy to see that the conditional distribution of  $z_i^l$  given  $y_i^l$ ,  $\alpha$  and  $\beta_j^l(t)$  is a truncated normal:

$$z_i^l | y_i^l, \alpha, \beta_j^l(t) \sim TN(\mu_z, 1) \{ I_{\{z_i^l < 0\}} I_{\{y_i^l = 1\}} + I_{\{z_i^l \geq 0\}} I_{\{y_i^l = 0\}} \}, \quad (5.4)$$

where  $\mu_z = (s_i^l)^T \alpha + \sum_{j=1}^J \int_{T_j} x_{ij}^l(t) \beta_j^l(t) dt$ . Since  $\{\phi_k^j\}_{k=1}^\infty$  is a complete orthonormal basis of  $L^2[T_j]$ , similar to (3.4) in Chapter 3, we can expand  $x_{ij}^l(t)$  and  $\beta_j^l(t)$  by

$$x_{ij}^l(t) = \sum_{k=1}^{\infty} c_{ijk}^l \phi_k^j(t), \quad \beta_j^l(t) = \sum_{k=1}^{\infty} b_{jk}^l \phi_k^j(t), \quad (5.5)$$

and use the truncated version of (5.5) to approximate them. If assuming that  $x_{ij}^l(t)$  has zero mean and  $\int_{T_j} E[x_{ij}^l(t)^2] dt < \infty$ , we can estimate eigenfunctions using functional principal component analysis and treat them as the orthonormal basis. The resulting coefficients  $\{c_{ijk}^l\}_{k=1}^\infty$  are the functional principal component (FPC) scores of  $x_{ij}^l(t)$ . These steps are similar to what we have done in Section 3.4 of Chapter 3.

Based on the estimated orthonormal basis coefficients or the FPC scores, we can

reduce (5.1) by applying the truncated approximations in (5.5), which gives

$$z_i^l = (s_i^l)^T \alpha + \sum_{j=1}^J \sum_{k=1}^{p_j} c_{ijk}^l b_{jk}^l + \epsilon_i^l, \quad (5.6)$$

where  $p_j$  is the truncation parameter for the  $j$ th functional predictor. We propose to determine  $p_j$ 's by setting a function approximation criterion as suggested in Section 3.5. The notation of Equation (5.6) can be simplified by concatenating coefficients of the  $J$  functions to make one vector  $b^l$ . The simplified form of Equation (5.6) is:

$$Z_l = S_l \alpha + C_l b_l + \epsilon_l, \quad (5.7)$$

where  $Z_l = (z_1^l, \dots, z_{n_l}^l)^T$  and  $\epsilon_l = (\epsilon_1^l, \dots, \epsilon_{n_l}^l)^T$ . Here  $S_l$  is a matrix of size  $n_l \times q$  with the  $i$ th row equals  $(s_i^l)^T$ , and  $C_l$  is a matrix of size  $n_l \times p$  ( $p = \sum_{j=1}^J p_j$ ) with the  $i$ th row equals

$$(c_{i11}^l, \dots, c_{i1p_1}^l, c_{i21}^l, \dots, c_{i2p_2}^l, \dots, c_{iJ1}^l, \dots, c_{iJp_J}^l)^T,$$

$i = 1, \dots, n_l$ . Similarly,  $b_l = (b_{11}^l, \dots, b_{1p_1}^l, b_{21}^l, \dots, b_{2p_2}^l, \dots, b_{J1}^l, \dots, b_{Jp_J}^l)^T$ . Based on (5.7), the conditional distribution of the latent variables in (5.4) becomes

$$Z_l | \alpha, b_l, Y_l \sim TN(S_l \alpha + C_l b_l, I_{n_l}) \prod_{i=1}^{n_l} (I_{\{z_i^l < 0\}} I_{\{y_i^l = 1\}} + I_{\{z_i^l \geq 0\}} I_{\{y_i^l = 0\}}), \quad (5.8)$$

where  $Y_l = (y_1^l, \dots, y_{n_l}^l)$ . The truncated orthonormal basis expansion or FPC analysis also reduces the Gaussian process priors for  $\beta_j^l(t)$  and  $\beta_j^0(t)$  to the following multivariate normal priors

$$\begin{aligned} b_l | b_0, \sigma_b^2, \tau &\sim N(b_0, \sigma_b^2 \Sigma_\tau), \\ b_0 | \tau &\sim N(0, \sigma_0^2 \Sigma_\tau), \end{aligned} \quad (5.9)$$

where  $\Sigma_\tau = D_\tau W^{1/2} R W^{1/2} D_\tau$ . Here  $R$  is the prior correlation matrix of  $b_l$  and  $b_0$ .

By the assumption in Section 5.2.1 that  $\beta_j^l(t)$ 's are independent for all  $j$ 's,  $R = I_p$ , an identity matrix.  $W$  is also a diagonal matrix of size  $p$ , with positive diagonal components  $(w_1^1, \dots, w_{p_1}^1, \dots, w_1^J, \dots, w_{p_J}^J)$ . In other words, the diagonal of  $W$  concatenates the first  $p_j$  components of the weight sequence  $\{w_k^j\}_{k=1}^\infty, j = 1, \dots, J$ .  $D_\tau$  is another diagonal matrix with diagonal components

$$(u_1^1, \dots, u_{p_1}^1, \dots, u_1^J, \dots, u_{p_J}^J),$$

where  $u_k^j = \nu_1 \tau_j + \nu_0 (1 - \tau_j)$ , for all  $k = 1, \dots, p_j$ ,  $j = 1, \dots, J$ . Note that  $u_k^j$  does not depends on  $k$ .

With the conditional distribution (5.8), the priors for  $\alpha$ ,  $\tau$  and  $\sigma_b^2$  in (5.2), and the reduced multivariate priors for  $b_l$  and  $b_0$  in (5.9), we get the joint conditional posterior distribution of  $\alpha$ ,  $b_l$ 's,  $b_0$ ,  $\sigma_b^2$ ,  $\tau$  given  $Z_l$ 's and  $Y_l$ 's by

$$\begin{aligned} & \pi(\alpha, b_1, \dots, b_L, b_0, \sigma_b^2, \tau | Z_l, Y_l, l = 1, \dots, L) \\ & \propto \left[ \prod_l \pi(Z_l | \alpha, b_l, b_0, \sigma_b^2, \tau, Y_l) \pi(b_l | b_0, \sigma_b^2, \tau) \right] \pi(b_0 | \tau) \pi(\alpha) \pi(\tau) \pi(\sigma_b^2). \end{aligned} \tag{5.10}$$

The parameters  $\alpha$ ,  $b_l$ 's and  $b_0$  can all be integrated out sequentially from (5.10), which gives the marginal conditional posterior density

$$\pi(\sigma_b^2, \tau | Z_l, Y_l, l = 1, \dots, L). \tag{5.11}$$

See Appendix A for details of the integration. Based on (5.8), (5.10) and (5.11), we design MCMC algorithms to obtain posterior samples of the parameters. The

posterior samples of  $b_l$ 's can then be used to estimate  $\beta_j^l(t)$ 's. For new observations, we use the estimated  $\beta_j^l(t)$ 's and the posterior samples of  $\alpha$  for prediction.

### 5.3 Markov Chain Monte Carlo

Based on the model constructed in Section 5.2, we propose two MCMC algorithms for posterior sampling. The first one is a hybrid Metropolis-Hastings/Gibbs sampler, and the second one is a modified version of algorithm 1 which uses the EMC algorithm to improve the mixing when the number of functional predictors is relatively large.

#### 5.3.1 Algorithm 1

(A Hybrid Metropolis-Hastings/Gibbs sampler)

Step 0. Set initial values for  $b_l$ 's,  $\alpha$ ,  $\tau$  and  $\sigma_b^2$ .

Step 1. For  $l = 1, \dots, L$ , conditional on  $Y_l$ , and current values of  $b_l$  and  $\alpha$ , update  $Z_l$  from the truncated normal distribution described in Equation (5.8) of Section 5.2.2.

Step 2. Update  $\sigma_b^2$  based on  $\pi(\sigma_b^2 | \tau, Z_l, Y_l, l = 1, \dots, L)$ . Sample a proposal  $\tilde{\sigma}_b^2$  by  $\log \tilde{\sigma}_b^2 = \log \sigma_b^2 + \epsilon$ , with  $\epsilon \sim N(0, \delta^2)$ .  $\delta$  is an adjustable step size. Compute the ratio

$$R_\sigma = \frac{\pi(\tilde{\sigma}_b^2 | \tau, Z_l, Y_l, l = 1, \dots, L) \tilde{\sigma}_b^2}{\pi(\sigma_b^2 | \tau, Z_l, Y_l, l = 1, \dots, L) \sigma_b^2}$$

and update  $\sigma_b^2 = \tilde{\sigma}_b^2$  with probability  $\min(1, R_\sigma)$ .

Step 3. Update  $\tau$  based on  $\pi(\tau|\sigma_b^2, Z_l, Y_l, l = 1, \dots, L)$ . Generate a proposal  $\tilde{\tau}$  by “switch/swap”, i.e., with probability  $\xi$ , randomly swap one 1 term with one 0 term; and with probability  $1 - \xi$ , randomly pick one position and switch it.

Then let

$$R_\tau = \frac{\pi(\tilde{\tau}|\sigma_b^2, Z_l, l = 1, \dots, L)}{\pi(\tau|\sigma_b^2, Z_l, l = 1, \dots, L)}$$

and update  $\tau = \tilde{\tau}$  with probability  $\min(1, R_\tau)$ .

Step 4. Update  $\alpha$  conditional on current values of  $\sigma_b^2$ ,  $\tau$  and  $Z_l$  through the conditional distribution  $\alpha|\sigma_b^2, \tau, Z_l \sim N(\mu_\alpha, V_\alpha)$ , where  $\mu_\alpha$  and  $V_\alpha$  are defined in Web Appendix B.

Step 5. Conditional on current values of  $\alpha$ ,  $\sigma_b^2$ ,  $\tau$ ,  $Z_l$ , update  $b_0$  by  $b_0|\alpha, \sigma_b^2, \tau, Z_l \sim N(\mu_0, V_0)$  where  $\mu_0$  and  $V_0$  are defined in Web Appendix B.

Step 6. Conditional on current values of  $b_0$ ,  $\alpha$ ,  $\sigma_b^2$ ,  $\tau$  and  $Z_l$ , update  $b_l$ ,  $l = 1, \dots, L$  by  $b_l|b_0, \alpha, \sigma_b^2, \tau, Z_l \sim N(\mu_l, V_l)$  where  $\mu_l$  and  $V_l$  are defined in Web Appendix B.

Repeat Step 1 – 6 until convergence.

In Appendix C, we verify that MCMC algorithm 1 converges to a unique equilibrium distribution, which is our posterior distribution defined in Section 5.2. The “switch/swap” proposal used in Step 6 is similar to the methods used in Brown et al. ([8], [9]). Our simulation shows that if the number of functional predictors is small, this type of proposal can locate the correct value of  $\tau$  within a few iterations. However, when the number of functional predictors is large, the size of the searching

space for  $\tau$  increases at an exponential rate. The “switch/swap” proposal can hardly find successful proposals because of the discrete nature of the large state space, thus results in extremely low acceptance rate (e.g., acceptance rate less than 0.1%).

In order to obtain better mixing for  $\tau$ , we construct a more effective EMC algorithm based on algorithm 1. The EMC algorithm is a MCMC scheme that inherits the attractive features from both simulated annealing and genetic algorithm. It simulates a population of  $I$  Markov chains in parallel, each with a different “temperature”. The temperatures are ordered decreasingly to form a “ladder”. For each chain, the posterior is transformed according to its temperature. Denote the target posterior distribution as  $\pi(\theta)$  and the temperature for the  $i$ th chain as  $t_i$ , the transformed posterior for the  $i$ th chain is  $\pi_i(\theta) \propto \pi(\theta)^{1/t_i}$ . Depending on  $t_i$ , such a transformation makes the unnormalized target posterior density more flat or more spiky. The EMC algorithm improves the Metropolis-Hastings updates by introducing three operations: mutation, crossover and exchange. These operations allow both independent updates for each chain and interactions between neighboring chains. We introduce more details of the EMC algorithm in Appendix D. More information about EMC can be found in Liang and Wong [39], Liu [40], Goswami and Liu [24], and Bottolo and Richardson [7].

When using the EMC algorithm, there are several crucial parameters need to be determined: the number of chains  $I$ , temperature of each chain and the maximum temperature. We adopt a simple method suggested by Bottolo and Richardson (2008)

to set temperature for each chain, which uses a geometric sequence and adjusts the common ratio in a burn-in period so that the acceptance rate for the exchange operation is close to 50%. For the number of chains and the maximum temperature, we suggest to choose the number of chains to be around  $J/2$ , and choose the maximum temperature between 10 and  $10^3$  according to experience. The algorithm stated below gives details of the EMC algorithm for our proposed model. In this algorithm, we borrow the idea of Bottolo and Richardson [7], where they update the main parameter of interest (the  $\gamma$  parameter in their setup) using EMC with multiple chains, and update the nuisance parameter (the  $\tau$  parameter in their setup) conditional on the main parameter obtained from the chain with temperature 1.

### 5.3.2 Algorithm 2 (EMC)

Step 0. Set initial values for  $b_l$ 's,  $\alpha$ ,  $\tau$  and  $\sigma_b^2$ . And set up an initial temperature ladder:  $t_1 > t_2 > \dots > t_I > 0$  with the initial ratio of the geometric sequence  $a = t_{i+1}/t_i, i = 1, \dots, I$ . We adjust the temperature ladder so that  $t_1$  is bounded by the maximum temperature and set one temperature to be exactly

1. Let the step-size for adjusting temperature be  $\delta_a = \log_2(a)/\tilde{n}$ , where  $\tilde{n}$  is the ratio of the burn-in period to a block size (usually 100). Set value for parameter  $q$ , the probability of mutation and crossover, and for  $\xi$ , the probability of switch and swap within the mutation step.

Step 1. Run step 1 – 2 in algorithm 1 based on the chain with temperature equals 1,

obtain samples of  $Z_l$ 's and  $\sigma_b^2$ . These steps should be identical with those in algorithm 1 since temperature value 1 does not modify the posterior density.

Step 2. Conditional on current values of  $Z_l$ 's and  $\sigma_b^2$ , update  $\tau$  according to the following steps in 2.1 and 2.2. For convenience, here we denote  $\pi(\tau|\sigma_b^2, Z_l, Y_l, l = 1, \dots, L)$  as  $\pi(\tau|\cdot)$ .

Step 2.1. (mutation/crossover) With probability  $q$ , perform a mutation step independently for each chain, i.e. “switch” or “swap” with probability  $\xi$ , as described in step 3 of algorithm 1. Denote the mutated value as  $\tilde{\tau}$  and compute the log ratio  $\log r_m = [\log \pi(\tilde{\tau}|\cdot) - \log \pi(\tau|\cdot)]/t$ , where  $t$  is the temperature of the chain. Update  $\tau = \tilde{\tau}$  with probability  $\min(1, r_m)$ .

With probability  $1 - q$ , perform a crossover step  $[I/2]$  times, where  $[I/2]$  denotes the integer part of  $I/2$ . The crossover is conducted as follows: selecting a pair of chains  $(i, j)$  according to some selection rules (see Liu (2001)), and exchange the right segment of the two  $\tau$ 's from a random point. Denote the old values as  $(\tau^i, \tau^j)$ , and the crossed values as  $(\tilde{\tau}^i, \tilde{\tau}^j)$ , we then compute the log ratio:

$$\log r_c = \frac{\log \pi(\tilde{\tau}^i|\cdot) - \log \pi(\tau^i|\cdot)}{t_i} + \frac{\log \pi(\tilde{\tau}^j|\cdot) - \log \pi(\tau^j|\cdot)}{t_j} + \log \frac{T((\tau^i, \tau^j)|(\tilde{\tau}^i, \tilde{\tau}^j))}{T((\tilde{\tau}^i, \tilde{\tau}^j)|(\tau^i, \tau^j))}$$

where  $T(x|y)$  is the transition probability from  $y$  to  $x$ .  $(\tilde{\tau}^i, \tilde{\tau}^j)$  are accepted with probability  $\min(1, r_c)$ .

Step 2.2. (exchange) Exchange  $\tau$  values from two adjacent chains  $I$  times, i.e.,

randomly choose  $\tau^i$  and  $\tau^j$  from neighboring chains, and compute the log ratio:

$$\log r_e = [\log \pi(\tau^j | \cdot) - \log \pi(\tau^i | \cdot)] \left( \frac{t_j - t_i}{t_i t_j} \right)$$

exchange  $\tau^i$  with  $\tau^j$  with probability  $\min(1, r_e)$ .

Step 3. Conditional on current values of  $Z_l$ 's,  $\sigma_b^2$ , and current sample of  $\tau$  from the chain with temperature 1, run Step 4 – 6 of algorithm 1 based on the chain with temperature equals 1, obtain samples of  $\alpha$ ,  $b_0$  and  $b$ . This step should be identical with Step 4 – 6 in algorithm 1.

Step 4. For every block of iterations within the burn-in period, we adjust the temperature ladder according to the acceptance rate of the exchange operations within this block. A new geometric ratio  $\tilde{a}$  is computed by  $\log_2 \tilde{a} = \log_2 a \pm \delta_a$ , where the “+” sign is used when we would like to reduce the acceptance rate of exchange. The new temperature ladder then is applied to the next block of iterations.

Repeat Step 1 – 4 until convergence.

The above algorithm is an extension of algorithm 1. We have applied the EMC algorithm to the step of updating  $\tau$ , while keeping the update of all other parameters the same as in algorithm 1, similar to the algorithm in Bottolo and Richardson [7]. As shown in simulation 2 and real data application, this algorithm seems work well. However, by now we haven't been able to figure out what the target posterior

distribution looks like under this algorithm setup, and we haven't been able to prove that the target distribution associated with this algorithm will result in a stationary distribution for the whole chain. The proof of the convergence remains an open problem.

## 5.4 Setting Parameters

In Section 5.2.1 and Section 5.2.2, we suggest to determine the truncation parameters  $p_j$  and the weights  $\{w_k^j\}_{k=1}^\infty$  using the method in Section 3.5. Besides  $p_j$  and  $\{w_k^j\}_{k=1}^\infty$ , there are several other priors need to be set, including  $\sigma_1^2$ ,  $\sigma_0^2$ ,  $(d_1, d_2)$ ,  $\omega_j$ 's and  $(\nu_1, \nu_0)$ .

Among these parameters,  $\sigma_1^2$  and  $\sigma_0^2$  are scaling parameters in the covariance of  $\alpha$  and  $\beta_j^0(t)$ 's. We usually set them between 10 and 100. Larger values also work but don't have significant influence to the posterior estimation of  $\alpha$  and  $\beta_j^0(t)$ 's. The parameter  $\omega_j$  reflects the *a priori* belief on the probability that the  $j$ th functional predictor is selected. If no further information is available on the preference of selecting certain functional predictor, we can set  $\omega_j$  to be a constant across all  $j$ 's, which is the proportion of functional predictors we expect to select.  $d_1$  and  $d_2$  are the parameters of the inverse-gamma prior for the scaling parameter  $\sigma_b^2$ . To determine these two parameters, our suggestion is to set up a mean and variance for the inverse-gamma prior and solve for  $d_1$  and  $d_2$ . For example, if one set the inverse-gamma prior for  $\sigma_\beta^2$  with mean 1 and variance 80, the resulting solution is  $d_1 = 2.01$ ,  $d_2 = 0.9$ . On the

setting of  $(\nu_1, \nu_0)$ , since we have scaling parameters  $\sigma_b^2$  and  $\sigma_0^2$  for  $\gamma_{\tau_j}$ , we usually fix  $\nu_1 = 1$  and set  $\nu_0$  close to zero (e.g,  $\nu_0^2 = 10^{-6}$ ).

Other parameters, such as  $\delta$ ,  $q$ ,  $\xi$  and  $a$ , also need to be determined in the two MCMC algorithms. Parameter  $\delta$  affects the acceptance rate of  $\sigma_b^2$ . It turns out that an empirical value of  $\delta$  between 0.5 and 2 yields acceptance rate approximately between 20% and 60%. Parameter  $q$  in algorithm 2 determines the probability of mutation, which is usually set to be 0.5. Another parameter  $\xi$  determines the swapping probability in step 3 of algorithm 1 and in the mutation step in algorithm 2. No significant improvement on the acceptance rate of  $\tau$  is found when adjusting the values of  $\xi$ , so we usually set it to be 0.5. The geometric ratio  $a$  in Algorithm 2 controls the temperature ladder, and the initial value of  $a$  is usually set to be 4.

## 5.5 Simulation Results

We conduct two simulation studies to evaluate the performance of the proposed model for functional data classification. In both simulations, we generate data with random effects and fixed effects. Simulation 1 uses only 4 functional predictors, in which case Algorithm 1 is expected to work well. Simulation 2 raises the number of functional predictors to 20, and algorithm 1 suffers slow mixing. Algorithm 2 is used, which improves the mixing for posterior samples of  $\tau$ .

### 5.5.1 Simulation 1

We generate  $n = 1000$  i.i.d. observations, using 2 non-functional predictors and 4 functional predictors. For the non-functional predictors, one of them is generated from a uniform distribution on  $[0, 1]$ , the other is a binary variable. The 4 functional predictors are generated using the first 10 orthonormal cosine bases on interval  $[0, 1]$ , i.e., using  $\phi_0(t) = 1, \phi_k(t) = \sqrt{2} \cos(k\pi t), k = 1, \dots, 9$  (see Eubank (1999) for details of cosine series). The random effect has two levels, which result in two vectors of coefficients:  $b_l, l = 1, 2$ . We set the true value of  $\tau$  to be  $(0, 1, 0, 1)$ , indicating that the first and the third function do not contribute to the model, i.e.,  $\beta_1^l(t) = \beta_3^l(t) \equiv 0, \forall l$ . Other parameters used to generate the data are set as  $\sigma_0^2 = 10, \sigma_1^2 = 10, \sigma_b^2 = 5$ , and  $\nu_1^2 = 1$ . The weights  $\{w_k^j\}_{k=1}^\infty$  used for the prior covariance are determined using parameters  $m_1 = 0.8, m_2 = 3$ . The binary responses are generated based on (5.1) using numerical integration. After data generation, we randomly split the data into a training set with 800 observations and a test set with 200 observations.

The proposed model in Section 2 is applied to the training data. We use FPC to construct the orthonormal basis and set the approximation criterion described in Section 5.4 to be  $c_1 = 0.99$ , which results in  $p_j = 4$  for all  $j$ . Based on the FPC scores, the model is trained using Algorithm 1 with the following prior parameters:  $\sigma_0^2 = \sigma_1^2 = 100, d_1 = 2.01, d_2 = 0.9, w_j \equiv 0.5, \nu_1^2 = 1$ , and  $\nu_0^2 = 10^{-6}$ . The prior parameters for the weight matrix  $W$  is set by letting  $m_1 = 0.9, m_2 = 2$ . Other parameters in the MCMC are set as follows:  $\delta = 0.9$ , which gives an acceptance

rate of  $\sigma_b^2$  around 45%;  $\xi = 0.5$ , which is the swapping probability in step 3 of algorithm 1. After 10000 iterations with a burn in period of 4000, we find that the posterior samples of  $\tau$  converge to the true  $\tau$  within 50 iterations. The estimated marginal posterior probability  $P\{\tau_j = 1, j = 1, \dots, 4\} = (0, 1, 0, 1)$ , indicating that our algorithm has successfully selected the second and the fourth functional predictor as expected. Figure 5.2 shows the autocorrelation plot of the posterior samples of  $\sigma_b^2$  and the corresponding histogram plot. We check the convergence of  $\sigma_b^2$  using the Geweke convergence diagnostic test (Geweke 1992). This test uses the first 10% and last 50% of the posterior  $\sigma_b^2$  samples, and yields a Z-score of  $-0.67$ , indicating appropriate convergence. Note that since the orthonormal bases used for estimation and data generation are different, the posterior estimates of  $b_l$ 's and  $b_0$  are not comparable with the true values. Figure 5.3 shows the posterior means of the coefficient functions and the corresponding simultaneous 95% credibility bands for the non-zero coefficient functions, together with the true functions. The simultaneous credibility band is obtained by finding a constant  $M$ , such that 95% of the simulated posterior functions fall into the interval  $\hat{\beta}_j^l(t) \pm M\hat{\sigma}_j^l(t)$ ,  $\forall t$ , where  $\hat{\beta}_j^l(t)$  and  $\hat{\sigma}_j^l(t)$  are the posterior mean and standard deviation of the coefficient functions. From Figure 5.3, we see that the true coefficient functions lie in the 95% confidence bands.

After the training step, the estimated coefficient functions are applied to the test set to get the posterior predictive probability. Treating  $y_i = 1$  as diseased and  $y_i = 0$  as normal, the prediction on the test set gives sensitivity 93% and specificity 99%,

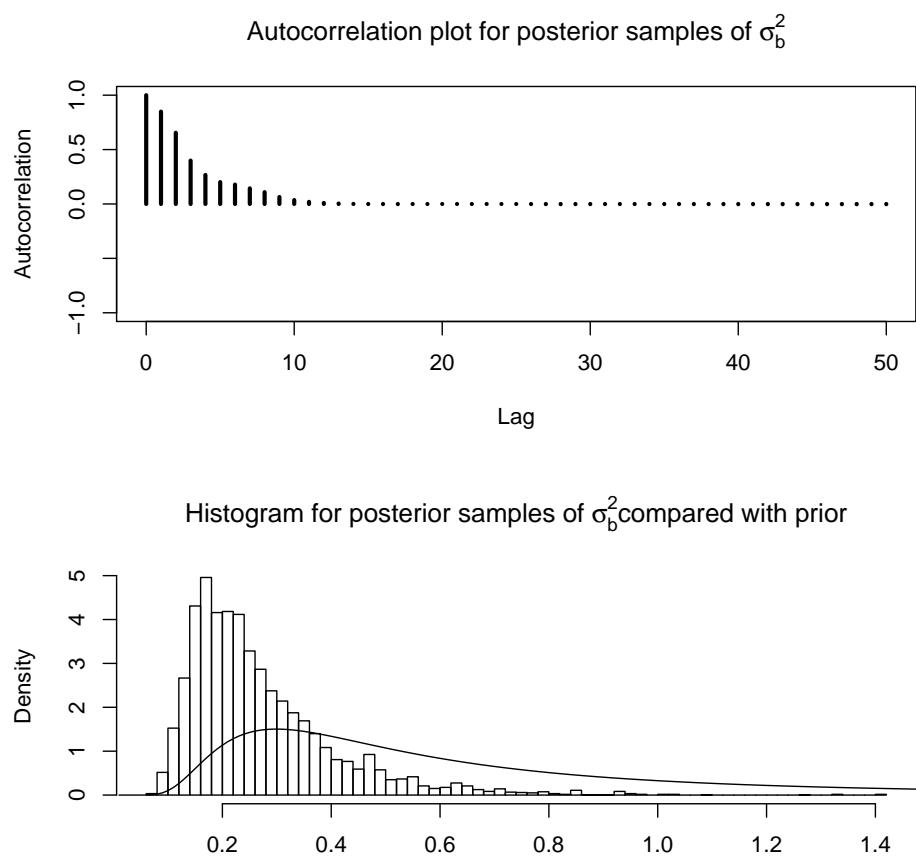


Figure 5.2: Result of Simulation 1: The autocorrelation plot for posterior samples of  $\sigma_b^2$  and the corresponding histogram plot. On the bottom panel, the curve on top of the histogram is the prior density of  $\sigma_b^2$ .

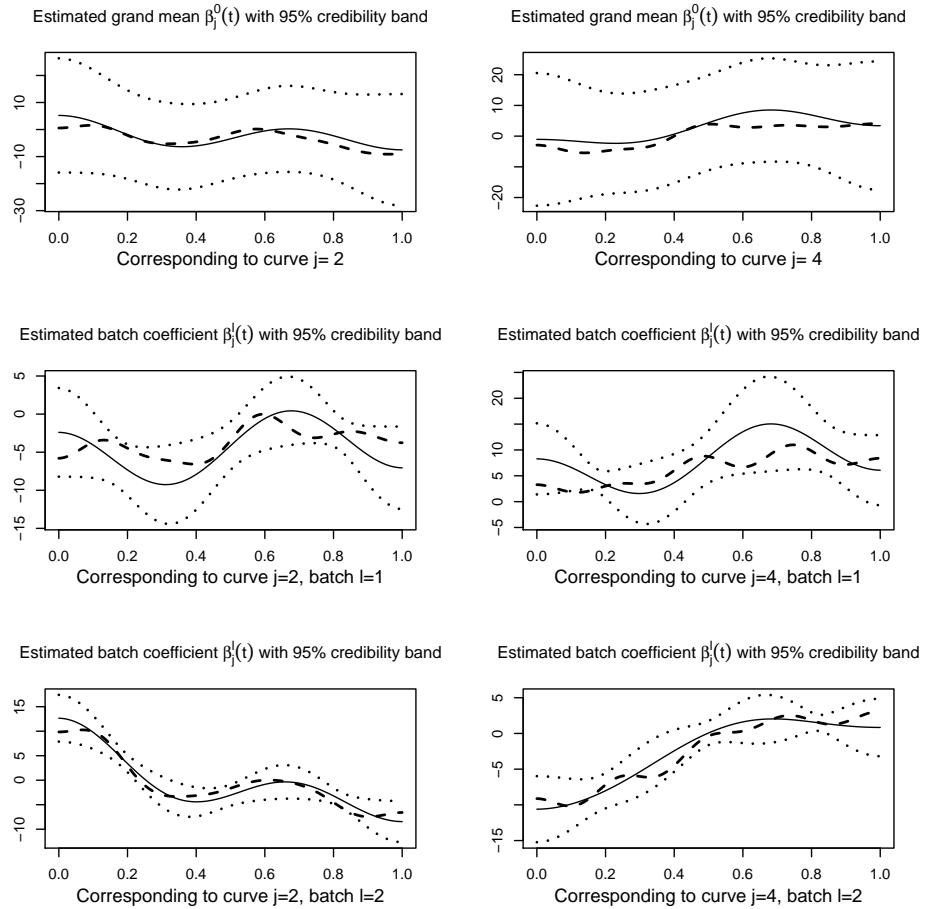


Figure 5.3: The posterior estimation of the non-zero coefficient functions  $\beta_j^l(t)$  and their 95% credibility band, compared with the true coefficient functions used to generate the data. Here  $j$  is the index for multiple functional predictors, and  $l$  is the index for batch.  $\beta_j^0(t)$ 's are the grand means of all batch coefficients. The solid lines denote the posterior mean; the dotted lines denote the 95% credibility bands; the dashed lines denote the true coefficient functions. We only listed the estimations for  $j = 2, 4$  since the functional predictors 1 and 3 are unselected and thus the associated coefficient estimations are close to zero.

with a total misclassification rate 4%. Note that the results reported here are obtained by maximizing the sum of sensitivity and specificity on the empirical ROC curve (see Zweig and Campbell (1993) for an introduction to ROC Curves).

As mentioned in Section 5.3, in Algorithm 1 we use a Metropolis-Hastings step with a “switch/swap” proposal to update the parameter  $\tau$ . In this simulation, the searching space for  $\tau$  only has  $2^4$  possible values. The tracing of the posterior samples of  $\tau$  shows that Algorithm 1 starts from a random value, reaches the correct value in only 6 iterations and stays there afterwards. However, as the length of  $\tau$  increases, the size of the state space increases exponentially, and the samples proposed by “switch/swap” can hardly be accepted. Simulations show that when the length of  $\tau$  goes beyond 8, Algorithm 1 suffers extremely low acceptance rate for  $\tau$  and the MCMC mixes very slowly. Therefore we suggest to use Algorithm 2 when more than 8 functional predictors are involved.

### 5.5.2 Simulation 2

To evaluate the performance of Algorithm 2 when there are a relatively large number of functional predictors, we generate  $n = 1000$  i.i.d. observations using the first 10 cosine bases but increase the number of functional predictors per observation to 20. We set the true  $\tau$  to be a binary vector such that 8 out of the 20 components are 1’s. Other parameters are set to be the same as in simulation 1. Again, we split the data into training and test set as in simulation 1.

Similarly as in simulation 1, in the dimension reduction step, we set the approximation criterion  $c_1 = 0.99$ , which results in  $p_j = 4$  for all  $j$ . Eight parallel chains are used in Algorithm 2 with a maximum temperature of 100. To construct the temperature ladder, we set the geometric ratio starting at 4. Other prior parameters are set similarly as in Simulation 1. We perform 20000 MCMC iterations, in which the first 5000 iterations are used as a burn-in period to adjust the temperature ladder, and another 5000 are treated as a second-stage burn-in period. Therefore the posterior inference is based on the last 10000 iterations. Coded in R language, the simulation takes about 11 hours when running on one dual-processor (900MHz Intel Itanium 2 for each) login node (8GB RAM) of a computing cluster. The final temperature ladder after the burn-in period adjustment is  $(100, 6.79, 1, 0.031, 0.002, 1.4 \times 10^{-4}, 9.8 \times 10^{-6}, 6.7 \times 10^{-7})$ . We obtain several acceptance rates for diagnosis. The acceptance rate of  $\sigma_b^2$  is 31%. The acceptance rates of  $\tau$  for different chains in the mutation operation are  $(0.25, 0.02, 0.001, 9 \times 10^{-4}, 8 \times 10^{-4}, 6 \times 10^{-4}, 5 \times 10^{-4}, 4 \times 10^{-4})$ , in the order of the temperature ladder. The acceptance rates for crossover and exchange operations are 38% and 78%, respectively. We plot the estimated marginal posterior probability  $P\{\tau_j = 1, j = 1, \dots, 20\}$  under three selected temperatures in Figure 5.4, together with the true value of  $\tau$ . This figure shows that at temperature 100 the marginal posterior probabilities are non-zero for all components of  $\tau$ . The chains with temperature 1 and with the lowest temprature produce similar marginal posterior probilities, and they both pick out the correct functional predictors. The

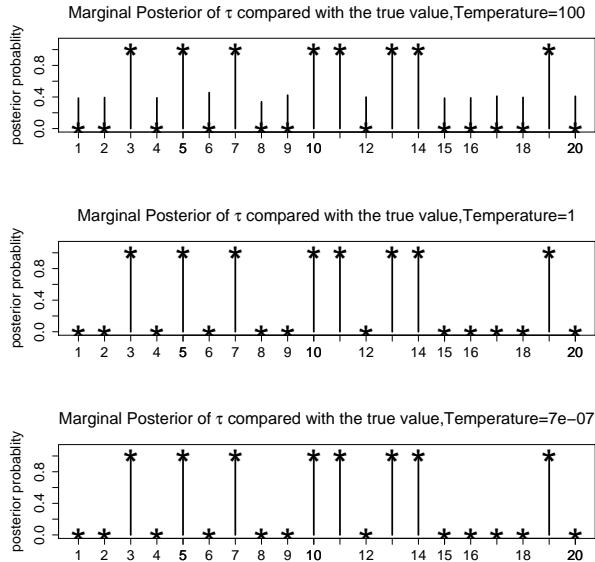


Figure 5.4: The marginal posterior probabilities  $P\{\tau_j = 1, j = 1, \dots, J\}$  at 3 different selected temperatures. The symbol  $\star$  indicates the true value of each component of  $\tau$ .

estimated regression coefficient functions are obtained and applied to the test set for prediction, with a resulting sensitivity of 91%, specificity of 99% and misclassification error of 5%.

## 5.6 Fluorescence Spectroscopy Data Application

The proposed model is applied to the fluorescence spectroscopy data introduced in Section 1.2. In this dataset, every EEM measurement is an observation with 16 functional predictors, corresponding to the 16 excitation wavelengths. Our goal is to select a subset of the 16 curves in the EEM to reduce the cost of data collection, and perform classification based on the selected subset.

There are totally 2414 measurements taken from 1006 patients. Each patient has 1 or more (up to 6) sites measured and some patients may have repeated measurements. All the measurements come from 6 device-clinic combinations, which we treat as the sources of random effects. We also consider two fixed effects: tissue-types, coded as 1, 2 and menopausal status, coded as 1, 2, 3, and treat them as non-functional predictors in the proposed model. After pre-processing (background correction, smoothing, etc), the total 2414 measurements are randomly split into a training set with 1353 observations and a test set with 1061 observations. This partition is conducted at patient level, i.e., measurements from the same patient cannot exist in both training set and test set. The proportion of diseased observations in the training and test set are 10% and 9%, respectively. We use both cosine basis expansion and FPC to approximate functional predictors. To avoid possible bias, the computation of FPC scores for the test set is based on the eigenfunctions estimated from the training set. We determine the number of basis used for each curve by setting the approximation criterion  $c_1 = 0.998$  for FPC, and  $c_2 = 0.992$  for cosine basis expansion. The resulting  $p_j$ 's lie between 2 and 4 for each functional predictor. The priors are set as:  $\sigma_0^2 = \sigma_1^2 = 100$ ,  $d_1 = 2.01$ ,  $d_2 = 0.9$ ,  $w = 0.5$ ,  $\nu_1 = 1$ , and  $\nu_0 = 0.001$ . Using the way described in Section 5.4, the weight matrix  $W$  is determined by setting  $m_1 = 0.8$ ,  $m_2 = 3$ . For both FPC and cosine basis expansion, we use 9 parallel chains, and set the initial geometric ratio  $a = 4$ . The maximum temperature is 10 in the FPC case and 5 in the cosine expansion case. Other parameters are set as:  $\delta = 0.9$ ,  $q = 0.5$ ,

Table 5.1: Real Data Application: The acceptance rates for the EMC algorithm based on two different function approximation methods. M-H denotes the Metropolis-Hastings update. The vector values correspond to the acceptance rates of all chains at the temperature ladder stated in the text.

Accept. rate	Method using cosine basis	Method using FPC's
M-H for $\sigma_b^2$	0.457	0.439
Mutation for $\tau$	$(31, 18, 7, 8, 8, 6, 6, 6, 5) \times 10^{-2}$	$(39, 28, 18, 10, 5, 6, 5, 4, 4) \times 10^{-2}$
Crossover for $\tau$	0.23	0.20
Exchange for $\tau$	0.44	0.48

$\xi = 0.5$ . Similary as in Simulation 2, we perform 20000 MCMC iterations with 5000 burn-in iterations for temperature ladder adjustment, and treat an additional 5000 iterations as a second-stage burn-in period. The acceptance rates in both cases are listed in Web Table 1. In Figure 5.5, we plot the estimated marginal posterior probabilities  $P\{\tau_j = 1, j = 1, \dots, 16\}$  for both cases. From Figure 5.5, we see that the two basis expansion methods provide similar marginal posterior probabilities for  $\tau$ , and both methods show high probability of selecting functions at excitation 340 and 400nm, followed by functions at excitation 470 and 480nm and others. The marginal posterior probabilities suggest the selection order of the functional predictors, higher quantities indicating higher priority of being selected. For example, if we would like to select 4 functional predictors, both methods of basis expansion suggest to select functions at excitation 340, 400, 470 and 480nm. The posterior estimate for  $\sigma_b^2$  is 0.253 using FPC, and is 0.248 using cosine basis expansion.

The posterior inference for functional predictor selection can also be based on

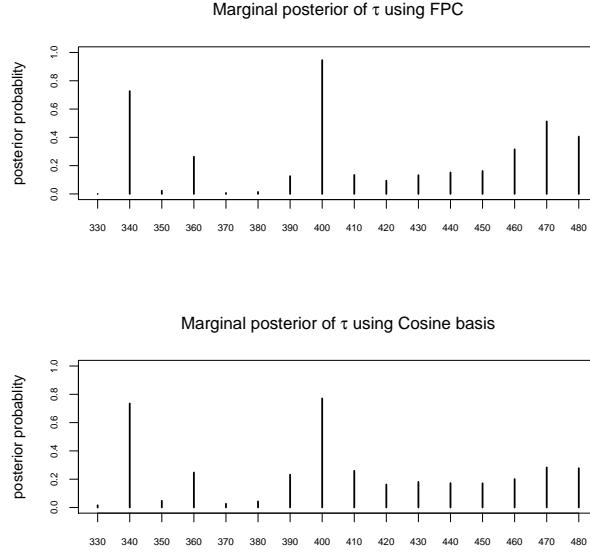


Figure 5.5: The marginal posterior probabilities  $P\{\tau_j = 1, j = 1, \dots, 16\}$  for both cases basis expansions. The top panel is based on FPC, the bottom panel is based on Cosine basis expansion.

the joint posterior distribution of  $\tau$  rather than the marginals. In Figure 5.6, we plot the most frequently visited models for the two function approximation methods. Figure 5.6 shows that both methods select curves at excitation wavelength 340 and 400nm with high frequency. The curves at excitation wavelength 470 or 480nm are also selected frequently but they rarely appear in the same model.

The estimated regression coefficients are applied to the test set for prediction. Table 5.2 lists the prediction results in comparison with 5 other classifiers. Note that all the classifiers in Table 5.2 use both non-functional and all 16 functional predictors. In particular, the BVS model is the Bayesian variable selection method proposed in Chapter 3, which does not consider random effects and functional predictor selection.

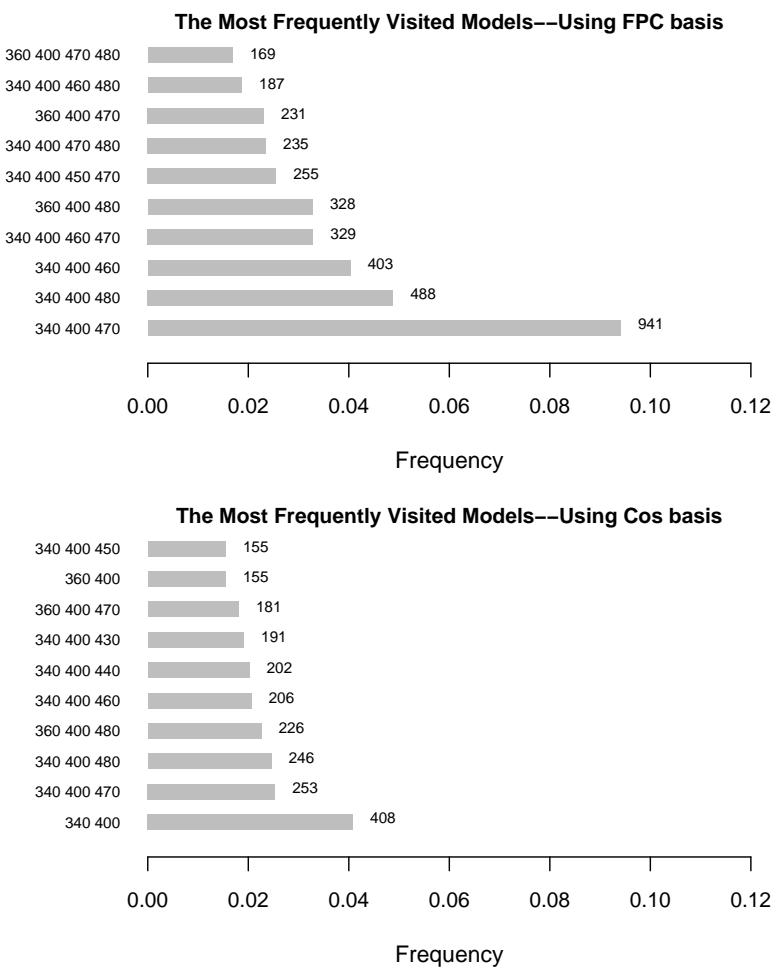


Figure 5.6: The top 10 most frequently visited models for both methods of basis expansion.

Table 5.2: The prediction on test set results using the proposed model(BHFPS) compared with 5 other methods. Two methods of dimension reduction are used: cosine series expansion and functional principal component analysis. AUC: Area under ROC curve; MisR: misclassification rate; Sens: sensitivity; Speci: specificity; BHFPS: the proposed Bayesian hierarchical functional predictor selection model; BHVS: Bayesian hierarchical variable selection model; BVS: regular Bayesian variable selection model; KNN: K-nearest neighbor; LDA: linear discriminant analysis; SVM: support vector machine. See text for explanation of BVS and BHVS models

Method	Using Cosine basis expansion				Using FPC			
	AUC	MisR	Sens	Spec	AUC	MisR	Sens	Spec
BHFPS	0.817	24.2%	74.7%	75.9%	0.822	21.2%	72.6%	79.4%
BHVS	0.819	25.6%	76.8%	74.1%	0.824	27.2%	77.9%	72.3%
BVS	0.802	28.1%	76.8%	71.4%	0.819	30.5%	84.2%	68.0%
KNN	0.697	27.7%	62.1%	73.3%	0.718	32.1%	71.8%	74.7%
LDA	0.796	27.3%	74.7%	72.5%	0.804	25.0%	75.8%	74.9%
SVM	0.657	56.6%	85.3%	39.2%	0.679	38.4%	68.4%	61.0%

The Bayesian hierarchical variable selection (BHVS) is an extension of the BVS model which considers random effects by a hierarchical setup, but does not perform functional predictor selection. From Table 5.2, we see that the proposed method (BHFPS) provides comparable prediction results with BHVS. Both BHFPS and BHVS obtain slightly higher AUC scores than the BVS model does. Table 5.2 also shows that the two orthonormal basis expansion methods are comparable in their prediction ability, although the cosine basis expansion method has slightly lower AUC than the FPC method. In Figure 5.7, we compare the empirical ROC curves for models listed in Table 5.2 based on the FPC method.

Based on the functional predictors selected by the proposed model, other classification algorithms can be trained independently using the selected curves only. For

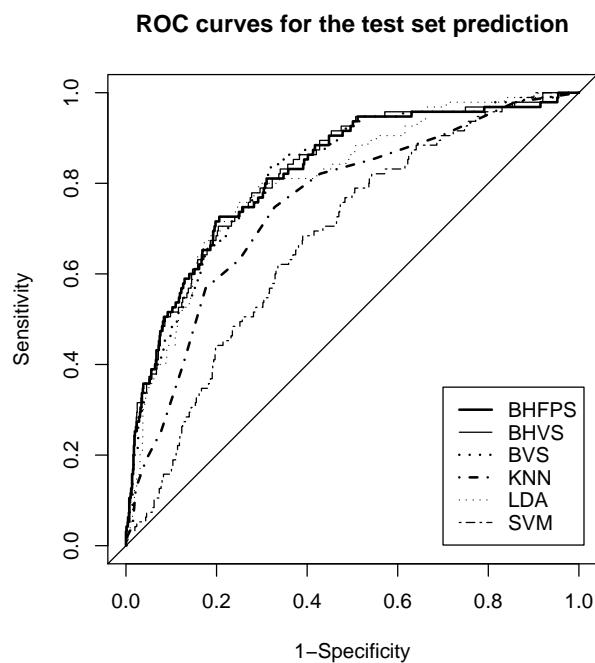


Figure 5.7: ROC curves obtained by test set prediction using the proposed model compared with 5 other classifiers, where BHFPS, BHVS, BVS, KNN, LDA, SVM are defined in table 5.2.

example, training the BHVS model on the first 4 functional predictors selected by the proposed model (based on the marginal posterior of  $\tau$ ) gives sensitivity 77.9% and specificity 70.0%, with corresponding AUC 0.819 and misclassification rate 20.7%. Compared with those in Table 5.2, we see that these prediction results are as good as those based on all the 16 curves. Hence it is possible to achieve a high prediction power by using a subset of functional predictors. Using the selected curves, a new device can be constructed which reduces cost and saves measurement time.

## 5.7 Discussion

Motivated by practical problems on functional data classification, we have proposed a Bayesian hierarchical model to deal with the situations when functional predictors are contaminated by random batch effects. Inferences based on this model help to select a subset of functional predictors for classification. This model is applied to an application problem which uses fluorescence spectroscopy data for pre-cervical cancer diagnosis. The results suggest that it is possible to build more cost-effective device with less spectral curves. In this section, we discuss some issues related to the proposed model.

The first one is about the prior correlation matrix of  $\beta_j^l(t)$ . When setting priors for the coefficient functions in (5.2), we assume that  $\beta_j^l(t)$  are independent for all  $j$  and  $l$ , which leads to the prior correlation matrix  $R = I_p$  in (5.9) after approximation by basis expansion. This is just a simplified prior choice. It is possible to allow the

priors for  $\beta_j^l(t)$  to be correlated. For example, we may assume that  $(\beta_1^l(t), \dots, \beta_J^l(t))$  has a multivariate Gaussian process, as done in Morris and Carroll (2006). In such a case, it may be difficult to determine the prior correlations and the resulting posterior computation may be complex.

Another issue is about the necessity of using a hierarchical structure to adjust for batch effects. As we have pointed out in Section 5.1, for data obtained from an unbalanced experimental design, classification can be easily biased by batch effects. Algorithms which do not adjust for batch effects may result in classification based on batch difference, rather than the disease information. Using a hierarchical model is a natural way to model the batch structure. In our real data application, the hierarchical models (BHFPS and BHVS) are more preferable as they account for possible batch effects, although they may not necessarily improve prediction over models like BVS (see Table 5.2 and Figure 5.7). In fact, we should not always expect to improve the prediction by accounting for batch effects, since with a bad experimental design, a classification algorithm can get prediction as good as 100% sensitivity and specificity, by simply using the batch information (Baggerly et al., 2004).

As a side note, in our simulation and real data applications, we train the proposed model using data from all batches, and make predictions based on observations with the same batch information. Prediction on observations from new batches is also applicable. However, it is natural to expect that the prediction will be worse when predicting on new batches, since the random effect of the new batch is unknown when

training the model.

Finally, like many other regression problems, when there exists severe collinearity between the functional predictors, a unique solution for the “best” subset may not be guaranteed using our proposed model. In this case, exploring functional predictor selection from a Bayesian decision theory point of view may provide a solution.

# Chapter 6

## Priors for Covariance Operators in Functional Data Analysis

In this chapter, we discuss the properties of covariance operators of functional data and the conditions for formulating appropriate priors for such covariance operators. We also propose a prior and prove some of its mathematical properties.

### 6.1 Grid Refinement Invariance Principle

Although functional data ideally live in infinite dimensional space, they can only be collected and stored in finite dimensional (multivariate) form. They are typically recorded either on some fine grids or in forms of finite linear combinations of basis functions. For example, for a random function  $X(t)$  defined on a compact domain  $T \subset \mathbb{R}$ , one can discretize  $T$  on a grid of  $p$  points,  $T_p = (t_1, \dots, t_p)^T$ . A realization

of  $X(t)$ ,  $x(t)$  can thus be stored in a vector form  $\vec{x} = (x(t_1), \dots, x(t_p))^T$ , although  $p$  can be very large and  $x(t_i)$  can be very close to  $x(t_{i+1})$ . A linear interpolation of  $\vec{x}$  on the grid  $T_p$  provides an approximation of  $x(t)$ . Statistical methods which treat functional data as multivariate fail to make use of the “functional structure” of the data. The study in this chapter is motivated by a general principle of functional data analysis stated as follows:

**Grid Refinement Invariance Principle(GRIP)** *As the order of approximation becomes more exact, i.e., the grids become finer or the upper limit of the basis function expansion tends to infinity, the functional data analysis method should approach the appropriate limiting analogue of the true functional (infinite dimensional) observations.*

Under GRIP, we would like to look for functional data analysis models that are appropriately defined in the infinite dimensional space and project them down to finite dimensional space in implementation. This makes it necessary to investigate the properties of functional data in infinite dimensional space. We study these properties based on the theoretical structure of Gaussian measures.

## 6.2 Gaussian Measures

We follow Prato [56] to define Gaussian measures but use slightly different notations. Let  $H$  be a separable Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and norm  $|\cdot| = \sqrt{\langle \cdot, \cdot \rangle}$ . In this chapter, we assume that  $H$  is associated with a real scalar field. For convenience,

we write a sequence  $\{x_k\}_{k=1}^{\infty}$  in  $H$  as  $(x_k)$ . Let  $\mathcal{B}(H)$  be the Borel  $\sigma$ -field on  $H$ . We use  $L(H)$  to denote the Banach algebra of all continuous linear operators from  $H$  to  $H$ , and  $L^+(H)$  represents the subset of  $L(H)$  which contains all symmetric and nonnegative definite operators, i.e.,

$$L^+(H) = \{A \in L(H) : \langle Ax, y \rangle = \langle x, Ay \rangle, \forall x, y \in H, \text{ and } \langle Ax, x \rangle \geq 0, \forall x \in H\}.$$

Furthermore, we denote as  $L_{(1)}$  the subset of  $L(H)$  that are trace class operators, in the sense that if  $A \in L_{(1)}$ , then  $(A^*A)^{1/2}$  has eigenvalues  $\{\lambda_k\}_{k=1}^{\infty}$  with  $\sum_{k=1}^{\infty} \lambda_k < \infty$ . The trace of  $A \in L_{(1)}$  is defined as

$$\text{Tr} A = \sum_{k=1}^{\infty} \langle Ae_k, e_k \rangle, \quad (6.1)$$

where  $\{e_k\}_{k=1}^{\infty}$  is an arbitrary complete orthonormal sequence (c.o.n.s.) of  $H$ .  $L_{(1)}^+(H)$  represents the set of all operators in  $L^+(H) \cap L_{(1)}(H)$ . We call operators in  $L_{(1)}^+(H)$   $S$ -operators.

### 6.2.1 Gaussian Measures Defined on Finite-dimensional Hilbert Space

For a pair of real numbers  $(m, s)$  with  $s > 0$ , we define the one-dimensional Gaussian measure (with mean  $m$  and variance  $s$ ) on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  by

$$\mu_{m,s}(dx) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s}} dx.$$

We also allow  $s = 0$ , in which case, for all  $A \in \mathcal{B}(\mathbb{R})$ ,

$$\mu_{m,0}(A) = \delta_m(A) = \begin{cases} 1 & \text{if } m \in A, \\ 0 & \text{if } m \notin A. \end{cases}$$

For a  $d$ -dimensional Hilbert space  $H$  and  $S \in L_{(1)}^+(H)$ , we can find the set of eigenvectors of  $S$ , denoted as  $(e_1, \dots, e_d)$ , which is orthonormal and satisfies

$$Se_k = \lambda_k e_k, k = 1, \dots, d, \text{ for some } \lambda_k \geq 0.$$

For any  $x \in H$ , if  $x_k = \langle x, e_k \rangle, k = 1, \dots, d$ ,  $H$  can be identified with  $\mathbb{R}^d$  through an isomorphism  $\gamma$ :

$$\gamma : H \longrightarrow \mathbb{R}^d, \text{ and } \gamma(x) = (x_1, \dots, x_d), \forall x \in H.$$

We then define the Gaussian measure on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , hence on  $(H, \mathcal{B}(H))$  by

$$\mu_{m,S} = \times_{k=1}^d \mu_{m_k, \lambda_k}, \quad (6.2)$$

which is a product measure formed by  $d$  one-dimensional measures. It is easy to show the following properties of finite dimensional Gaussian measures:

**Proposition 6.2.1.** *Let  $m \in H$ ,  $S \in L_{(1)}^+(H)$ . For  $\mu_{m,S}$  defined in (6.2), we have*

$$\begin{aligned} \int_H x \mu_{m,S}(dx) &= m, \\ \int_H \langle y, x - m \rangle \langle z, x - m \rangle \mu_{m,S}(dx) &= \langle Sy, z \rangle, \forall y, z \in H. \end{aligned}$$

*The characteristic function (Fourier transform) of  $\mu_{m,S}$  is*

$$\widehat{\mu_{m,S}}(h) := \int_H e^{i\langle h, x \rangle} \mu_{m,S}(dx) = e^{\langle m, h \rangle - \frac{1}{2}\langle Sh, h \rangle}, h \in H.$$

*$m$  and  $S$  are called the mean and covariance operator of  $\mu_{m,S}$ . Furthermore, the Gaussian measure is uniquely determined by its characteristic function.*

### 6.2.2 Gaussian Measures Defined on Infinite-dimensional Hilbert Space

Now assume that  $H$  is a infinite dimensional separable Hilbert space. We first define the mean and covariance for a measure  $\mu$  on  $(H, \mathcal{B}(H))$ . Suppose  $\int_H |x| \mu(dx) < \infty$ , for any  $h \in H$ , the linear functional  $f : H \rightarrow \mathbb{R}$  with

$$f(h) = \int_H \langle x, h \rangle \mu(dx), h \in H,$$

is continuous since

$$|f(h)| \leq \int_H |x| \mu(dx) |h|, h \in H.$$

By Riesz representation theorem ([81], page 90), there exists a unique  $m \in H$  such that

$$\langle m, h \rangle = \int_H \langle x, h \rangle \mu(dx), h \in H.$$

We call  $m$  the mean of  $\mu$  and write  $m = \int_H x \mu(dx)$ . Now suppose  $\int_H |x|^2 \mu(dx) < \infty$ .

We consider the bilinear map  $g : H \times H \rightarrow \mathbb{R}$  such that

$$g(h, k) = \int_H \langle h, x - m \rangle \langle k, x - m \rangle \mu(dx), h, k \in H.$$

It is easy to see that  $g$  is continuous since

$$|g(h, k)| \leq \int_H |x - m|^2 \mu(dx) |h| |k|, h, k \in H.$$

Again, by Riesz theorem, there is a unique linear bounded operator  $S \in L(H)$  such that

$$\langle Sh, k \rangle = \int_H \langle h, x - m \rangle \langle k, x - m \rangle \mu(dx), h, k \in H. \quad (6.3)$$

We call  $S$  the covariance of  $\mu$ . It is easy to show that  $S$  is symmetric and nonnegative definite. Also, by the definition of trace in (6.1),

$$\mathrm{Tr}S = \sum_{k=1}^{\infty} \langle Se_k, e_k \rangle = \sum_{k=1}^{\infty} \int_H \langle e_k, x - m \rangle^2 \mu(dx) = \int_H |x - m|^2 \mu(dx) < \infty,$$

where the last equality is by Parseval identity (and monotone convergence theorem), therefore  $S \in L_{(1)}^+(H)$ .

**Definition 6.2.2. Gaussian Measure** Let  $m \in H$  and  $S \in L_{(1)}^+(H)$ . A Gaussian measure  $\mu := \mu_{m,S}$  on  $(H, \mathcal{B}(H))$  is a measure  $\mu$  with mean  $m$ , covariance operator  $S$  and characteristic function

$$\widehat{\mu_{m,S}}(h) = \exp\left\{i\langle m, h \rangle - \frac{1}{2}\langle Sh, h \rangle\right\}, h \in H.$$

The Gaussian measure  $\mu_{m,S}$  is called non-degenerate if  $\mathrm{Ker}(S) = \{x \in H : Sx = 0\} = \{0\}$ . [56]

Prato [56] shows the existence and uniqueness of a Gaussian measure through the following proposition:

**Proposition 6.2.3.** For any  $m \in H$  and  $S \in L_{(1)}^+(H)$ , there exists a unique Gaussian measure  $\mu = \mu_{m,S}$  on  $(H, \mathcal{B}(H))$ . [56]

*Proof.* We summarize Prato's proof here. First, since  $H$  is a infinite dimensional separable Hilbert space, we can define a projection mapping  $P_n : H \longrightarrow P_n(H)$  by  $P_n x = \sum_{k=1}^n \langle x, e_k \rangle e_k, \forall x \in H$ . Then we have  $\lim_n P_n x = x, \forall x \in H$ . This holds for any c.o.n.s.  $(e_k)$  of  $H$ . Since  $S \in L_{(1)}^+(H)$ , there exists a c.o.n.s.  $(e_k)$  and a sequence

of non-negative numbers  $(\lambda_k)$  such that

$$Se_k = \lambda_k e_k, k \in \mathbb{N}.$$

The existence of such  $(e_k)$  and  $(\lambda_k)$  is shown in Theorem 1.5 (spectral representation) by Kuo[34].  $\lambda_k$ 's are called eigenvalues and  $e_k$ 's are called eigenvectors. For any  $x \in H$ , set  $x_k = \langle x, e_k \rangle$ . This constructs an isomorphism  $\gamma$  between  $H$  and  $l^2$  (The space of square summable sequences) defined by

$$\gamma : H \longrightarrow l^2,$$

and  $\gamma(x) = (x_k), \forall x \in H$ . Thus we can identify  $H$  with  $l^2$ . Now we construct the product measure  $\mu := \times_{k=1}^{\infty} \mu_{m_k, \lambda_k}$  over the product space  $\mathbb{R}^{\infty} := \times_{k=1}^{\infty} \mathbb{R}$ . The existence of  $\mu$  is guaranteed by the extension theorem stated in Prato's book([56], Theorem 1.9). So it remains to show that  $\mu$  is a Gaussian measure with mean  $m$ , covariance  $S$ .

For  $h \in H$ ,  $|\langle x, h \rangle| \leq |x||h|$  and

$$\begin{aligned} (\int_H |x| \mu(dx))^2 &< \int_H |x|^2 \mu(dx) = \int_{\mathbb{R}^{\infty}} \sum_{k=1}^{\infty} x_k^2 \mu(dx) \\ &= \sum_{k=1}^{\infty} \int_{\mathbb{R}} x_k^2 \mu_{m_k, \lambda_k}(dx_k) = \sum_{k=1}^{\infty} (\lambda_k + m_k^2) = \text{Tr}S + |m|^2 < \infty. \end{aligned}$$

Hence by dominated convergence theorem,

$$\int_H \langle x, h \rangle \mu(dx) = \lim_n \int_H \langle P_n x, h \rangle \mu(dx).$$

But

$$\begin{aligned} \int_H \langle P_n x, h \rangle \mu(dx) &= \sum_{k=1}^n \int_H x_k h_k \mu(dx) \\ &= \sum_{k=1}^n h_k \int_{\mathbb{R}} x_k \mu_{m_k, \lambda_k}(dx_k) = \sum_{k=1}^n h_k m_k = \langle P_n m, h \rangle \longrightarrow \langle m, h \rangle, \end{aligned}$$

as  $n \rightarrow \infty$ . Therefore  $m$  is the mean of  $\mu$ .

To determine the covariance (operator) of  $\mu$ , we fix  $y, z \in H$  and let

$$\int_H \langle x - m, y \rangle \langle x - m, z \rangle \mu(dx) = \lim_n \int_H \langle P_n(x - m), y \rangle \langle P_n(x - m), z \rangle \mu(dx).$$

Since

$$\begin{aligned} \int_H \langle P_n(x - m), y \rangle \langle P_n(x - m), z \rangle \mu(dx) &= \sum_{k=1}^n \int_H (x_k - m_k)^2 y_k z_k \mu(dx) \\ &= \sum_{k=1}^n y_k z_k \int_{\mathbb{R}} (x_k - m_k)^2 \mu_{m_k, \lambda_k}(dx_k) = \sum_{k=1}^n y_k z_k \lambda_k = \langle P_n S y, z \rangle \longrightarrow \langle S y, z \rangle, \end{aligned}$$

as  $n \rightarrow \infty$ . Therefore  $S$  is the covariance of  $\mu$ .

Finally, we verify that the characteristic function of  $\mu$  is that of a Gaussian measure. For  $h \in H$ ,

$$\begin{aligned} \int_H e^{i\langle x, h \rangle} \mu(dx) &= \lim_{n \rightarrow \infty} \int_H e^{i\langle P_n x, h \rangle} \mu(dx) = \lim_{n \rightarrow \infty} \prod_{k=1}^n \int_{\mathbb{R}} e^{ix_k h_k} \mu_{m_k, \lambda_k}(dx_k) \\ &= \lim_{n \rightarrow \infty} \prod_{k=1}^n e^{im_k h_k - \frac{1}{2}\lambda_k h_k^2} = \lim_{n \rightarrow \infty} e^{i\langle P_n m, h \rangle} e^{-\frac{1}{2}\langle P_n S h, h \rangle} \\ &= e^{i\langle m, h \rangle} e^{-\frac{1}{2}\langle S h, h \rangle}. \end{aligned}$$

So the characteristic function of  $\mu$  is that of a Gaussian measure with mean  $m$ , covariance operator  $S$ . Therefore  $\mu = \mu_{m,S}$ . Since the product measure on  $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$  is a unique extension of  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ ,  $\mu_{m,S}$  is unique.  $\square$

**Proposition 6.2.4.** *A non-degenerate Gaussian measure on  $H$  is fully supported.*

[56]

*Proof.* Let  $B(x, r) \in \mathcal{B}(H)$  be an arbitrary ball with center  $x \in H$  and radius  $r > 0$ .

We just need to show that  $\mu_{m,S}(B(x, r)) > 0$ . Let  $A_n = \{x \in H : \sum_{k=1}^n x_k^2 \leq \frac{r^2}{2}\}$  and  $B_n = \{x \in H : \sum_{k=n+1}^{\infty} x_k^2 < \frac{r^2}{2}\}$ . Then  $\mu(B(0, r)) \geq \mu(A_n \cap B_n) = \mu(A_n)\mu(B_n)$  because  $A_n$  and  $B_n$  are independent([56],example 1.22). Clearly  $\mu(A_n) > 0$ . It suffices to show that  $\mu(B_n) > 0$  for  $n$  large enough. Now, by Markov inequality,

$$\begin{aligned} \mu(B_n) &= 1 - \mu(B_n^c) \geq 1 - \frac{2}{r^2} \sum_{k=n+1}^{\infty} \int_H x_k^2 \mu(dx) \\ &= 1 - \frac{2}{r^2} \sum_{k=n+1}^{\infty} (\lambda_k + m_k^2) > 0, \end{aligned}$$

for  $n$  large enough.  $\square$

### 6.3 A Possible Prior for Covariance Operators

Suppose  $\{X_i\}_{i=1}^n$  are i.i.d. random elements taking values in a separable Hilbert space  $H$ . Let  $\mu(\cdot)$  be a Borel measure defined on  $(H, \mathcal{B}(H))$  such that  $\int_H |X_i| \mu(dX_i) < \infty$  and  $\int_H |X_i|^2 \mu(dX_i) < \infty$ . Let the mean of  $\mu(\cdot)$  be zero and the covariance operator of  $\mu(\cdot)$  be  $A_\mu$ . Then,

$$\langle A_\mu x, y \rangle = \int_H \langle x, z \rangle \langle y, z \rangle \mu(dz)$$

and  $A_\mu \in L_{(1)}^+(H)$ . In order to construct a prior for  $A_\mu$ , we propose the following expansion

$$A = \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j, \quad (6.4)$$

where  $w_j > 0$  and  $\sum_j w_j < \infty$ . The operation  $\otimes$  is defined as

$$(u \otimes v) x = u \langle v, x \rangle, \quad (6.5)$$

for all  $u, v, x \in H$ .  $Z_j$ 's are *a priori* assumed to be i.i.d. zero mean Gaussian random elements in  $H$  with a known covariance operator  $B \in L_{(1)}^+(H)$ . We will show that the right hand side of (6.4) converges almost surely in  $L_{(1)}$ , and  $A$  is in  $L_{(1)}^+(H)$ . Therefore  $A$  is a  $L_{(1)}^+(H)$  random variable. The distribution of  $A$  can be used as a prior for  $A_\mu$ .

To construct a prior for the distribution of (6.4), several conditions need to be satisfied, which should be able to guarantee that the resulting posterior is consistent. We say that a posterior distribution is consistent if the posterior measure on an arbitrary  $\epsilon$ -neighborhood (under some metric such as Hellinger metric) of the true underlying distribution approaches to a point mass almost surely when the number of observed samples approaches infinity. Proofs for posterior consistency under different assumptions can be found in some Bayesian nonparametric literature, such as Barron, Schervish and Wasserman [4], Ghosal, Ghosh and Van Der Vaart[23], Walker[75], and Walker, Lijoi and Prünster[76]. Most proofs for posterior consistency assume that the probability measures under study are absolutely continuous with respect to a  $\sigma$ -finite dominating measure. It remains an open question how to construct the consistency for random functions with infinite-dimensional Gaussian measures.

For a prior in the form of (6.4), we conjecture that  $A$  must take values from the space of  $L_{(1)}^+(H)$  and the distribution of  $A$  must be fully supported on the whole space of  $L_{(1)}^+(H)$ . Intuitively, if we want the posterior distribution to be close to the true density, the prior distribution must put positive mass over a neighborhood of the true density. An supportive example can be found in Schervish's book ([67], page 430, Example 7.79). We will show in Theorem 6.3.1 and Theorem 6.3.4 that  $\sum_{j=1}^{\infty} w_j Z_j \otimes Z_j$  converges in  $L_{(1)}^+(H)$  almost surely and its distribution is fully-supported on  $L_{(1)}^+(H)$ .

**Theorem 6.3.1.** *Let  $Z_j \in L^2$  be i.i.d. zero mean Gaussian random functions taking values in  $H$ , where  $H$  is a separable Hilbert space associated with norm  $|\cdot|$ , then the random covariance operator  $\sum_{j=1}^n w_j Z_j \otimes Z_j$  is in  $L_{(1)}^+(H)$  for every finite  $n$ , and*

$$\sum_{j=1}^n w_j Z_j \otimes Z_j \xrightarrow{a.s.} A$$

as  $n \rightarrow \infty$  for some  $A \in L_{(1)}^+(H)$ .

*Proof.* Since the scalar field associated with  $H$  is real,  $\langle x, ay \rangle = \langle ax, y \rangle = a\langle x, y \rangle$ .

1. First, we show that  $Z_j \otimes Z_j$  is a random operator taking values in  $L_{(1)}^+(H)$ .

$\forall x, y \in H$ , we have

$$\langle Z_j \otimes Z_j x, x \rangle = \langle Z_j \langle Z_j, x \rangle, x \rangle = \langle Z_j, x \rangle^2 \geq 0,$$

$$\langle Z_j \otimes Z_j x, y \rangle = \langle Z_j \langle Z_j, x \rangle, y \rangle = \langle Z_j, x \rangle \langle Z_j, y \rangle = \langle x, Z_j \otimes Z_j y \rangle.$$

This proves that  $Z_j \otimes Z_j$  is positive definite and self-adjoint. To show that it is trace class, let  $(e_i)$  be a c.o.n.s. of  $H$ , if we denote  $\|\cdot\|_{L(1)}$  as the trace class

norm, then

$$\|Z_j \otimes Z_j\|_{L_{(1)}} = \sum_{i=1}^{\infty} |\langle Z_j \otimes Z_j e_i, e_i \rangle| = \sum_{i=1}^{\infty} \langle Z_j, e_i \rangle^2 = |Z_j|^2 < \infty.$$

Now, for  $n$  fixed and  $w_j > 0, \forall j \leq n$ , and  $\forall x, y \in H$ , we have

$$\left\langle \sum_{j=1}^n w_j Z_j \otimes Z_j x, x \right\rangle = \sum_{j=1}^n w_j \langle Z_j, x \rangle^2 \geq 0, \quad (6.6)$$

$$\left\langle \sum_{j=1}^n w_j Z_j \otimes Z_j x, y \right\rangle = \sum_{j=1}^n w_j \langle Z_j, x \rangle \langle Z_j, y \rangle = \langle x, \sum_{j=1}^n Z_j \otimes Z_j y \rangle, \quad (6.7)$$

$$\left\| \sum_{j=1}^n w_j Z_j \otimes Z_j \right\|_{L_{(1)}} = \sum_{i=1}^{\infty} \left| \left\langle \sum_{j=1}^n w_j Z_j \otimes Z_j e_i, e_i \right\rangle \right| = \sum_{j=1}^n w_j |Z_j|^2 < \infty. \quad (6.8)$$

This proves that  $\sum_{j=1}^n w_j Z_j \otimes Z_j \in L_{(1)}^+(H)$  for every finite  $n$ .

2. Now let  $A_n = \sum_{j=1}^n w_j Z_j \otimes Z_j$  and  $A = \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j$ . Note that we will also need to show that  $A$  exists. The idea is to show that  $A_n$  is a Cauchy sequence almost surely. Let  $m > n$ , then

$$\|A_m - A_n\|_{L_{(1)}} = \left\| \sum_{j=n+1}^m w_j Z_j \otimes Z_j \right\|_{L_{(1)}} = \sum_{j=n+1}^m w_j |Z_j|^2.$$

Therefore we just need to show that  $\sum_{j=n+1}^m w_j |Z_j|^2 \rightarrow 0$ , as  $m$  and  $n$  approaches infinity, which is equivalent to show that the series  $\sum_{j=1}^{\infty} w_j |Z_j|^2$  converges (since  $Z_j$ 's are independent). This will be shown in the following (i.e., (a)-(c)) when we prove that  $A$  is trace class.

Firstly, we have,

$$A = \lim_{n \rightarrow \infty} \sum_{j=1}^n w_j Z_j \otimes Z_j = \lim_{n \rightarrow \infty} A_n,$$

and

$$\langle A_n x, x \rangle \geq 0 \text{ (by (6.6))} \Rightarrow \langle \lim_n A_n x, x \rangle = \lim_n \langle A_n x, x \rangle \geq 0,$$

by the continuity of inner-product. Similarly,

$$\langle A_n x, y \rangle = \langle x, A_n y \rangle \text{ (by (6.7))} \Rightarrow \langle \lim_n A_n x, y \rangle = \langle x, \lim_n A_n y \rangle.$$

Hence  $A$  is positive definite and self-adjoint. To show that  $A$  is trace class, we just need to show that  $\|A\|_{L_{(1)}} < \infty$ . For a c.o.n.s.  $(e_k)$  of  $H$ , since

$$\|A\|_{L_{(1)}} = \left\| \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j \right\|_{L_{(1)}} = \sum_{j=1}^{\infty} w_j \sum_{i=1}^{\infty} \langle Z_j, e_i \rangle^2 = \sum_{j=1}^{\infty} w_j |Z_j|^2,$$

it suffices to show the a.s convergence of the random series  $\sum_{j=1}^{\infty} w_j |Z_j|^2$ . We use Kolmogrov three series theorem [63] to show this.  $\forall c > 0$ , we have

$$(a) \sum_j P[w_j |Z_j|^2 > c] = \sum_j P[|Z_j|^2 > \frac{c}{w_j}] \leq \sum_j \frac{E[|Z_j|^2]}{c/w_j} = \frac{E[|Z_1|^2]}{c} (\sum_j w_j) <$$

$\infty$ , by Markov inequality and  $Z_j \in L^2$ .

$$(b) \sum_j E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] \leq \sum_j w_j E[|Z_j|^2] = E[|Z_1|^2] (\sum_j w_j) < \infty, \text{ by the fact that } Z_j \text{ are i.i.d. and } Z_j \in L^2.$$

(c) We have

$$\begin{aligned} & \sum_j Var(w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}) \\ &= \sum_j E[w_j^2 |Z_j|^4 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] - \sum_j E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}]^2 \end{aligned}$$

with

$$\begin{aligned} 0 \leq \sum_j E[w_j^2 |Z_j|^4 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] &\leq \sum_j c E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] \\ &\leq \sum_j c E[w_j |Z_j|^2] \leq c E[|Z_1|^2] (\sum_j w_j) < \infty \end{aligned}$$

and

$$0 \leq \sum_j E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}]^2 \leq \left( \sum_j E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] \right)^2 < \infty,$$

by  $E[w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}] \geq 0, \forall j$  and results of part (b). Therefore,

$$\sum_j Var(w_j |Z_j|^2 \mathbf{1}_{\{w_j |Z_j|^2 < c\}}) < \infty \text{ a.s.}$$

Thus  $\|A\|_{L_{(1)}} = \sum_{j=1}^{\infty} w_j |Z_j|^2$  converges a.s. in  $L_{(1)}^+(H)$ .

□

Before stating Theorem 6.3.4, we first give the definition for the support of a measure as follows:

**Definition 6.3.2.** *Let  $\mu$  be a measure defined on a measurable space  $(\Omega, \mathcal{B})$ . The support of  $\mu$  (denoted as  $supp(\mu)$ ) is the set of all points  $\omega$  in  $\Omega$  for which every open neighborhood  $N_\omega$  of  $\omega$  has positive measure, i.e.,*

$$supp(\mu) = \{\omega \in \Omega | \omega \in \Omega \implies \mu(N_\omega) > 0\}.$$

In some cases, we refer to the support of a random element as the support of the induced measure. In Definition 6.3.3, we define the induced probability measure for a random element following Resnick ([63], page 75).

**Definition 6.3.3.** Let  $(\Omega, \mathcal{B}, \mu)$  be a probability space, and suppose

$$X : (\Omega, \mathcal{B}) \longmapsto (\Omega', \mathcal{B}')$$

is measurable. For  $A' \subset \Omega'$ , let

$$[X \in A'] := X^{-1}(A') = \{\omega : X(\omega) \in A'\}.$$

Define the set function  $\mu \circ X^{-1}$  on  $\mathcal{B}'$  by

$$\mu \circ X^{-1}(A') = \mu(X^{-1}(A')).$$

Then  $\mu \circ X^{-1}$  is a probability on  $(\Omega', \mathcal{B}')$  called the induced probability or distribution of  $X$ , denoted as  $\text{Law}[X]$ .

According to Definition 6.3.3, it is clear that  $\text{supp}(\text{Law}[X]) \subset \Omega'$ .

**Theorem 6.3.4.** If we denote the measure of the random covariance operator  $A = \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j$  as  $\text{Law}[A]$ , then  $\text{Law}[A]$  is fully supported on the whole  $L_{(1)}^+(H)$  space, i.e.,

$$\text{supp}(\text{Law}[A]) = L_{(1)}^+(H).$$

*Proof.* Let  $A_0$  be a fixed operator in  $L_{(1)}^+(H)$ , it suffices to show that  $\forall \epsilon > 0$

$$P[\|A - A_0\|_{L_{(1)}} < \epsilon] > 0,$$

where  $A$  denotes the random operator above.

1. We first show that the above statement holds for  $A_0$  being a finite rank operator in  $L_{(1)}^+(H)$  with rank  $p$ . Since  $A_0 \in L_{(1)}^+(H)$ , there exists orthonormal eigenfunctions  $\{e_j\}_{j=1}^p$  and eigenvalues  $\{b_j\}_{j=1}^p$  such that

$$A_0 e_k = b_k e_k,$$

where  $b_j > 0, \forall j$ . Then we can write  $A_0 = \sum_{j=1}^p b_j e_j \otimes e_j$  and write

$$\begin{aligned} P[\|A - A_0\|_{L_{(1)}} < \epsilon] &= P[\|\sum_{j=1}^{\infty} w_j Z_j \otimes Z_j - \sum_{j=1}^p b_j e_j \otimes e_j\|_{L_{(1)}} < \epsilon] \\ &\geq P[\|\sum_{j=n+1}^{\infty} w_j Z_j \otimes Z_j\|_{L_{(1)}} < \frac{\epsilon}{3}] \cdot \prod_{j=p+1}^n P[\|w_j Z_j \otimes Z_j\|_{L_{(1)}} < \frac{\epsilon}{3(n-p)}] \\ &\quad \cdot \prod_{j=1}^p P[\|w_j Z_j \otimes Z_j - b_j e_j \otimes e_j\|_{L_{(1)}} < \frac{\epsilon}{3p}] \end{aligned} \tag{6.9}$$

Now we show that all the three factors in (6.9) are strictly positive.

(a)

$$\begin{aligned} P[\|\sum_{j=n+1}^{\infty} w_j Z_j \otimes Z_j\|_{L_{(1)}} < \frac{\epsilon}{3}] &= 1 - P[\|\sum_{j=n+1}^{\infty} w_j Z_j \otimes Z_j\|_{L_{(1)}} \geq \frac{\epsilon}{3}] \\ &\geq 1 - \frac{3}{\epsilon} E[\|\sum_{j=n+1}^{\infty} w_j Z_j \otimes Z_j\|_{L_{(1)}}] \\ &= 1 - \frac{3}{\epsilon} E[\sum_{j=n+1}^{\infty} w_j |Z_j|^2]. \end{aligned}$$

Since  $\sum_{j=1}^{\infty} w_j |Z_j|^2$  converge a.s., we can take  $n > p$  and  $n$  large enough so that  $\sum_{j=n+1}^{\infty} w_j |Z_j|^2 < \epsilon/6$ . Therefore,

$$P \left[ \|\sum_{j=n+1}^{\infty} w_j Z_j \otimes Z_j\|_{L_{(1)}} < \frac{\epsilon}{3} \right] > 1 - \frac{3}{\epsilon} \frac{\epsilon}{6} = \frac{1}{2} > 0$$

(b) For  $p < j \leq n$ ,

$$\begin{aligned} P[\|w_j Z_j \otimes Z_j\|_{L(1)} < \frac{\epsilon}{3(n-p)}] &= P\left[w_j |Z_j|^2 < \frac{\epsilon}{3(n-p)}\right] \\ &= P\left[|Z_j|^2 < \frac{\epsilon}{3w_j(n-p)}\right] > 0 \end{aligned}$$

by the fact that  $Z_j$  is fully supported on the whole Hilbert space  $H$  (Proposition 6.2.4).

(c) For  $j \leq p$ , we show that the map from  $(\sqrt{w_j/b_j} Z_j - e_j, |\cdot|)$  to  $(w_j/b_j Z_j \otimes Z_j - e_j \otimes e_j, \|\cdot\|_{L(1)})$  is continuous so that  $\forall \epsilon/(3p) > 0$ , there exists  $\delta > 0$  such that

$$|\sqrt{w_j} Z_j - \sqrt{b_j} e_j| < \delta \implies \|w_j Z_j \otimes Z_j - b_j e_j \otimes e_j\|_{L(1)} < \frac{\epsilon}{3p}.$$

Let  $\tilde{Z} = \sqrt{w_j/b_j} Z_j$ , and let  $(e_i)$  be the orthonormal basis of  $H$  extended

from the eigen-basis of  $B$ , then

$$\begin{aligned}
& \|w_j Z_j \otimes Z_j - b_j e_j \otimes e_j\|_{L^{(1)}} = \sum_{i=1}^{\infty} |\langle w_j Z_j \otimes Z_j - b_j e_j \otimes e_j, e_i \rangle| \\
&= \sum_{i=1}^{\infty} |w_j \langle Z_j, e_i \rangle^2 - b_j \langle e_j, e_i \rangle^2| = b_j \sum_{i=1}^{\infty} |\langle \tilde{Z}, e_i \rangle^2 - \langle e_j, e_i \rangle^2| \\
&= b_j \left( \sum_{i \neq j} \langle \tilde{Z}, e_i \rangle^2 + |\langle \tilde{Z}, e_j \rangle^2 - \langle e_j, e_j \rangle^2| \right) \\
&= b_j \left( |\tilde{Z} - \langle \tilde{Z}, e_j \rangle e_j|^2 + |\langle \tilde{Z}, e_j \rangle + \langle e_j, e_j \rangle| |\langle \tilde{Z}, e_j \rangle - \langle e_j, e_j \rangle| \right) \\
&\leq b_j \left( |\tilde{Z} - e_j|^2 + |\langle \tilde{Z} + e_j, e_j \rangle| |\langle \tilde{Z} - e_j, e_j \rangle| \right) \\
&\leq b_j \left( |\tilde{Z} - e_j|^2 + (|\tilde{Z} - e_j| + |2e_j|) |\tilde{Z} - e_j| |e_j|^2 \right) \\
&\leq b_j \left( \frac{\delta^2}{b_j} + \left( \frac{\delta}{\sqrt{b_j}} + 2 \right) \frac{\delta}{\sqrt{b_j}} \right) \\
&= 2\delta^2 + 2\sqrt{b_j}\delta. \quad (\text{note } |\sqrt{w_j}Z_j - \sqrt{b_j}e_j| < \delta \implies |\tilde{Z} - e_j| < \frac{\delta}{\sqrt{b_j}})
\end{aligned}$$

Therefore, we can let  $\delta$  be small enough so that  $\|w_j Z_j \otimes Z_j - b_j e_j \otimes e_j\|_{L^{(1)}} < \frac{\epsilon}{3p}$ . Hence

$$P[\|w_j Z_j \otimes Z_j - b_j e_j \otimes e_j\|_{L^{(1)}} < \frac{\epsilon}{3p}] > P[|\sqrt{w_j}Z_j - \sqrt{b_j}e_j| < \delta] > 0,$$

by the fact that  $Z_j$  is fully supported on the whole  $H$  space.

In summary, (a)-(c) show that all components in (6.9) are strictly positive.

Thus the theorem has been proved for  $A_0$  being finite rank.

2. If  $A_0$  is not finite rank, the set of finite rank  $L_{(1)}^+(H)$ -operators is dense in  $L_{(1)}^+(H)$ ,  $\forall \epsilon > 0$ , so the  $\epsilon$ -neighborhood of  $B$  contains at least one finite rank

operator, say  $A_k$ . Let  $\|A_0 - A_k\|_{L_{(1)}} = r$ , then

$$\{\|A - A_k\|_{L_{(1)}} < \frac{r}{2}\} \subset \{\|A - A_0\|_{L_{(1)}} < \epsilon\}.$$

Hence  $P\{\|A - A_0\|_{L_{(1)}} < \epsilon\} > P\{\|A - A_k\|_{L_{(1)}} < \frac{r}{2}\} > 0$ . By 1. and 2., we have shown that the random operator  $A = \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j$  is fully supported on the whole space of  $L_{(1)}^+(H)$ .

□

## 6.4 A Markov Chain Monte Carlo

In this section, we restrict the separable Hilbert space  $H$  to be  $L^2(T)$  where  $T = [0, 1]$ . A random element  $X$  taking values in  $H$  is called a stochastic process and is usually denoted by  $X(t)$ . Suppose there are  $n$  such random processes  $\{X_i(t)\}_{i=1}^n$ , which are i.i.d. with Gaussian measure  $\mu_m, \Sigma$ , where  $m$  is the mean and  $\Sigma$  is the covariance operator such that  $\text{Ker}(\Sigma) = \{0\}$ . We construct a prior for  $\Sigma$  using the expansion in (6.4). The likelihood and priors are:

$$X_i(t) \mid m, \Sigma \sim \mu_{m, \Sigma}, \quad (6.10)$$

$$m \mid \Sigma \sim \mu_{0, k\Sigma}, \quad (6.11)$$

$$\Sigma = c \sum_{j=1}^{\infty} w_j Z_j \otimes Z_j, \quad (6.12)$$

$$Z_j(t) \sim \mu_{0, B}, \quad (6.13)$$

$$c \sim \text{Inv-}\chi^2(d_a, d_b). \quad (6.14)$$

Here  $k$ ,  $B$ ,  $d_a$  and  $d_b$  are pre-specified prior parameters, and  $\text{Inv-}\chi^2(\nu, s^2)$  represents a scaled inverse chi-square distribution with density function:

$$f(x; \nu, s^2) = \frac{(s^2 \nu / 2)^{\nu/2}}{\Gamma(\nu/2)} x^{-\nu/2-1} \exp\left\{-\frac{\nu s^2}{2x}\right\}. \quad (6.15)$$

Note that an  $\text{Inv-}\chi^2(\nu, s^2)$  is equivalent to an Inv-Gamma with  $(\nu/2, \nu s^2/2)$ . Here  $d_a = \nu, d_b = s^2$ . We assume that  $Z_j(t)$ 's are independent Gaussian with zero mean and known covariance operator  $B$ . The operation  $Z_j \otimes Z_j$  is defined as

$$(Z_j \otimes Z_j)h(t) = Z_j(t)\langle Z_j(s), h(s) \rangle = Z_j(t) \int_T Z_j(s)h(s)ds. \quad (6.16)$$

We also assume that the scaling parameter  $c$  is independent of  $Z_j(t)$ 's.

The posterior inference based on the above likelihood and priors can be conducted using finite dimensional projection, which is discussed in detail in the following section.

#### 6.4.1 Derivations of the Posterior Distribution

Based on the likelihood and prior settings from (6.10) to (6.14), we can do posterior inference by projecting  $X_i(t)$ 's on a finite grid  $T_p = (t_1, \dots, t_p)^T$ . Denote the discretized version of  $X_i(t)$  as  $\vec{X}_i = (\vec{X}_i(t_1), \dots, \vec{X}_i(t_p))^T$ ,  $\vec{X}$  provides an approximation for  $X_i(t)$  as  $p$  approaches infinity. After discretization, the covariance operator  $\Sigma$  becomes a  $p$  by  $p$  covariance matrix  $\vec{\Sigma}$ , and the likelihood in (6.10) becomes a multivariate normal with density

$$\pi(\vec{X} | \vec{m}, \vec{\Sigma}) \propto |\vec{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (\vec{X}_i - \vec{m})^T \vec{\Sigma}^{-1} (\vec{X}_i - \vec{m})\right\}, \quad (6.17)$$

where  $\vec{X} = (\vec{X}_1, \dots, \vec{X}_n)^T$  and  $\vec{m}$  is the discretized mean. The expansion in (6.12) can be approximated by first projecting  $Zj(t)$ 's on  $T_p$ , then truncating the infinite sum at a fixed number  $J$ . According GRIP in Section 6.1, if we let  $J \rightarrow \infty$  and  $p \rightarrow \infty$ , then our posterior should converge to the “functional posterior” obtained from (6.10)-(6.14). We write the approximated version of the priors in (6.11)-(6.13) as follows:

$$\begin{aligned}\vec{m} \mid \vec{\Sigma} &\sim N(0, k\vec{\Sigma}), \\ \vec{\Sigma} &= c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T, \\ \vec{Z}_j &\sim N(0, \vec{B}).\end{aligned}\tag{6.18}$$

Here  $N(\cdot, \cdot)$  represents multivariate normal distribution. After finite dimensional projections of the likelihood and priors, we obtain the posterior in multivariate form as follows:

$$\begin{aligned}\pi(\vec{m}, \vec{\Sigma} \mid \vec{X}) &\propto \pi(\vec{X} \mid \vec{m}, \vec{\Sigma}) \pi(\vec{m} \mid \vec{\Sigma}) \pi(\vec{\Sigma}) \\ &\propto |\vec{\Sigma}|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\vec{X}_i - \vec{m})^T \vec{\Sigma}^{-1} (\vec{X}_i - \vec{m}) \right\} \\ &\quad \cdot |k\vec{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \vec{m}^T (k\vec{\Sigma})^{-1} \vec{m} \right\} \pi(\vec{\Sigma}).\end{aligned}\tag{6.19}$$

We now integrate out  $\vec{m}$  from the above joint posterior to obtain the marginal posterior of  $\vec{\Sigma}$  by  $\pi(\vec{\Sigma} \mid \vec{X}) = \int \pi(\vec{m}, \vec{\Sigma} \mid \vec{X}) d\vec{m}$ . Note that we can reformulate the quadratic terms in (6.19) to get:

$$\begin{aligned}\pi(\vec{m}, \vec{\Sigma} \mid \vec{X}) &\propto |\vec{\Sigma}|^{-\frac{n}{2}} |k\vec{\Sigma}|^{-\frac{1}{2}} \pi(\vec{\Sigma}) \\ &\quad \cdot \exp \left\{ -\frac{1}{2} \left[ \vec{m}^T \left( n\vec{\Sigma}^{-1} + (k\vec{\Sigma})^{-1} \right) \vec{m} - 2\vec{m}^T \vec{\Sigma}^{-1} \left( \sum_{i=1}^n \vec{X}_i \right) + \sum_{i=1}^n \vec{X}_i^T \vec{\Sigma}^{-1} \vec{X}_i \right] \right\}.\end{aligned}$$

Let  $K_1 = n\vec{\Sigma}^{-1} + (k\vec{\Sigma})^{-1} = (n + \frac{1}{k})\vec{\Sigma}^{-1}$  and  $M_1 = \vec{\Sigma}^{-1} \left( \sum_{i=1}^n \vec{X}_i \right)$ , the multivariate normal density in the above expression can be split as:

$$\begin{aligned} \pi(\vec{m}, \vec{\Sigma} | \vec{X}) &\propto |K_1^{-1}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\vec{m} - K_1^{-1} M_1)^T K_1 (\vec{m} - K_1^{-1} M_1) \right\} \\ &\cdot |K_1^{-1}|^{\frac{1}{2}} |\vec{\Sigma}|^{-\frac{n}{2}} |k\vec{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \vec{X}_i^T \vec{\Sigma}^{-1} \vec{X}_i + \frac{1}{2} M_1^T K_1^{-1} M_1 \right\} \pi(\vec{\Sigma}). \end{aligned} \quad (6.20)$$

The first two factors can be integrated out w.r.t.  $\vec{m}$  since they form a multivariate normal density. This gives the resulting marginal posterior as

$$\pi(\vec{\Sigma} | \vec{X}) \propto |K_1^{-1}|^{\frac{1}{2}} |\vec{\Sigma}|^{-\frac{n}{2}} |k\vec{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \vec{X}_i^T \vec{\Sigma}^{-1} \vec{X}_i + \frac{1}{2} M_1^T K_1^{-1} M_1 \right\} \pi(\vec{\Sigma}).$$

The above form can be further simplified by combining the  $|\vec{\Sigma}|$  terms, drop the constant terms, and using the simplified terms of

$$K_1 = (n + \frac{1}{k})(\vec{\Sigma})^{-1}$$

and

$$M_1^T K_1^{-1} M_1 = (n + \frac{1}{k})^{-1} \left( \sum_{i=1}^n \vec{X}_i \right)^T \vec{\Sigma}^{-1} \left( \sum_{i=1}^n \vec{X}_i \right).$$

The simplified form is

$$\begin{aligned} \pi(\vec{\Sigma} | \vec{X}) &\propto \\ &\exp \left\{ -\frac{1}{2} \sum_{i=1}^n \vec{X}_i^T \vec{\Sigma}^{-1} \vec{X}_i + \frac{1}{2} (n + \frac{1}{k})^{-1} \left( \sum_{i=1}^n \vec{X}_i \right)^T \vec{\Sigma}^{-1} \left( \sum_{i=1}^n \vec{X}_i \right) \right\} |\vec{\Sigma}|^{-\frac{n}{2}} \pi(\vec{\Sigma}). \end{aligned} \quad (6.21)$$

Note that in (6.21) the prior for  $\vec{\Sigma}$  has not been given a particular form yet. If we write  $\vec{Z} = (\vec{Z}_1, \vec{Z}_2, \dots, \vec{Z}_J)^T$ , according to the prior assumption in (6.18),  $\vec{\Sigma}$  is a

deterministic function of  $c$  and  $\vec{Z}$ . Thus  $\vec{\Sigma}$  can be replaced by  $c$  and  $\vec{Z}$  in the likelihood and the conditional prior  $\pi(\vec{m}|\vec{\Sigma})$ . Instead of setting up prior for  $\vec{\Sigma}$ , we set up priors for  $\vec{Z}$  as

$$\begin{aligned}\pi(\vec{Z}) &= \pi(c) \prod_{j=1}^J \pi(\vec{Z}_j) \\ &\propto c^{-d_a/2-1} \exp\left\{-\frac{d_a d_b}{2c}\right\} \exp\left\{-\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j\right\}\end{aligned}$$

The posterior samples of  $\vec{Z}$  can be used to construct samples for  $\vec{\Sigma}$ . To get the joint posterior distribution for  $c$  and  $\vec{Z}$ , we just need to replace  $\pi(\vec{\Sigma})$  by  $\pi(\vec{Z})$  in (6.21), and replace other terms of  $\vec{\Sigma}$  by the linear expansion in (6.18), which gives

$$\begin{aligned}\pi(c, \vec{Z} | \vec{X}) &\propto \left| c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right|^{-\frac{n}{2}} c^{-d_a/2-1} \exp\left\{-\frac{d_a d_b}{2c}\right\} \exp\left\{-\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j\right\} \\ &\quad \cdot \exp\left\{-\frac{1}{2} \sum_{i=1}^n \vec{X}_i^T \left( c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \vec{X}_i\right\} \\ &\quad \cdot \exp\left\{\frac{1}{2} \left(n + \frac{1}{k}\right)^{-1} \left(\sum_{i=1}^n \vec{X}_i\right)^T \left( c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \left(\sum_{i=1}^n \vec{X}_i\right)\right\}\end{aligned}\quad (6.22)$$

The above posterior distribution can be simplified once more by integrating  $c$  out.

Separate all the terms containing  $c$ :

$$\begin{aligned}\pi(c, \vec{Z} | \vec{X}) &\propto c^{-\frac{np+d_a}{2}-1} \exp\left\{-\frac{\tilde{A} - \tilde{B} + d_a d_b}{2c}\right\} \\ &\quad \cdot \left| \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j\right\},\end{aligned}\quad (6.23)$$

where

$$\tilde{A} = \sum_{i=1}^n \vec{X}_i^T \left( \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \vec{X}_i \quad (6.24)$$

and

$$\tilde{B} = \left( n + \frac{1}{k} \right)^{-1} \left( \sum_{i=1}^n \vec{X}_i \right)^T \left( \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \left( \sum_{i=1}^n \vec{X}_i \right). \quad (6.25)$$

The first two factors in (6.23) indicate that, conditional on  $\vec{Z}$ ,  $c$  is Inv- $\chi^2(v1, v2)$ , where  $v1 = np + d_a$  and  $v2 = (\tilde{A} - \tilde{B} + d_a d_b)/v1$ . Therefore we can integrate them out, which gives:

$$\begin{aligned} \pi(\vec{Z}|\vec{X}) &\propto \left( \frac{\tilde{A} - \tilde{B} + d_a d_b}{np + d_a} \right)^{-\frac{np+d_a}{2}} \left| \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j \right\} \\ &\propto \left( \tilde{A} - \tilde{B} + d_a d_b \right)^{-\frac{np+d_a}{2}} \left| \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j \right\} \end{aligned} \quad (6.26)$$

Based on the above posterior distribution, we describe our MCMC algorithm below.

In this algorithm,  $N$  is a pre-defined maximum number of iterations,  $i$  is the iteration index and we write  $\theta^{(i)}$  as the posterior sample for parameter  $\theta$  in iteration  $i$ .

Step 0. Set initial values for  $\vec{Z}_j, j = 1, \dots, J$ .

For  $i = 1, \dots, N$ , run Step 1-3.

Step 1. Conditional on  $\vec{X}$ , update  $\vec{Z} = (\vec{Z}_1, \dots, \vec{Z}_1)^T$ . For  $j = 1, \dots, J$ , sample a new observation from the proposal distribution  $\vec{Z}_j^* \sim N(\vec{Z}_j^{(i-1)}, \delta I)$ , and calculate

$$r = \frac{\pi(\vec{Z}_1^{(i)}, \dots, \vec{Z}_{j-1}^{(i)}, \vec{Z}_j^*, \vec{Z}_{j+1}^{(i-1)}, \dots, \vec{Z}_J^{(i-1)} | \vec{X})}{\pi(\vec{Z}_1^{(i)}, \dots, \vec{Z}_{j-1}^{(i)}, \vec{Z}_j^{(i-1)}, \vec{Z}_{j+1}^{(i-1)}, \dots, \vec{Z}_J^{(i-1)} | \vec{X})}.$$

Note that the numerator and the denominator can be computed using (6.26). Update  $\vec{Z}_j^{(i)} = \vec{Z}_j^*$  with probability  $\min(1, r)$ .

Step 2. Conditional on  $\vec{Z}^{(i)}$  and  $\vec{X}$ , sample  $c^{(i)}$  from  $\pi(c|\vec{Z}, \vec{X}) = \text{inv}-\chi^2(v1, v2)$ , where  $v1 = np + d_a$  and  $v2 = (\tilde{A} - \tilde{B} + d_a d_b)/v1$ , where  $\tilde{A}$  and  $\tilde{B}$  are defined in (6.24) and (6.25).

Step 3. Conditional on  $\vec{Z}^{(i)}, c^{(i)}$  and  $\vec{X}$ , sample  $\vec{m}^{(i)}$  from  $N(\mu_0, V_0)$  distribution, where

$$\mu_0 = K_1^{-1} M_1 = \left(n + \frac{1}{k}\right)^{-1} \left( \sum_{i=1}^n \vec{X}_i \right)$$

and

$$V_0 = K_1^{-1} = \left(n + \frac{1}{k}\right)^{-1} \vec{\Sigma} = \left(n + \frac{1}{k}\right)^{-1} c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T.$$

This conditional distribution can be easily observed from the joint distribution (6.20).

### 6.4.2 Notes on Some Computational Tricks

This section collects some computational tricks which are helpful to improve the MCMC algorithm. We focus on the posterior distribution derived in (6.26). Since the variables are all in discretized form, for simplicity, we remove the vector symbol (the arrow on top of a variable), i.e.,  $X$  and  $Z$  are the same as  $\vec{X}$  and  $\vec{Z}$  defined in Section 6.4.1. Note that  $X$  is a  $n \times p$  data matrix with each row a discretized functional observation, and  $Z$  is a  $J \times p$  matrix with the  $j$ th row being  $Z_j^T$ .

**NOTE 1.** Let  $W = \text{diag}(\sqrt{w_1}, \dots, \sqrt{w_J})$ ,

$$WZ = \begin{pmatrix} \sqrt{w_1} & & & \\ & \sqrt{w_2} & & \\ & & \ddots & \\ & & & \sqrt{w_J} \end{pmatrix} \begin{pmatrix} Z_1^T \\ Z_2^T \\ \vdots \\ Z_J^T \end{pmatrix} = \begin{pmatrix} \sqrt{w_1} Z_1^T \\ \sqrt{w_2} Z_2^T \\ \vdots \\ \sqrt{w_J} Z_J^T \end{pmatrix}.$$

Therefore  $\sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T = (WZ)^T (WZ)$ . In real computation, this is done by performing QR decomposition for  $WZ$  so that  $WZ = QR$ , where  $Q$  is a matrix with orthonormal columns and  $R$  is a upper triangular matrix. Note that such a decomposition always exists, see, for example, Trefethen and Bau [74]. Now the linear expansion becomes

$$\sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T = (WZ)'(WZ) = R' Q' Q R = R' R.$$

Hence the covariance expansion in (6.18) becomes  $\vec{\Sigma} = c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T = c R' R$ . Note that  $\sqrt{c}R$  is the Cholesky decomposition of the covariance matrix  $\vec{\Sigma}$ . In each iteration, the MCMC algorithm updates the rows of  $WZ$  one by one using the built-in functions *qrdelete* and *qrinsert* of MATLAB (The Math works, Inc., Natick, Mass., U.S.A.)

**NOTE 2.** For the factor  $\exp\left\{-\frac{1}{2} \sum_{j=1}^J \vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j\right\}$  in (6.26). To compute matrix inversion  $B^{-1}$  efficiently, we first perform cholesky decomposition for  $B$ , i.e.  $R_1^T R_1 = B$  for some upper triangular matrix  $R_1$ . Then  $B^{-1}$  is obtained by

$$B^{-1} = (R_1^T R_1)^{-1} = R_1^{-1} (R_1^T)^{-1} = R_1^{-1} (R_1^{-1})^T.$$

So  $\vec{Z}_j^T \vec{B}^{-1} \vec{Z}_j = ((R_1^{-1})^T Z_j)^T ((R_1^{-1})^T Z_j)$ .

**NOTE 3.** For the term

$$\sum_{i=1}^n \vec{X}_i^T \left( c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \vec{X}_i,$$

we have seen from **NOTE 1** that the Cholesky decomposition of the middle term  $c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T$  is  $\sqrt{c}R$ . Let  $T = ((\sqrt{c}R)^{-1})^T X^T$ , and write  $T = (T_1, \dots, T_n)$ , where  $T_i$  are the columns of  $T$ , we have

$$\sum_{i=1}^n \vec{X}_i^T \left( c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \vec{X}_i = \sum_i T_i^T T_i = \text{Trace}(T^T T).$$

For the same  $T$ ,

$$\left( \sum_{i=1}^n \vec{X}_i \right)^T \left( c \sum_{j=1}^J w_j \vec{Z}_j \vec{Z}_j^T \right)^{-1} \left( \sum_{i=1}^n \vec{X}_i \right) = (\sum_i T_i)^T (\sum_i T_i).$$

### 6.4.3 Simulation Results

Based on the prior proposed in Section 6.3 and the MCMC algorithm proposed in Section 6.4, we conduct a simulation study in this section. Our data come from  $n = 50$  Brownian Motion paths on a time grid of  $[0, 1]$ , with the number grid points  $p = 60$ . Note that the covariance function of the Brownian Motion is  $K(s, t) = \min(s, t), s, t \in [0, 1]$ . Figure 6.1 shows the plot of the sample paths and Figure 6.2 shows the corresponding true covariance function.

The proposed MCMC in Section 6.4 is applied to the simulated data, with 10000 iterations and a 4000 burn-in period. We set the parameters in the priors and other related parameters to be:  $k = 100$ ,  $d_a = 4.01$ ,  $d_b = 10$ ,  $\delta = 0.005$ ,  $J = 150$ . Initial

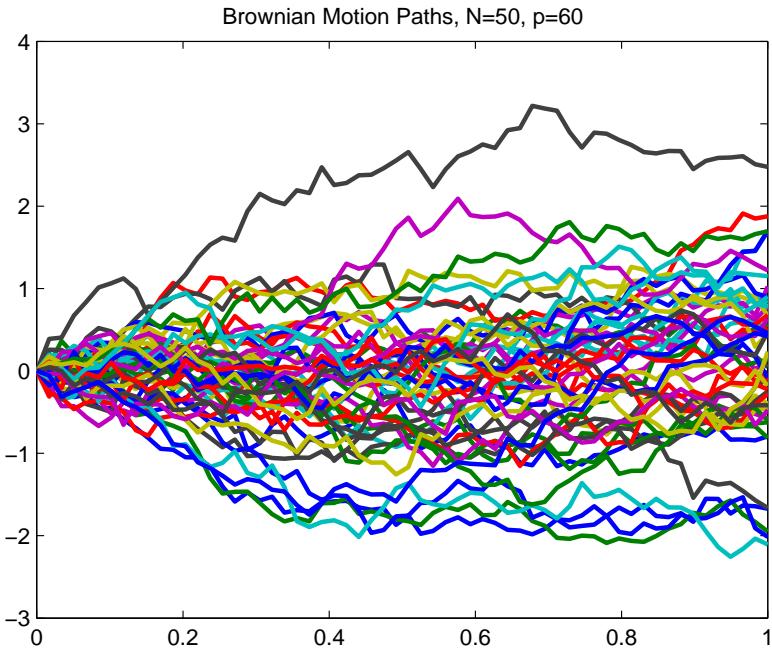


Figure 6.1: Plot of  $N = 50$  sample paths of a Brownian Motion.

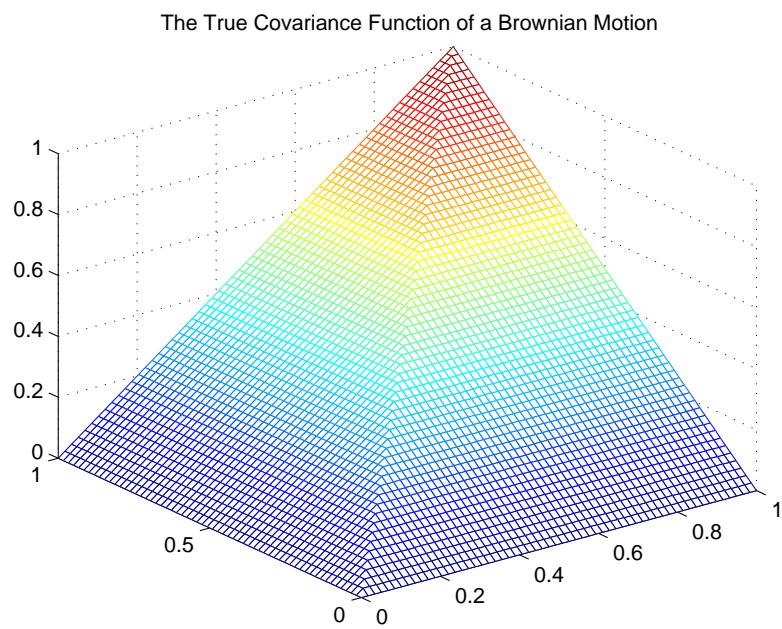


Figure 6.2: The true covariance function of Brownian Motion.

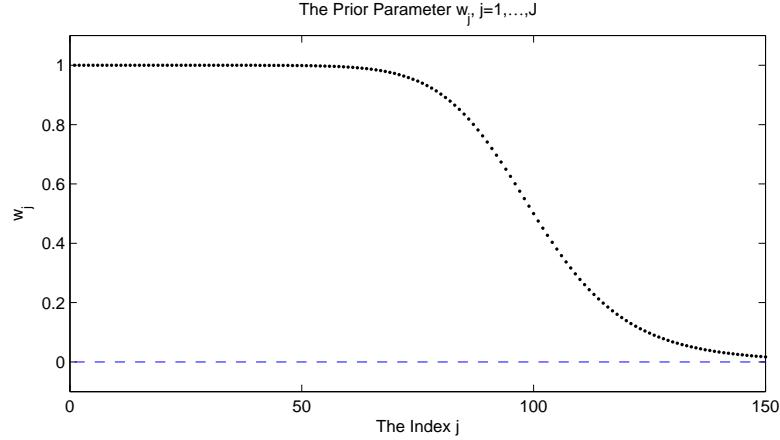


Figure 6.3: The plot of prior parameter  $w_j$ .

values of  $Z_j$ 's are generated from normal distribution with zero mean and identity covariance. For the weight  $w_j$ , we use the following form:

$$w_j = \left(1 + \left(\frac{j}{\alpha}\right)^q\right)^{-1}, j = 1, \dots, J,$$

where  $\alpha = 100, q = 10$  for this simulation study. The values of  $w_j$ 's are plotted in Figure 6.3. The prior covariance for  $Z_j$  is set to be

$$B(t_i, t_j) = \exp(-0.6145|t_i - t_j|^{1/2}). \quad (6.27)$$

Figure 6.4 shows the plot of the prior covariance  $B$ . We use the posterior  $Z_j$  samples in each iteration to compute the posterior  $\vec{\Sigma}$  samples, and average the posterior samples of  $\vec{\Sigma}$  to obtain the final estimate. Figure 6.5 plots the posterior sample average of  $\vec{\Sigma}$ . The trace plot of the posterior samples of  $c$ , together with its histogram, is shown in Figure 6.6. Figure 6.7 shows the posterior mean of  $\vec{m}$  and its 95% credible interval. The acceptance rates of the  $Z_j$ 's is between 22% and 39%.

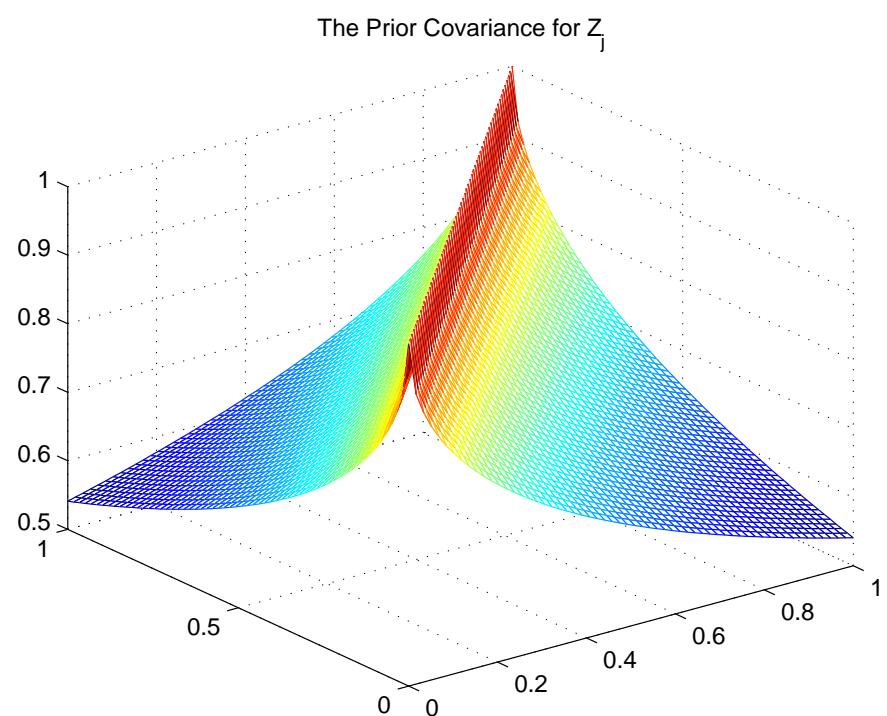


Figure 6.4: The prior covariance function for  $Z_j, j = 1, \dots, J$ .

The Posterior Mean Estimate of the Covariance Using the Proposed Prior

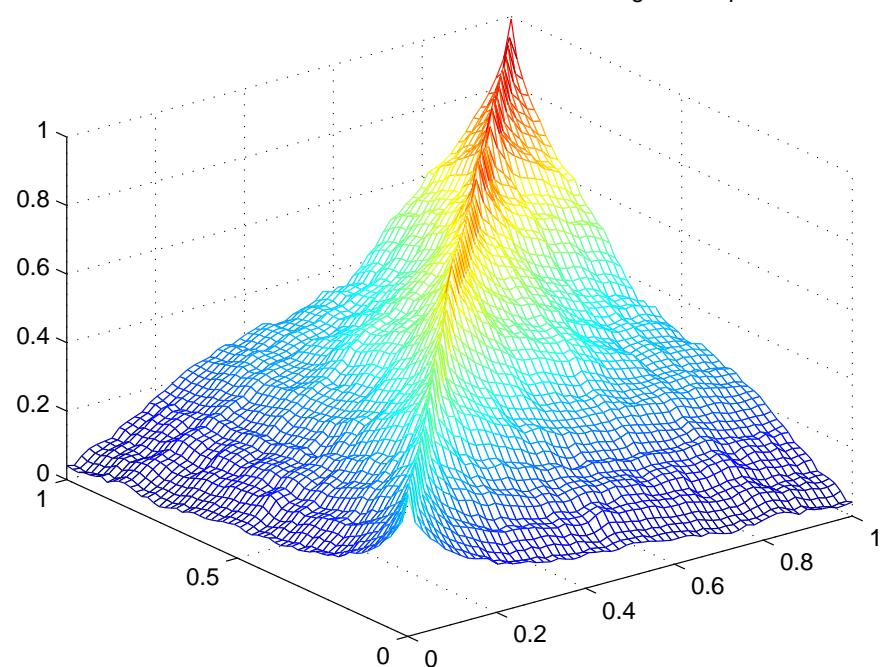


Figure 6.5: The posterior mean of the covariance function using the proposed prior.

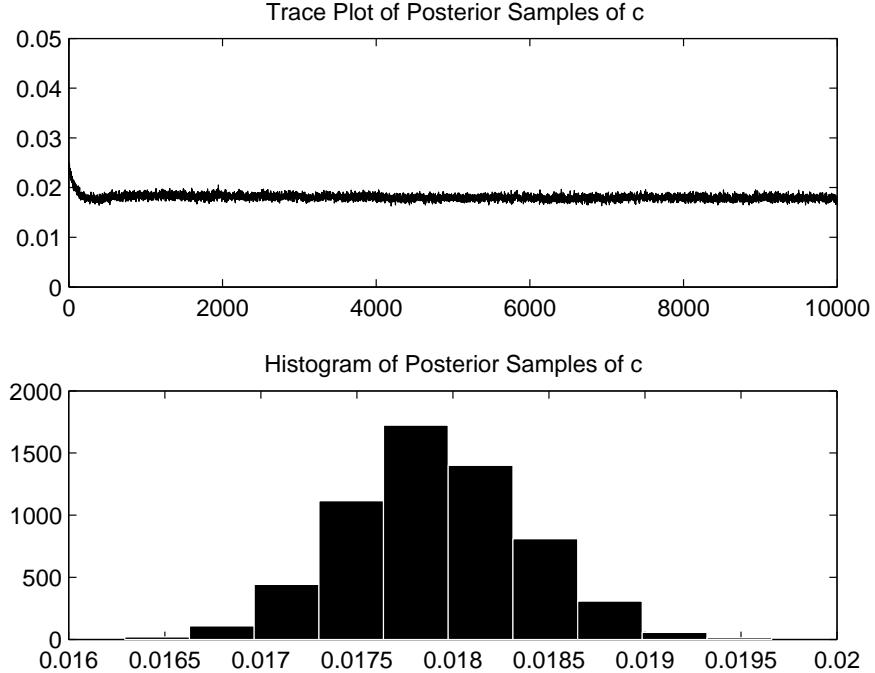


Figure 6.6: The trace plot of the posterior samples of  $c$  and its histogram.

To compare the estimated covariance function with the true, we use two metrics for measuring the estimation error. One is the averaged squared-error(ASE) defined by

$$ASE(\hat{\Sigma}, \Sigma) = \frac{1}{p^2} \sum_i \sum_j (\hat{\sigma}_{ij}^2 - \sigma_{ij}^2)^2, \quad (6.28)$$

where  $\hat{\sigma}_{ij}, \sigma_{ij}$  is the  $(i, j)$ th component of the estimated and true covariance matrix, respectively. The second metric is called the averaged absolute error (AL1E) defined by

$$AL1E(\hat{\Sigma}, \Sigma) = \frac{1}{p^2} \sum_i \sum_j |\hat{\sigma}_{ij}^2 - \sigma_{ij}^2|. \quad (6.29)$$

Table 6.1 lists the estimation error coming from the Bayesian estimate using the prior

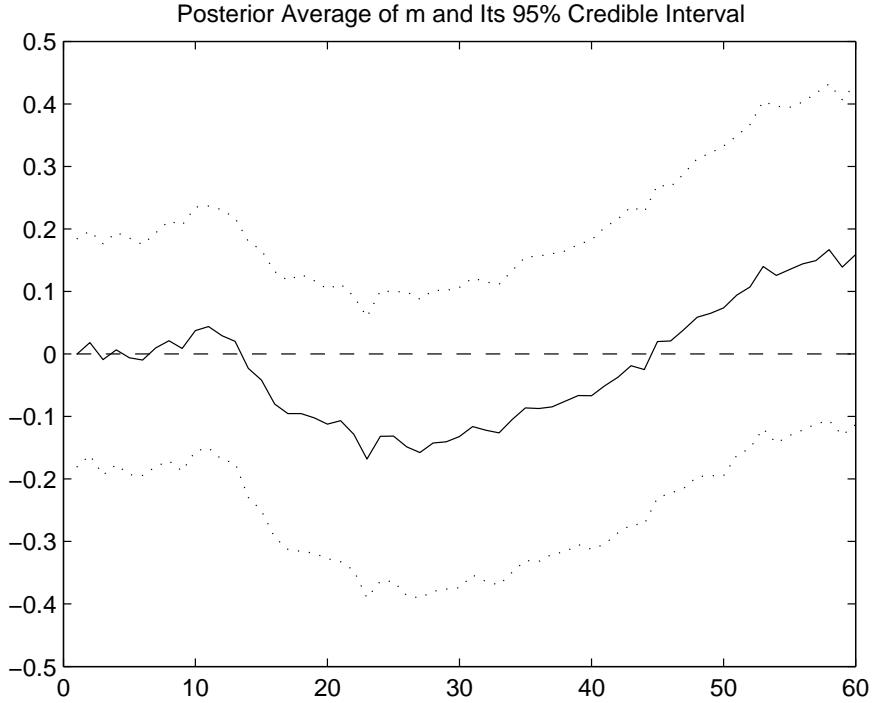


Figure 6.7: The posterior mean of the mean function  $\mu(t)$  and its 95% credible interval.

proposed in Section 6.3, and from the sample estimate, where the sample estimate  $\hat{\Sigma}_{\text{sample}}$  is obtained by

$$\frac{1}{n-1} \sum_{i=1}^n (\vec{X}_i - \bar{\vec{X}})(\vec{X}_i - \bar{\vec{X}})^T. \quad (6.30)$$

We see from Table 6.1 that, using the suggested MCMC algorithm, the Bayes estimate based on the proposed prior gives slightly smaller error than the sample estimate. More details of the estimation error are illustrated in Figure 6.8 and Figure 6.9, where we plot  $(\hat{\sigma}_{ij}^2 - \sigma_{ij}^2)$  at all  $(i, j)$  pairs for both estimation methods.

Method	ASE	AL1E
Bayes estimate using proposed prior	0.0129	0.0881
Sample estimate	0.0193	0.1146

Table 6.1: The Estimation Error Comparison.

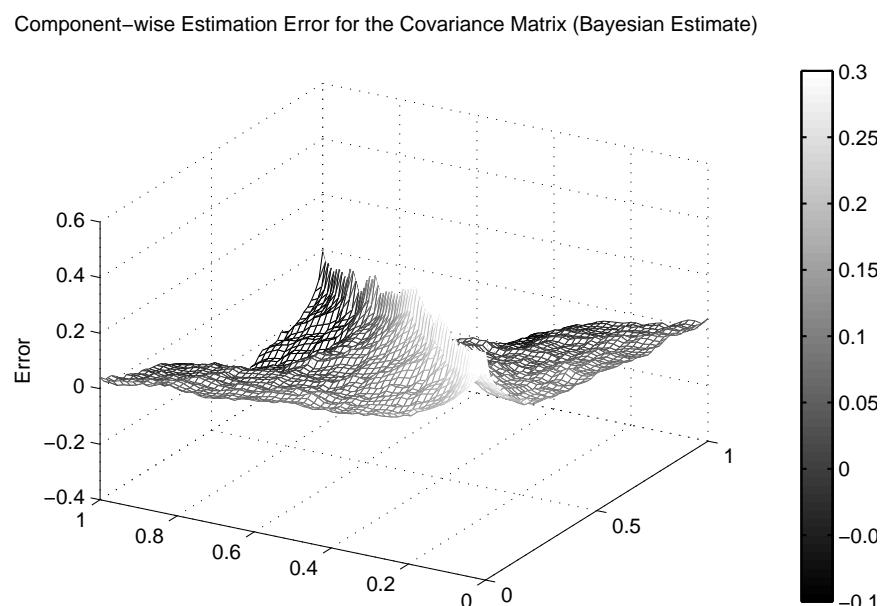


Figure 6.8: Plot of the Component-wise Estimation Error for the Covariance Matrix using the Bayesian Method.

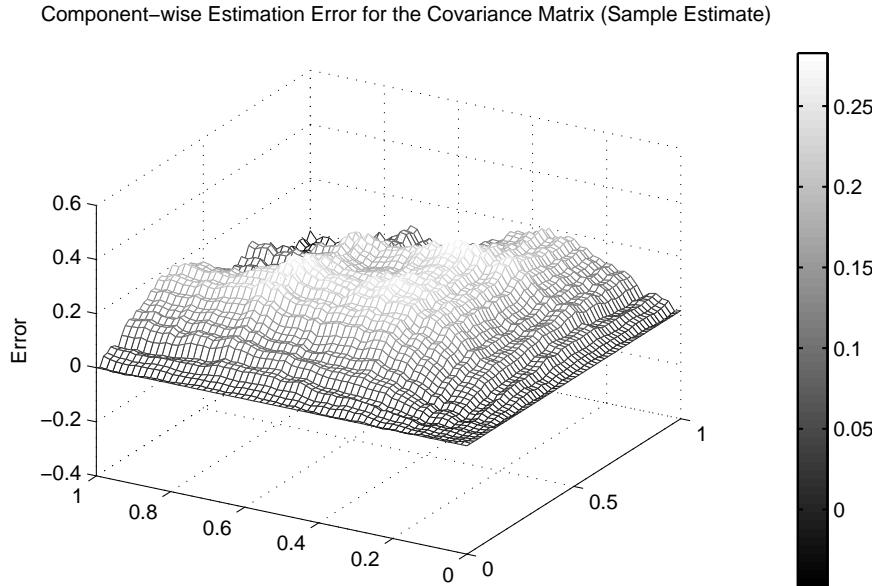


Figure 6.9: Plot of the Component-wise Estimation Error for the Covariance Matrix using the Sample Estimation Method.

## 6.5 Inverse-Wishart Prior and its limiting Behavior

### Prior

In order to look for appropriate priors for covariance operators in infinite-dimensional setup, we study the limiting behavior of Inverse-Wishart prior as the dimension (i.e., the number of grid points) goes to infinity. It is not clear whether there exists an infinite-dimensional counterpart of Inverse-Wishart distribution. We start from deriving the limits of the first two moments of multivariate Inverse-Wishart distribution in this section.

### 6.5.1 Definition and Some Facts about Wishart and Inverse Wishart distribution

We first give the definition of Wishart and Inverse Wishart distribution in multivariate setup.

- (1) Wishart distribution. Let  $\Sigma$  be a  $p$  by  $p$  positive definite and symmetric random matrix. We say  $\Sigma$  is of Wishart distribution with degree of freedom  $\nu$  and scale matrix  $S$ , and write  $\Sigma \sim \text{Wishart}(\nu, S)$ , if the pdf of  $\Sigma$  is

$$f(\Sigma|\nu, S) = \left( 2^{\nu p/2} \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\left(\frac{\nu+1-i}{2}\right) \right)^{-1} |S|^{-\nu/2} |\Sigma|^{(\nu-p-1)/2} \exp\left(-\frac{1}{2} \text{tr}(S^{-1}\Sigma)\right)$$

where  $\nu \geq p + 1$ ,  $S$  is positive definite and symmetric. It can be shown that  $E[\Sigma] = \nu S$ ,  $\text{mode}(\Sigma) = (\nu-p-1)S$  for  $\nu > p+1$ , and the characteristic function  $\phi(U) = E[\exp(i \cdot \text{tr}(\Sigma U))] = |I - 2iUS|^{-\nu/2}$ , where  $I$  and  $U$  are matrices of the same size of  $S$ .

- (2) Inverse-Wishart distribution. Suppose  $\Sigma$  is a  $p \times p$  positive definite random matrix,  $\Sigma \sim IW(\nu, S)$  with d.f.  $\nu$  and scaling matrix  $S$ , if

$$f(\Sigma|\nu, S) = \left( 2^{\nu p/2} \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\left(\frac{\nu+1-i}{2}\right) \right)^{-1} |S|^{\nu/2} |\Sigma|^{-(\nu+p+1)/2} \exp\left(-\frac{1}{2} \text{tr}(S\Sigma^{-1})\right).$$

It can be shown that  $E[\Sigma] = \frac{1}{\nu-p-1}S$ . (Note that some literature denote  $\Sigma \sim IW(\nu, S^{-1})$  for the same density stated above. The form of the density functions will be clear if one indicates the form of  $E[\Sigma]$ ).

(3) Relation of Wishart and Inverse Wishart distribution. If  $A \sim \text{Wishart}(\nu, S)$ , by change of variables, we can easily show that  $A^{-1} \sim \text{IW}(\nu, S^{-1})$  with  $E[A^{-1}] = \frac{1}{\nu-p-1}S^{-1}$ . Note: the Jacobian  $|\frac{\partial A}{\partial A^{-1}}| = |A|^{m+1}$ .

(4) Moments of Inverse-Wishart Matrix. Siskind [69] stated the following results about the general second-order moment of an Inverse-Wishart matrix: if  $t$  is a  $p \times 1$  constant vector,  $A$  is a  $p \times p$  Wishart matrix with  $\nu > p + 3$  degree of freedom and expectation  $\nu S$  (i.e.,  $A \sim \text{Wishart}(\nu, S)$ , by (3),  $A^{-1} \sim \text{IW}(\nu, S^{-1})$  with  $E[A^{-1}] = \frac{1}{\nu-p-1}S^{-1}$ ), so

$$(\nu - p)(\nu - p - 3)E[A^{-1}tt^T A^{-1}] = S^{-1}tt^TS^{-1} + S^{-1}(t^TS^{-1}t)/(\nu - p - 1),$$

$$\text{i.e., } E[A^{-1}tt^T A^{-1}] = \frac{S^{-1}tt^TS^{-1}}{(\nu-p)(\nu-p-3)} + \frac{S^{-1}(t^TS^{-1}t)}{(\nu-p-1)(\nu-p)(\nu-p-3)}.$$

### 6.5.2 Conjugate Inverse Wishart Priors for the Covariance in Multivariate Normal Model

Suppose  $X_i, i = 1, \dots, n$ , are i.i.d. normally distributed random vectors with unknown mean  $m$  and unknown variance matrix  $\Sigma$ . If we construct a Bayesian model as:

$$\begin{aligned}\pi(X_i|m, \Sigma) &= N(m, \Sigma), \\ \pi(m|\Sigma) &= N(m_0, 1/k_0\Sigma), \\ \pi(\Sigma) &= \text{IW}(\nu_0, \Lambda_0),\end{aligned}\tag{6.31}$$

the resulting posterior distribution

$$\pi(\Sigma|X_1, \dots, X_n) = \text{IW}(\tilde{\nu}, \tilde{\Lambda}),$$

where  $\tilde{\nu} = \nu_0 + n$ ,  $\tilde{\Lambda} = \Lambda_0 + S_n + \frac{n k_0}{n+k_0} (\bar{X} - m_0)(\bar{X} - m_0)^T$ ,  $S_n = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$ .

Therefore

$$E[\Sigma | X_1, \dots, X_n] = 1/(\tilde{\nu} - m - 1) \tilde{\Lambda}, \quad (6.33)$$

and conditional on  $\Sigma$ , we have  $\pi(m | \Sigma, X_1, \dots, X_n) = N(\tilde{m}, \tilde{V})$ , where  $\tilde{m} = \frac{n}{n+k_0} \bar{X} + \frac{k_0}{n+k_0} m_0$  and  $\tilde{V} = \frac{1}{n+k_0} \Sigma$ .

### 6.5.3 A Simulation Study using the Bayesian Model with Conjugate Inverse-Wishart Prior

In Section 6.4.3, we conducted a simulation to estimate the covariance of Brownian Motions using priors proposed in Section 6.3. In comparison, the simulation is repeated in this simulation by using the Bayesian model stated in Section 6.5.2. We set the scaling matrix  $\Lambda_0$  in (6.32) to be the prior matrix  $B$  used in (6.27), and set the other two prior parameters in (6.32) and (6.31) as  $\nu_0 = 65$  and  $k_0 = 1/10^6$ , respectively. The prior  $m_0$  is set to be a zero vector. For the same data generated in Section 6.4.3, we obtain 3000 posterior samples for  $\Sigma$  and use their average as the final estimate. Alternatively, since the posterior mean has an explicit form (as shown in (6.33)), we can also compute the posterior mean directly and use it as the estimate of  $\Sigma$ . In Table 6.2, the estimation errors defined by (6.28) and (6.29) are computed for both the sample average and the posterior mean. Comparing with Table 6.1, we find that for this simulated data, the estimation error obtained from Inverse-Wishart prior is very similar to that from the prior proposed in Section 6.3, and both estima-

Method	MSE	ML1E
Bayesian estimate (IW prior, based on 3000 sample)	0.0126	0.0929
Bayesian estimate (IW prior, the posterior mean)	0.0123	0.0917
Sample estimate	0.0193	0.1146

Table 6.2: The estimation error of the Bayes model with Inverse-Wishart prior compare with that of the sample estimate.

Posterior mean of the covariance using Inverse Wishart prior

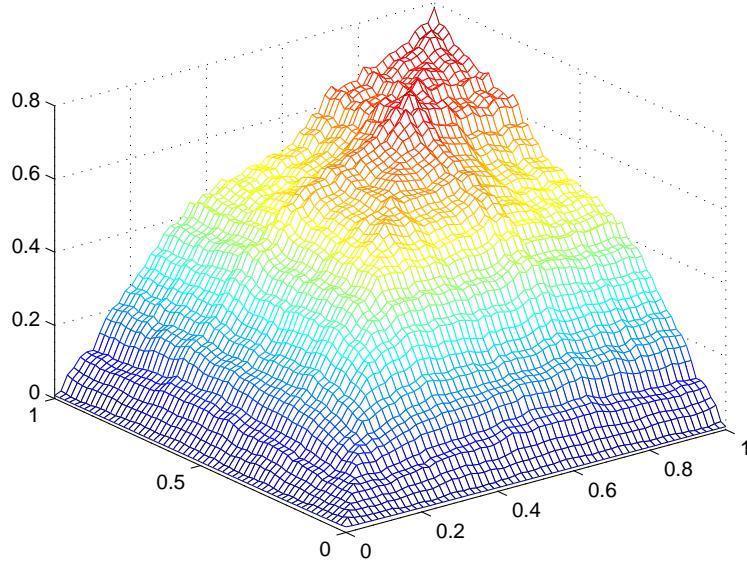


Figure 6.10: The posterior average of the covariance using the Bayesian model with an Inverse-Wishart prior.

tion methods give slightly smaller estimation error than the sample estimate. The posterior mean estimate is plotted in Figure 6.10.

As the number of grid points increases, the estimation error (defined in 6.28 and 6.29) using inverse Wishart prior is supposed to increase too. To show this, we generate  $n = 50$  Brownian Motion paths on  $[0, 1]$  but sample them at 3 different grid levels with the number of grid points: 11, 101 and 1000. Figure 6.11 plots the first Brownian Motion path at all three grid levels. We set the prior parameter  $\nu_0 = 2p$  and

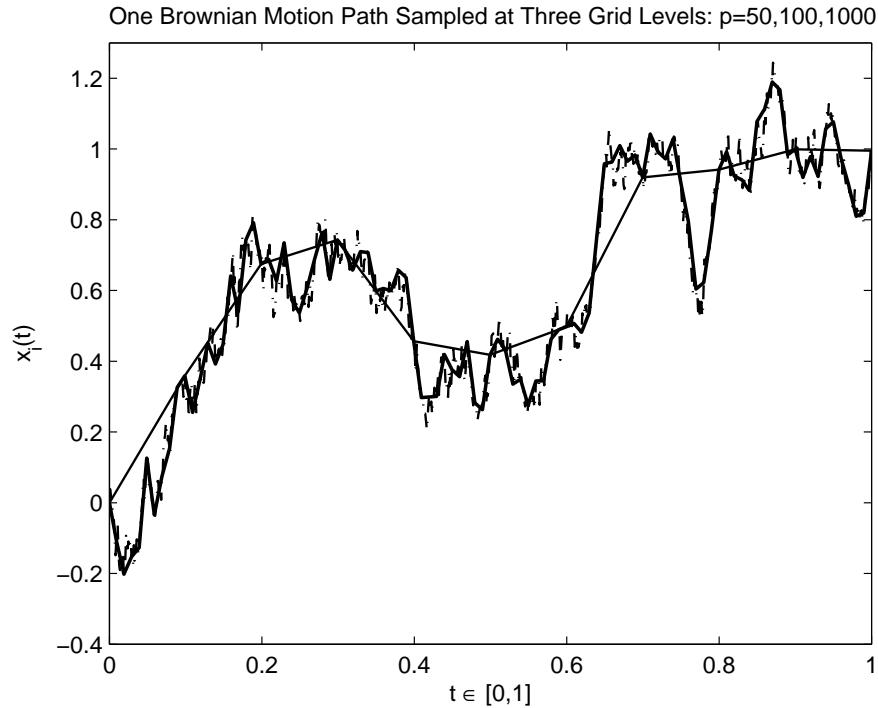


Figure 6.11: One Brownian Motion path sampled at three grid levels:  $p=50,100,1000$ .

$k_0 = 1/10^3$ . The estimation errors of the sample estimates and the Bayesian estimates with inverse-Wishart prior at all grid levels are listed in Table 6.3, which suggests that the estimating errors increase as  $p$  increases, and the estimating error of Bayes estimates (with Inverse-Wishart prior) increases faster than the sample estimates.

p	Bayes Est.(IW prior)		Sample Estimate	
	ASE	AL1E	ASE	AL1E
10	0.0023	0.0392	0.0048	0.0581
100	0.0314	0.1303	0.0057	0.0647
1000	0.1339	0.2965	0.0058	0.0655

Table 6.3: The estimation error comparison (IW prior) when the sampling grid gets finer.

### 6.5.4 Limiting Behavior of the First Two Moments of the Inverse-Wishart Distribution

Suppose that the covariance operator  $\Sigma \in L_{(1)}^+(H)$  and  $H = L^2[0, 1]$ . We have

$$\Sigma : L^2[0, 1] \longmapsto L^2[0, 1]$$

and for any  $f \in L^2[0, 1]$ ,

$$\Sigma f(s) = \int_0^1 k(s, t)f(t)dt,$$

where  $k(\cdot, \cdot)$  is the covariance kernel of  $\Sigma$ . Denote the discretized version of  $\Sigma$  on a finite grid as  $\vec{\Sigma}_p$ , which is a random matrix of size  $p$ , where  $p$  is the number of grid points. Our purpose is to find conditions such that, as  $p \rightarrow \infty$ , the limiting covariance operator maps any function on  $[0, 1]$  to some function with a non-degenerate measure.

Let  $\vec{\Sigma}_p \sim \text{IW}(\nu_p, \vec{B}_p)$ . We write the discretized version of  $f$  as  $\vec{f}_p = (f(t_1), \dots, f(t_p))^T$ .  $\vec{f}_p$  can be used to approximate  $f$  by linear interpolation over the grid. Let  $\vec{g}_p = \vec{\Sigma}_p \vec{f}_p$ , we will find the first two moments of  $\vec{g}_p$  and investigate their limits as  $p \rightarrow \infty$ . Since  $\vec{\Sigma}_p \sim \text{IW}(\nu_p, \vec{B}_p)$ , we have

$$E[\vec{\Sigma}_p] = \frac{\vec{B}_p}{\nu_p - p - 1} \tag{6.34}$$

and

$$E[\vec{\Sigma}_p x x^T \vec{\Sigma}_p] = \frac{\vec{B}_p x x^T \vec{B}_p}{(\nu_p - p)(\nu_p - p - 3)} + \frac{\vec{B}_p (x^T \vec{B}_p x)}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)}, \tag{6.35}$$

for all  $x \in R^p$ , by our previous definition of Inverse Wishart in Section 6.5.1. For the first moment of  $\vec{g}_p$ ,

$$E[\vec{g}_p] = E[\vec{\Sigma}_p \vec{f}_p] = \frac{\vec{B}_p \vec{f}_p}{\nu_p - p - 1}$$

by (6.34). Suppose that  $\vec{B}_p$  is the discretization of a covariance operator  $B$  with kernel  $b(s, t)$ ,

$$\begin{aligned} (E[\vec{g}_p])_i &= \frac{1}{(\nu_p - p - 1)} \sum_{j=1}^p (\vec{B}_p)_{ij} (\vec{f}_p)_j \\ &= \frac{1}{(\nu_p - p - 1)} \sum_{j=1}^p b(t_i, t_j) f(t_j) \\ &= \frac{p}{(\nu_p - p - 1)} \sum_{j=1}^p b(t_i, t_j) f(t_j) \frac{1}{p}. \end{aligned}$$

Therefore,  $E[\vec{g}_p(s)] \rightarrow Bf(s) = \int_0^1 b(s, t) f(t) dt$  as  $p \rightarrow \infty$ , provided that  $\frac{p}{\nu_p - p - 1} \rightarrow$

1. For the second moment of  $\vec{g}_p$ ,

$$\begin{aligned} E[\vec{g}_p \vec{g}_p^T] &= E[\vec{\Sigma}_p \vec{f}_p \vec{f}_p^T \vec{\Sigma}_p] = \frac{\vec{B}_p \vec{f}_p \vec{f}_p^T \vec{B}_p}{(\nu_p - p)(\nu_p - p - 3)} \\ &\quad + \frac{\vec{B}_p (\vec{f}_p^T \vec{B}_p \vec{f}_p)}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)}, \end{aligned} \quad (6.36)$$

by (6.35). For the first term of (6.36), since

$$\begin{aligned} (\vec{B}_p \vec{f}_p \vec{f}_p^T \vec{B}_p)_{ij} &= (\vec{B}_p \vec{f}_p)_i (\vec{B}_p \vec{f}_p)_j \\ &= \left( \sum_{l=1}^p b(t_i, t_l) f(t_l) \right) \left( \sum_{\tau=1}^p b(t_j, t_{\tau}) f(t_{\tau}) \right) \\ &= p^2 \left( \sum_{l=1}^p b(t_i, t_l) f(t_l) \frac{1}{p} \right) \left( \sum_{\tau=1}^p b(t_j, t_{\tau}) f(t_{\tau}) \frac{1}{p} \right), \end{aligned}$$

we have

$$\frac{\vec{B}_p \vec{f}_p \vec{f}_p^T \vec{B}_p}{(\nu_p - p)(\nu_p - p - 3)} \rightarrow Bf \otimes Bf$$

as  $p \rightarrow \infty$ , provided that  $\frac{p^2}{(\nu_p - p)(\nu_p - p - 3)} \rightarrow 1$ . Note that we have defined the  $\otimes$  operation in (6.5) and (6.16). Here  $Bf \otimes Bf$  is defined in a similar way, i.e.,  $(Bf \otimes Bf)x = Bf\langle Bf, x \rangle$ , and  $\langle Bf, x \rangle = \int_0^1 Bf(s)x(s)ds$ . For the second term of (6.36),

$$\begin{aligned} \vec{f}_p^T \vec{B}_p \vec{f}_p &= \sum_{i=1}^p f(t_i) \left( \sum_{j=1}^p b(t_i, t_j) f(t_j) \right) \\ &= p^2 \sum_{i=1}^p \sum_{j=1}^p f(t_i) b(t_i, t_j) f(t_j) \frac{1}{p^2} \end{aligned}$$

where  $\sum_{i=1}^p \sum_{j=1}^p f(t_i) b(t_i, t_j) f(t_j) \frac{1}{p^2} \longrightarrow \int_0^1 \int_0^1 f(s) b(s, t) f(t) dt ds = \langle f, Bf \rangle$ , as  $p \rightarrow \infty$ . Therefore, for any  $x \in L^2[0, 1]$  with a discretized version  $\vec{x}_p = (x(t_1), \dots, x(t_p))^T$ ,

$$\begin{aligned} &\frac{1}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)} (\vec{B}_p (\vec{f}_p^T \vec{B}_p \vec{f}_p) \vec{x}_p)_l \\ &= \frac{p^3}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)} \left( \sum_{i=1}^p \sum_{j=1}^p f(t_i) b(t_i, t_j) f(t_j) \frac{1}{p^2} \right) \left( \sum_{\tau=1}^p b(t_l, t_\tau) x(t_\tau) \frac{1}{p} \right) \\ &\longrightarrow 1 \cdot \langle f, Bf \rangle Bx(t_l) \end{aligned}$$

as  $p \rightarrow \infty$ , provided that  $\frac{p^3}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)} \rightarrow 1$ . Thus the second term of (6.36) satisfies

$$\frac{\vec{B}_p (\vec{f}_p^T \vec{B}_p \vec{f}_p)}{(\nu_p - p - 1)(\nu_p - p)(\nu_p - p - 3)} \longrightarrow \langle f, Bf \rangle B.$$

The above results shows that  $E[\vec{g}_p] \rightarrow Bf$  and  $E[\vec{g}_p \vec{g}_p^T] \rightarrow Bf \otimes Bf + \langle f, Bf \rangle B$

under the condition that  $\frac{p}{\nu_p - p} \rightarrow 1$ . This implies

$$Cov(\vec{g}_p) = E[\vec{g}_p \vec{g}_p^T] - E[\vec{g}_p]E[\vec{g}_p]^T$$

$$\longrightarrow Bf \otimes Bf + \langle f, Bf \rangle B - Bf \otimes Bf,$$

hence  $Cov(\vec{g}_p) \rightarrow \langle f, Bf \rangle B$  as  $p \rightarrow \infty$ .

To summarize, we have obtained the limit of the first two moments of  $\vec{g}_p = \vec{\Sigma}_p \vec{f}_p$ .

As the number of grid points  $p \rightarrow \infty$  and  $\frac{p}{\nu_p - p} \rightarrow 1$ , we have

$$E[\vec{\Sigma}_p \vec{f}_p) \longrightarrow Bf,$$

$$Cov(\vec{\Sigma}_p \vec{f}_p) \longrightarrow \langle f, Bf \rangle B.$$

# Chapter 7

## Conclusion and Discussion

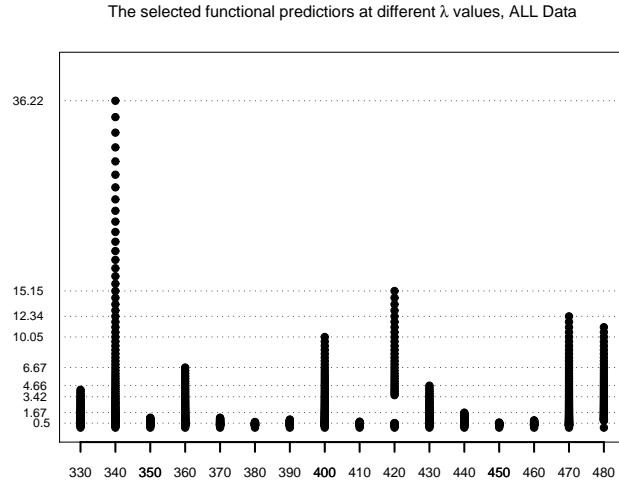
In summary, we have proposed three statistical models on the topic of functional data classification, and presented a study on the covariance operator of functional data analysis. We compare the results from previous chapters and discuss some related issues in this chapter.

The Bayesian variable selection model proposed in Chapter 3 provides good classification performance compared with several other methods without variable selection. The functional predictors are approximated using orthonormal basis expansion, and variable selection is performed based on the coefficients of the orthonormal basis. This model is novel as a functional data classification method, however, it also has some drawbacks. First, the variable selection results depend on different choices of the orthonormal basis. Second, the variables selected are usually hard to explain and visualize in the original function space. Orthonormal basis such as Wavelets can

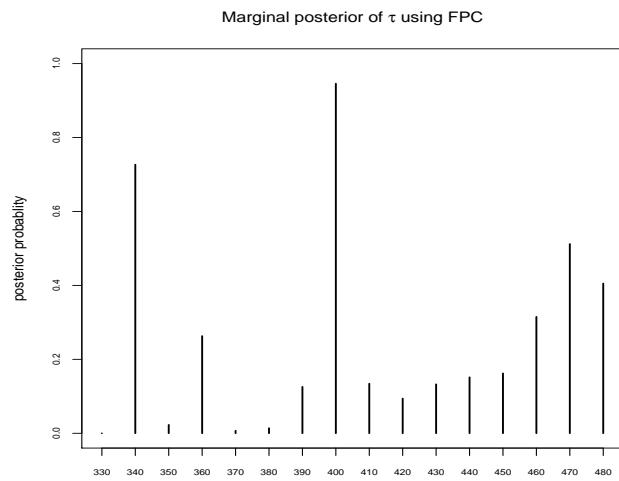
preserve some location information, therefore may improve the model and make it easier to explain.

The functional generalized linear model (FGLM) in Chapter 4 and the Bayesian hierarchical model in Chapter 5 both aim to select a subset of functional predictors in order to reduce the cost of data collection in the cervical cancer diagnosis application. However, the selection results reported in Chapter 4 are not comparable with those in Chapter 5 due to the fact that the FGLM model does not consider random effects, and the real data managed by FGLM are a subset of the whole dataset that are measured by a fixed device (and clinic). To compare FGLM with the Bayesian hierarchical model on their predictor selection performance, we re-trained the FGLM model based on all data and ignore the random effect. Figure 7.1 plots the predictor selection result of FGLM using all data together with the marginal posterior of  $\tau$  obtained in Chapter 5. Note that these two results are both based on the FPC method with approximation criterion  $c_1 = 0.998$ . Although Figure 7.1(a) and Figure 7.1(b) have different explanation for their own model, they show some similarities on the selection of functional predictors, i.e., the curves with excitation wavelengths at around 340-360, 400-420, 470-480nm have higher possibilities of being selected.

Finally, for the study of the covariance operator, besides the results obtained in Chapter 6, there are more theoretical work that worth further investigation. First, the consistency of the posterior needs to be constructed based on the priors introduced in Section 6.3. Second, more computationally efficient MCMC algorithms need to



(a) FGLM model trained on all data: the selected functional predictors at different  $\lambda$  using function approximation with FPC.



(b) BHFPS model: the marginal posterior probabilities  $P\{\tau_j = 1, j = 1, \dots, 16\}$  using function approximation with FPC.

Figure 7.1: A comparison of the functional predictor selection results of FGLM and BHFPS.

be developed to deal with data with large number of grids. It is also of interest to look for the infinite-dimensional counterpart for the Inverse-Wishart distribution and construct priors from there. Continuation of the covariance operator research will certainly enrich the field of functional data analysis.

## Appendix A

# Integrating $b_l$ 's, $b_0$ and $\alpha$ Out Sequentially from the Conditional Posterior (5.10).

From conditional posteriors in (5.10) and priors in (5.2) and (5.9), we have

$$\begin{aligned}
& \pi(\alpha, b_1, \dots, b_L, b_0, \sigma_b^2, \tau | Z_l, Y_l, l = 1, \dots, L) \\
& \propto \prod_l |K_l^{-1}|^{-1/2} \exp \left\{ -\frac{1}{2} \sum_l (b_l^T K_l b_l - 2b_l^T M_l + M_l^T K_l^{-1} M_l) \right\} \\
& \quad \cdot \exp \left\{ \frac{1}{2} \sum_l [M_l^T K_l^{-1} M_l - (Z_l - S_l \alpha)^T (Z_l - S_l \alpha)] \right\} \\
& \quad \cdot \exp \left\{ -\frac{1}{2} b_0^T [L(\sigma_b^2 \Sigma_\tau)^{-1} + (\sigma_0^2 \Sigma_\tau)^{-1}] b_0 \right\} \\
& \quad \cdot \exp \left\{ -\frac{1}{2} \alpha^T (\sigma_1^2 I)^{-1} \alpha \right\} (\prod_l |K_l^{-1}|^{1/2}) |\sigma_b^2 \Sigma_\tau|^{-L/2} |\sigma_0^2 \Sigma_\tau|^{-1/2} \pi(\sigma_b^2) \pi(\tau),
\end{aligned}$$

where  $K_l = C_l^T C_l + (\sigma_b^2 \Sigma_\tau)^{-1}$  and  $M_l = C_l^T (Z_l - S_l \alpha) + (\sigma_b^2 \Sigma_\tau)^{-1} b_0$ ,  $l = 1, \dots, L$ . From above, we find the conditional distribution  $b_l | \alpha, b_0, \sigma_b^2, \tau, Z_l, Y_l \sim N(\mu_l, V_l)$ , where  $\mu_l = K_l^{-1} M_l$  and  $V_l = K_l^{-1}$ , for  $l = 1, \dots, L$ . The  $b_l$ 's can be integrated out from the above conditional posterior since the first  $2L$  factors construct  $L$  normal density kernels. After integrating out  $b_l$ 's, we can expand  $M_l^T K_l^{-1} M_l$  and combine the terms with  $b_0$ , which gives the following:

$$\begin{aligned} & \pi(\alpha, b_0, \sigma_b^2, \tau | Z_l, Y_l, l = 1, \dots, L) \\ & \propto |K_0^{-1}|^{-1/2} \exp \left\{ -\frac{1}{2} (b_0^T K_0 b_0 - 2b_0^T M_0 + M_0^T K_0^{-1} M_0) \right\} \\ & \quad \cdot \exp \left\{ \frac{1}{2} M_0^T K_0^{-1} M_0 + \frac{1}{2} \sum_l (Z_l - S_l \alpha)^T (C_l K_l^{-1} C_l^T - I) (Z_l - S_l \alpha) \right\} \\ & \quad \cdot \exp \left\{ -\frac{1}{2} \alpha^T (\sigma_1^2 I)^{-1} \alpha \right\} |K_0^{-1}|^{1/2} \left( \prod_l |K_l^{-1}|^{1/2} \right) |\sigma_b^2 \Sigma_\tau|^{-L/2} |\sigma_0^2 \Sigma_\tau|^{-1/2} \pi(\sigma_b^2) \pi(\tau), \end{aligned}$$

where  $K_0 = (\sigma_0^2 \Sigma_\tau)^{-1} + L(\sigma_b^2 \Sigma_\tau)^{-1} - (\sigma_b^2 \Sigma_\tau)^{-1} (\sum_l K_l^{-1}) (\sigma_b^2 \Sigma_\tau)^{-1}$  and

$$M_0 = (\sigma_b^2 \Sigma_\tau)^{-1} \sum_l K_l^{-1} C_l^T (Z_l - S_l \alpha)$$

. It is easy to see from above that  $b_0 | \alpha, \sigma_b^2, \tau, Z_l, Y_l \sim N(\mu_0, V_0)$ , where  $\mu_0 = K_0^{-1} M_0$  and  $V_0 = K_0^{-1}$ . We can further integrate  $b_0$  out since the first two factors form a normal density kernel. After integrating out  $b_0$ , we can expand the term  $M_0^T K_0^{-1} M_0$ , combine terms of  $\alpha$  and factor out a normal kernel for  $\alpha$ , from where we obtain that

$\alpha | \sigma_b^2, \tau, Z_l, Y_l, \forall l \sim N(\mu_\alpha, V_\alpha)$ , where  $\mu_\alpha = \tilde{K}^{-1} \tilde{M}$ ,  $V_\alpha = \tilde{K}^{-1}$ ,

$$\begin{aligned} \tilde{K} &= \sum_l S_l^T S_l + (\sigma_1^2 I)^{-1} - \sum_l S_l^T C_l K_l^{-1} C_l^T S_l \\ &\quad - \left( \sum_l K_l^{-1} C_l^T S_l \right)^T (\sigma_b^2 \Sigma_\tau)^{-1} K_0^{-1} (\sigma_b^2 \Sigma_\tau)^{-1} \left( \sum_l K_l^{-1} C_l^T S_l \right), \end{aligned}$$

and

$$\tilde{M} = \sum_l S_l^T Z_l - \sum_l S_l^T C_l K_l^{-1} C_l^T Z_l - \left( \sum_l K_l^{-1} C_l^T S_l \right)^T (\sigma_b^2 \Sigma_\tau)^{-1} K_0^{-1} (\sigma_b^2 \Sigma_\tau)^{-1} \left( \sum_l K_l^{-1} C_l^T Z_l \right).$$

We finally can integrate out  $\alpha$  to obtain the marginal conditional posterior of  $\sigma_b^2$  and  $\tau$ , conditional on values of  $Z_l$ 's and  $Y_l$ 's, which gives

$$\begin{aligned} & \pi(\sigma_b^2, \tau | Z_l, Y_l, l = 1, \dots, L) \\ & \propto \exp \left\{ \frac{1}{2} \tilde{M}^T \tilde{K}^{-1} \tilde{M} + \frac{1}{2} \left( \sum_l K_l^{-1} C_l^T Z_l \right)^T (\sigma_b^2 \Sigma_\tau)^{-1} K_0^{-1} (\sigma_b^2 \Sigma_\tau)^{-1} \left( \sum_l K_l^{-1} C_l^T Z_l \right) \right\} \\ & \quad \cdot \exp \left\{ \frac{1}{2} \sum_l Z_l^T C_l K_l^{-1} C_l^T Z_l \right\} |\tilde{K}|^{-1/2} |K_0|^{-1/2} \left( \prod_l |K_l|^{-1/2} \right) |\sigma_b^2 \Sigma_\tau|^{-L/2} |\sigma_0^2 \Sigma_\tau|^{-1/2} \\ & \quad \cdot \pi(\sigma_b^2) \pi(\tau), \end{aligned}$$

where  $\tilde{K}$ ,  $\tilde{M}$ ,  $K_0$  and  $K_l$ 's are defined in the above derivation.

## Appendix B

### Proof of Proposition 4.2.1

The proof of Proposition 4.2.1 uses a result stated in the following lemma.

**Lemma B.0.1.** *Let  $f : \mathbb{R}^n \mapsto \mathbb{R}$  be a strictly convex function with a minimizer  $\tilde{x}$ , and let  $g : \mathbb{R}^n \mapsto [0, \infty)$  be a convex function. Then  $f + g$  has a unique minimizer  $x^*$  in  $\mathbb{R}^n$ .*

*Proof:* Let  $h(x) = f(x) + g(x)$ . It is easy to show that  $h(x)$  is strictly convex from the definition. We claim that the existence of a minimizer  $\tilde{x}$  of  $f$  implies that  $h$  is coercive, which means  $h(x) \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ . The coerciveness and strict convexity of  $h$  implies the existence of a unique minimizer  $x^*$ .

To show that  $h$  is coercive, it is sufficient to show that  $f$  is coercive (since  $g \geq 0$ ). The minimizer  $\tilde{x}$  of  $f$  is the unique minimizer of  $f$  by strict convexity. Also,  $f$  is convex hence is continuous on  $\mathbb{R}^n$  (see [66], page 82). Thus  $\forall r > 0, \forall x$  such that  $\|x - \tilde{x}\| > r$ , we claim

$$f(x) > \frac{b}{r} \|x - \tilde{x}\| + f(\tilde{x})$$

where  $b = \inf\{f(x) : \|x - \tilde{x}\| = r\} - f(\tilde{x})$ . Note that  $b$  exists and  $b > 0$  by continuity of  $f$ . To show this inequality, let  $x_0 = r(x - \tilde{x})/(\|x - \tilde{x}\|) + \tilde{x}$ , so that  $x_0$  lies on the line formed by  $x$  and  $\tilde{x}$ , with  $\|x_0 - \tilde{x}\| = r$  and  $\|x - x_0\| = \|x - \tilde{x}\| - r$ . Thus  $f(x_0) - f(\tilde{x}) \geq b$  by the definition of  $b$ . Now let  $\alpha = r/\|x - \tilde{x}\|$ . We see that  $x_0 = \alpha x + (1 - \alpha)\tilde{x}$ . By strict convexity of  $f$ ,

$$f(x_0) < \alpha f(x) + (1 - \alpha)f(\tilde{x})$$

Thus

$$\begin{aligned} \frac{b}{r}\|x - \tilde{x}\| + f(\tilde{x}) &\leq (f(x_0) - f(\tilde{x}))\frac{\|x - \tilde{x}\|}{r} + f(\tilde{x}) \\ &< (\alpha f(x) + (1 - \alpha)f(\tilde{x}) - f(\tilde{x}))\frac{\|x - \tilde{x}\|}{r} + f(\tilde{x}) \\ &= f(x) \end{aligned}$$

Since  $\|x - \tilde{x}\| \geq \|x\| - \|\tilde{x}\|$ ,  $\|x\| \rightarrow \infty$  implies  $\|x - \tilde{x}\| \rightarrow \infty$ , which implies  $f(x) \rightarrow \infty$  by the above inequality and the facts that  $b > 0, r > 0, f(\tilde{x})$  finite. Therefore,  $f$  is coercive, and so is  $h$ .

Since  $h$  is coercive, we have  $h(x) \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ . Therefore, if we pick an arbitrary point  $x_1 \in \mathbb{R}^n$ , there exists a constant  $\delta > 0$  such that  $h(x) > h(x_1)$  for all  $\|x - x_1\| > \delta$ . Since the domain  $\|x - x_1\| \leq \delta$  is compact and  $h(x)$  is strictly convex on it,  $h(x)$  has a unique minimizer in  $\|x - x_1\| \leq \delta$ , which we denote as  $x^*$ . (A strictly convex real valued function defined on a compact domain has a unique minimum on its domain.) This  $x^*$  is also the global minimizer since  $h(x) > h(x_1) \geq h(x^*)$  on  $\|x - x_1\| > \delta$ .

*Proof of Proposition 4.2.1:* Based on results in Lemma B.0.1, we let  $f$  to be  $-l(\theta)$  and  $g$  to be  $\lambda \sum_{j=1}^J s(\delta_j) \|b_j\|_2$ , therefore our objective function in (4.9) is the sum of  $f$  and  $g$ , where  $\theta = \{\alpha_0, \alpha, b_j, j = 1, \dots, J\}$ , and  $l(\theta) = \sum_{i=1}^n y_i \eta_i - \log(1 + \exp(\eta_i))$  with  $\eta_i = \alpha_0 + z_i^T \alpha + \sum_{j=1}^J \sum_{k=1}^{\delta_j} c_{ijk} b_{jk}$ .

Firstly, we show that  $-l(\theta)$  is strictly convex. It is sufficient to show that its Hessian is positive definite. Since the Hessian takes the form

$$\nabla_\theta^2(-l(\theta)) = X^T D X$$

where  $D = \text{diag}\{\exp(\eta_i)/(1 + \exp(\eta_i))^2, i = 1, \dots, n\}$ . It is positive definite since  $X$  is of rank  $m$  (full rank). Secondly, since the maximum likelihood estimator exists,  $-l(\theta)$  has an unique minimizer. The existence of maximum likelihood estimator for logistic regression requires some conditions for the design matrix  $X$ . Basically, the  $n$  rows of  $X$  can not be completely separated or quasi-completely separated in  $\mathbb{R}^m$ . See [1] for details. In practice, as long as we can find a numerical solution for the MLE at  $\lambda = 0$ , we would believe that the maximum likelihood estimator exists. Finally, let  $g(b) = \lambda \sum_{j=1}^J s(\delta_j) \|b_j\|_2$ ,  $b^T = (b_1^T, \dots, b_J^T)$ . It is easy to see that  $g(b)$  is convex by the triangle inequality. Therefore by Lemma B.0.1,  $Q_\lambda(\theta)$  has a unique minimizer  $\theta^*$ .

## Appendix C

# Verification for Convergence of the MCMC Algorithm 1 in Chapter 5.

### C.1 The Verification of Algorithm 1

Based Equation (5.8),(5.10) and (5.11) in Section 5.2, Algorithm 1 can be simplified as follows:

Step 0. Set initial values for  $b_l$ 's,  $\alpha$ ,  $\tau$  and  $\sigma_b^2$ .

Step 1.  $Z_l | \alpha, b_l, Y_l \sim TN, l = 1, \dots, L$ .

Step 2.  $\sigma_b^2 | \tau, Z_l, Y_l, l = 1, \dots, L$ .

Step 3.  $\tau | \sigma_b^2, Z_l, Y_l, l = 1, \dots, L$ .

Step 4.  $\alpha | \sigma_b^2, \tau, Z_l \sim N(\mu_\alpha, V_\alpha)$ .

Step 5.  $b_0|\alpha, \sigma_b^2, \tau, Z_l \sim N(\mu_0, V_0)$ .

Step 6.  $b_l|b_0, \alpha, \sigma_b^2, \tau, Z_l \sim N(\mu_l, V_l)$ .

Note that Step 2 and 3 are two Metropolis-Hastings steps within the larger Gibbs steps. Step 4-6 are simple Gibbs steps. Let  $Z = (Z_1, \dots, Z_L)$ ,  $Y = (Y_1, \dots, Y_L)$  and  $b = (b_1, \dots, b_L)$ . We firstly combine step 2 and 3 by letting  $E = (\sigma_b^2, \tau)$ . We can represent the transition kernel from Step 2–3 as  $P(E, A)$  with (conditional) transition density  $f(\tilde{E}|E, Z, Y) = p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2, Z, Y)p_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau, Z, Y)$ . Therefore,  $f(\tilde{E}|Z, Y) = \int f(\tilde{E}|E, Z, Y)f(E|Z, Y)dE$  in the bigger Gibbs steps in (C.1). Later on we will verify that  $P(E, A)$  is invariant with respect to the conditional measure  $f(E|Z, Y)$ .

First of all, we need to check that the transition kernel formed by the whole MCMC steps is invariant. Here we denote the domain of parameter  $x$  as  $\mathcal{D}(x)$ . Then

$$\int_{\mathcal{D}(b)} \int_{\mathcal{D}(b_0)} \int_{\mathcal{D}(\alpha)} \int_{\mathcal{D}(E)} \int_{\mathcal{D}(Z)} f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y) \quad (C.1)$$

$$\cdot f(\tilde{Z}|b, b_0, \alpha, W, Y) f(b, b_0, \alpha, W, Z|Y) dZ dW d\alpha db_0 db$$

$$= \int_{\mathcal{D}(b)} \int_{\mathcal{D}(b_0)} \int_{\mathcal{D}(\alpha)} \int_{\mathcal{D}(E)} f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y)$$

$$\cdot f(\tilde{Z}|b, b_0, \alpha, W, Y) f(b, b_0, \alpha, W|Y) dW d\alpha db_0 db$$

$$(\text{Since } \int_{\mathcal{D}(Z)} f(b, b_0, \alpha, W, Z|Y) dZ = f(b, b_0, \alpha, W|Y).)$$

$$= \int_{\mathcal{D}(b)} \int_{\mathcal{D}(b_0)} \int_{\mathcal{D}(\alpha)} \int_{\mathcal{D}(E)} f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y)$$

$$\cdot f(\tilde{Z}|b, b_0, \alpha, W|Y) dW d\alpha db_0 db$$

$$\begin{aligned}
&= \int_{\mathcal{D}(b)} \int_{\mathcal{D}(b_0)} \int_{\mathcal{D}(\alpha)} f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y) \\
&\quad \cdot f(\tilde{Z}, b, b_0, \alpha|Y) d\alpha db_0 db \\
&\quad (\text{Since } \int_{\mathcal{D}(E)} f(\tilde{Z}, b, b_0, \alpha, W|Y) dW = f(\tilde{Z}, b, b_0, \alpha|Y).) \\
&= f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y) \\
&\quad \cdot \int_{\mathcal{D}(b)} \int_{\mathcal{D}(b_0)} \int_{\mathcal{D}(\alpha)} f(\tilde{Z}, b, b_0, \alpha|Y) d\alpha db_0 db \\
&= f(\tilde{b}|\tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{b}_0|\tilde{\alpha}, \tilde{E}, \tilde{Z}, Y) f(\tilde{\alpha}|\tilde{E}, \tilde{Z}, Y) f(\tilde{E}|\tilde{Z}, Y) f(\tilde{Z}|Y) \\
&= f(\tilde{b}, \tilde{b}_0, \tilde{\alpha}, \tilde{E}, \tilde{Z}|Y)
\end{aligned}$$

This shows that the transition distribution formed by the larger Gibbs steps from step 1-6 is invariant.

Secondly, we look at step 2 and 3 in detail. we need to show that the transition density  $f(\tilde{E}|E, Z, Y) = p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2, Z, Y)p_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau, Z, Y)$  is invariant with respect to the conditional distribution  $f(E|Z, Y)$ . For simplicity, we remove the  $Z, Y$  from the transitional densities since all of them(within steps 2 – 3) are conditional on  $Z, Y$ .

Let  $q_1(\tilde{\sigma}_b^2|\sigma_b^2)$  be the proposal density for step 2, with the corresponding acceptance rate

$$\alpha_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau) = \min\left\{\frac{\pi(\tilde{\sigma}_b^2|\tau)q_1(\sigma_b^2|\tilde{\sigma}_b^2)}{\pi(\sigma_b^2|\tau)q_1(\tilde{\sigma}_b^2|\sigma_b^2)}, 1\right\}.$$

Therefore the transition density for step 2 is

$$p_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau) = q_1(\tilde{\sigma}_b^2|\sigma_b^2)\alpha_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau)\mathbf{1}_{\{\tilde{\sigma}_b^2 \neq \sigma_b^2\}}.$$

Then the Metropolis-Hastings routine gives us the following so called **reversibility condition**:

$$\pi(\sigma_b^2 | \tau) p_1(\tilde{\sigma}_b^2 | \sigma_b^2, \tau) = \pi(\tilde{\sigma}_b^2 | \tau) p_1(\sigma_b^2 | \tilde{\sigma}_b^2, \tau). \quad (\text{C.2})$$

Similarly, we let the proposal density for step 3 to be  $q_2(\tilde{\tau} | \tau)$ . The associated acceptance rate is

$$\alpha_2(\tilde{\tau} | \tau, \tilde{\sigma}_b^2) = \min\left\{\frac{\pi(\tilde{\tau} | \tilde{\sigma}_b^2) q_2(\tau | \tilde{\tau})}{\pi(\tau | \tilde{\sigma}_b^2) q_2(\tilde{\tau} | \tau)}, 1\right\}. \quad (\text{C.3})$$

Hence the transition density for step 3 is

$$p_2(\tilde{\tau} | \tau, \tilde{\sigma}_b^2) = q_2(\tilde{\tau} | \tau) \alpha_2(\tilde{\tau} | \tau, \tilde{\sigma}_b^2) \mathbf{1}_{\{\tilde{\tau} \neq \tau\}}.$$

Again, Metropolis-Hastings routine gives us the following reversibility condition:

$$\pi(\tau | \tilde{\sigma}_b^2) p_2(\tilde{\tau} | \tau, \tilde{\sigma}_b^2) = \pi(\tilde{\tau} | \tilde{\sigma}_b^2) p_2(\tau | \tilde{\tau}, \tilde{\sigma}_b^2). \quad (\text{C.4})$$

The proofs of Equation (C.2) and (C.4) are general for Metropolis-Hastings and can be done by following the theorem in the Section C.2.

Based on the above setup, the invariant transition distribution for step 2 and 3 can thus been shown as follows:

$$\begin{aligned}
& \int_{\mathcal{D}(E)} f(\tilde{E}|E) \pi(E) dE \\
&= \int_{\mathcal{D}(\tau)} \int_{\mathcal{D}(\sigma_b^2)} p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2) p_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau) \pi(\sigma_b^2|\tau) \pi(\tau) d\sigma_b^2 d\tau \\
&= \int_{\mathcal{D}(\tau)} \pi(\tau) p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2) \left[ \int_{\mathcal{D}(\sigma_b^2)} \pi(\sigma_b^2|\tau) p_1(\tilde{\sigma}_b^2|\sigma_b^2, \tau) d\sigma_b^2 \right] d\tau \\
&= \int_{\mathcal{D}(\tau)} \pi(\tau) p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2) \left[ \int_{\mathcal{D}(\tilde{\sigma}_b^2)} \pi(\tilde{\sigma}_b^2|\tau) p_1(\sigma_b^2|\tilde{\sigma}_b^2, \tau) d\sigma_b^2 \right] d\tau \quad (\text{by Equation (C.2)}) \\
&= \int_{\mathcal{D}(\tau)} \pi(\tau) p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2) \pi(\tilde{\sigma}_b^2|\tau) d\tau \\
&= \int_{\mathcal{D}(\tau)} \pi(\tilde{\sigma}_b^2) p_2(\tilde{\tau}|\tau, \tilde{\sigma}_b^2) \pi(\tau|\tilde{\sigma}_b^2) d\tau \\
&= \int_{\mathcal{D}(\tau)} \pi(\tilde{\sigma}_b^2) p_2(\tau|\tilde{\tau}, \tilde{\sigma}_b^2) \pi(\tilde{\tau}|\tilde{\sigma}_b^2) d\tau \quad (\text{by Equation (C.4)}) \\
&= \pi(\tilde{\sigma}_b^2) \pi(\tilde{\tau}|\tilde{\sigma}_b^2) \\
&= \pi(\tilde{\sigma}_b^2, \tilde{\tau}) \\
&= \pi(\tilde{E})
\end{aligned}$$

This proved that the Metropolis-Hastings Step in Step 2–3 has the right invariant density.

In addition to check invariance, we also need to check irreducibility and aperiodicity.(Note that irreducibility and existence of invariant distribution implies recurrence,

and thus implies positive recurrency when  $\pi$  has finite total mass([73],page 1712).)  
 Since the algorithm has component-wise transition, it suffices to check that each transition kernel (in each step) is irreducible and aperiodic. The irreducibility for transitions of  $Z, \sigma_b^2, \alpha, b_0$  and  $b$  is straight-forward since the transitions are fully supported on their convex domains. For  $\tau$ , it lies in a domain of finite number of points, for each pair of  $\tau$  and  $\tau'$ , there is a  $n$  such that  $P^n(\tau'|\tau) > 0$ . A simple strategy is let  $\tau$  firstly reduce to a vector of all 0's in  $\tau^T\tau$  steps, and let it increase to  $\tau'$  in  $(\tau')^T\tau$  steps, then  $n = \tau^T\tau + (\tau')^T\tau$ , and the transition probability is positive. Aperiodicity is trivial to check. Since we can not find a d-cycle for the transition kernel hence it is aperiodic.

To sum up, we have shown that the transition kernel formed by algorithm 1 has invariant distribution  $\pi(\cdot)$  and is irreducible and aperiodic, hence by Theorem 1 of Tierney([73]), it converges (in total variation) to a unique distribution  $\pi(\cdot)$ , which is our posterior density.

## C.2 Reversible Condition of Metropolis-Hastings

Assume that  $\pi$  has a density with respect to  $\mu$  and let  $Q$  be a transition kernel of the form

$$Q(x, dy) = q(x, y)\mu(dy).$$

Let  $E^+ = \{x : \pi(x) > 0\}$  and assume that  $Q(x, E^+) = 1$  for  $x \notin E^+$ . Also assume that  $\pi$  is not concentrated on a single point. For a given  $X_n = x$ , we propose a

candidate value  $Y = y$  for the next point  $X_{n+1}$  from the distribution  $Q(x, \cdot)$ , and accept it with probability

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\}.$$

Otherwise, the candidate is rejected and the chain remains at  $X_{n+1} = x$ .

If we define the off-diagonal density of a Metropolis kernel as

$$p(x, y) = q(x, y)\alpha(x, y)\mathbf{1}_{\{x \neq y\}},$$

and set  $r(x) = 1 - \int p(x, y)dy$ , then the Metropolis kernel  $P$  can be written as

$$P(x, dy) = p(x, y)\mu(dy) + r(x)\delta_x(dy), \quad (\text{C.5})$$

where  $\delta_x$  denote point mass at  $x$ . The value  $r(x)$  is the probability that the algorithm remains at  $x$ .

**Proposition** *For the Metropolis kernel defined above, we have*

$$\pi(x)p(x, y) = \pi(y)p(y, x), \quad (\text{C.6})$$

*which is called reversibility condition.*

**proof.** If  $x = y$ , then  $p(x, y) = 0$ , both sides equal 0. If  $x \neq y$  and  $\pi(y)q(y, x) \geq \pi(x)q(x, y)$ , we have  $\alpha(x, y) = 1$ . Therefore the left hand side(LHS) of Equation C.6 is

$$\text{LHS} = \pi(x)p(x, y) = \pi(x)q(x, y)\alpha(x, y) = \pi(x)q(x, y).$$

The right hand side(RHS) of Equation C.6 is

$$\text{RHS} = \pi(y)p(y, x) = \pi(y)q(y, x)\alpha(y, x) = \pi(y)q(y, x)\frac{\pi(x)q(x, y)}{\pi(y)q(y, x)} = \pi(x)q(x, y).$$

Therefore LHS=RHS, the equality holds. For the case of  $\pi(y)q(y, x) < \pi(x)q(x, y)$ , we can similarly show that the equality holds.

## Appendix D

### Some Details on EMC Algorithms

Here we give a more detailed introduction of EMC algorithm based on the work of Liang and Wong [39], Liu [40] and Goswami and Liu [24]. The basic goal of EMC algorithm is to generate Markov Chain samples from a target distribution  $\pi(x)$ , which can be a posterior distribution, or a conditional posterior distribution. In Liang and Wong [39], Liu [40] and Goswami and Liu [24], they focus on sampling from a target distribution with density function

$$f(x) \propto \exp\{-H(x)/t\}, \quad (\text{D.1})$$

where  $H(x)$  is called an *energy* function , which is equivalent to  $-\log \pi(x)$  in our Bayesian setup. The target function (D.1) is then a transformed version of  $\pi(x)$  since

$$\exp\{H(x)/t\} = \exp\{-(-\log \pi(x))/t\} = \pi(x)^{1/t}.$$

The  $t$  is called a *temperature*, which has the effect of making the target density more flat or more spiky, as shown in Figure D.1. Liang and Wong [39] assume that there

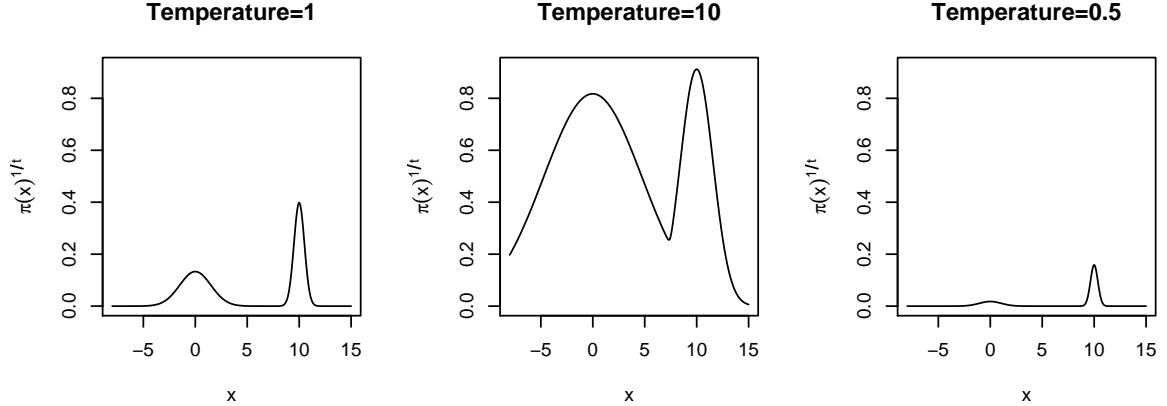


Figure D.1: The plot of  $\pi(x)^{1/t}$  for a two-mode mixture normal distribution. The density  $\pi(x) = 1/2\phi(x; 0, 1.5^2) + 1/2\phi(x; 10, 0.5^2)$ , where  $\phi(x; \mu, \sigma^2)$  is the normal density with mean  $\mu$  and variance  $\sigma^2$ .

are multiple  $t$ 's, denoted as  $t_i, i = 1, \dots, N$ , and  $t_i$ 's are ordered from high to low. The set  $\{t_1, \dots, t_N\}$  is called a temperature ladder. Assume that  $x \in \mathbb{R}^d$  and here we assume each component of  $x$  is either 1 or 0. EMC algorithm first expands the sample space from  $\mathbb{R}^d$  to  $\mathbb{R}^{Nd}$  by defining a new target density

$$\pi(\mathbf{x}) \propto \prod_{i=1}^N \pi(x_i)^{1/t_i},$$

where  $\mathbf{x} = (x_1, \dots, x_N)$  is called a population of samples. The Markov Chain samples is obtained based on  $\pi(\mathbf{x})$  with 3 types of operation: mutation, crossover and exchange. We summarize the details of the EMC algorithm stated in Liang and Wong [39] and Liu [40].

### An EMC algorithm

Step 0. Set the temperature ladder  $\{t_1, \dots, t_N\}$ , the initial values  $\mathbf{x} = (x_1, \dots, x_N)$

and the mutation rate  $q_m$ .

Step 1. With probability  $q_m$ , run mutation and with probability  $1 - q_m$ , run cross over.

(a) **Mutation.** Randomly select  $x_k$  from  $(x_1, \dots, x_k, \dots, x_N)$ . Propose  $x'_k$  by reversing some randomly selected bits of  $x_k$  (Note: it is called 1-point/2-points mutation based on the number of bits selected for switch). Denote  $\mathbf{x}' = (x_1, \dots, x'_k, \dots, x_N)$ , the new  $\mathbf{x}'$  is accepted with  $\min(1, r_m)$ , with

$$\log r_m = \log \left\{ \frac{\pi(\mathbf{x}') T(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x}) T(\mathbf{x}'|\mathbf{x})} \right\} = \frac{[\log \pi(x'_k) - \log \pi(x_k)]}{t_k} \frac{T(\mathbf{x}|\mathbf{x}')}{T(\mathbf{x}'|\mathbf{x})}.$$

Here  $T(\mathbf{x}|\mathbf{x}')$  denotes the transition probability of the proposal. Note that using 1-point or 2-point mutation will both result in symmetric transition probability ([39], Page 322).

(b) **Crossover.** First, randomly select a pair  $(x_i, x_j)$ , according to probability

$$p((x_i, x_j)|\mathbf{x}) = \frac{\pi(x_i)^{1/t} + \pi(x_j)^{1/t}}{\sum_{j=1}^N \pi(x_j)^{1/t}}, \quad x_i \neq x_j.$$

This can be done by firstly selecting  $x_i$  with probability

$$p(x_i|\mathbf{x}) = \pi(x_i)^{1/t} / \sum_j \pi(x_j)^{1/t},$$

then choosing  $x_j$  independent of  $x_i$ , but with the same sampling probability. If  $x_i = x_j$ , we discard them and repeat sampling until we obtain a distinct pair. ([40], Page 231). Here  $t$  is fixed (may not be the same with items in the temperature ladder). This selecting procedure is called

“roulette wheel” selection ([39], Page 319). After the pair  $(x_i, x_j)$  is chosen, randomly select a location  $k$  as a crossover point, and swap  $x_i$  with  $x_j$  starting to the right of the crossover point([39],Page 320). For example, if we denote  $x_i = (a_1, \dots, a_k, \dots, a_d)$ , and denote  $x_j = (b_1, \dots, b_k, \dots, b_d)$ .

Then after crossover at location  $k$ , we get

$$x'_i = (a_1, \dots, a_k, b_{k+1}, \dots, b_d),$$

$$x'_j = (b_1, \dots, b_k, a_{k+1}, \dots, a_d).$$

Denote the population of sample after crossover to be

$$\mathbf{x}' = (x_1, \dots, x'_i, \dots, x'_j, \dots, x_N),$$

the Metropolis ratio can be computed by

$$\begin{aligned} \log r_c &= \log \frac{\pi(\mathbf{x}') T(\mathbf{x} | \mathbf{x}')}{\pi(\mathbf{x}) T(\mathbf{x}' | \mathbf{x})} \\ &= \frac{\log \pi(x'_i) - \log \pi(x_i)}{t_i} + \frac{\log \pi(x'_j) - \log \pi(x_j)}{t_j} + \log \frac{T(\mathbf{x} | \mathbf{x}')}{T(\mathbf{x}' | \mathbf{x})} \end{aligned}$$

where  $T(\mathbf{x}' | \mathbf{x}) = P\{(x_i, x_j) | \mathbf{x}\} P\{(x'_i, x'_j) | (x_i, x_j)\}$ . Note that according to the selection rule, we have  $P\{(x'_i, x'_j) | (x_i, x_j)\} = P\{(x_i, x_j) | (x'_i, x'_j)\}$ , therefore the ratio of transition probabilities is reduced to the ratio of selection probabilities, i.e.,

$$\begin{aligned} \frac{T(\mathbf{x} | \mathbf{x}')}{T(\mathbf{x}' | \mathbf{x})} &= \frac{P\{(x'_i, x'_j) | \mathbf{x}'\}}{P\{(x_i, x_j) | \mathbf{x}\}} \\ &= \frac{\pi(x'_i)^{1/t} + \pi(x'_j)^{1/t}}{\pi(x'_i)^{1/t} + \pi(x'_j)^{1/t} + \sum_{k \neq i,j} \pi(x_k)^{1/t}} \cdot \frac{\sum_{k=1}^N \pi(x_k)^{1/t}}{\pi(x_i)^{1/t} + \pi(x_j)^{1/t}}. \end{aligned}$$

The new  $\mathbf{x}'$  is accepted with probability  $\min(1, r_c)$ .

Step 2. Selecting a pair  $(x_i, x_j)$  from the neighboring chains, i.e.,  $|i - j| = 1$ . Let

$x'_i = x_j$  and  $x'_j = x_i$ , and compute the Metropolis ratio

$$\log r_e = \log \frac{\pi(\mathbf{x}')T(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x})T(\mathbf{x}'|\mathbf{x})} = [\log \pi(x_j) - \log \pi(x_i)]\left(\frac{1}{t_i} - \frac{1}{t_j}\right) + \log \frac{T(\mathbf{x}|\mathbf{x}')}{T(\mathbf{x}'|\mathbf{x})}.$$

Note that the transition probability here is symmetric, since if we let  $p(x_i)$  be the probability of selecting  $x_i$ , and let  $w(x_j|x_i)$  be the probability that  $x_j$  is chosen to be exchanged with  $x_i$ , then

$$T(\mathbf{x}'|\mathbf{x}) = p(x_i)w(x_j|x_i) + p(x_j)w(x_i|x_j).$$

Therefore  $T(\mathbf{x}'|\mathbf{x}) = T(\mathbf{x}|\mathbf{x}')$ .

Note that in the EMC algorithm, each step can be run multiple times. For example, in the mutation step, Liang and Wong's algorithm ([39], Page 324) let each  $x_k$  to be updated independently using the mutation operation, and let the crossover operation repeat for  $[N/5]$  (the integer part of  $N/5$ ) times, and let the exchange operation repeat  $N$  times. Goswami and Liu's algorithm ([24], Page 25), however, performs mutation updates  $M$  times for each  $x_k$ , and performs crossover updates  $[N/2]$  times, and exchange updates  $N$  times.

# Bibliography

- [1] A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.
- [2] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.
- [3] R. B. Ash and M. F. Gardner. *Topics in Stochastic Processes*. Academic Press, New York, 1975.
- [4] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27:536–561, 1999.
- [5] S. Behseta, R. E. Kass, and Wallstrom G. L. Hierarchical models for assessing variability among functions. *Biometrika*, 95:419–434, 2005.
- [6] P. C. Besse, H. Cardot, R. Faivre, and M. Goulard. Statistical modelling of functional data. *Applied Stochastic Models in Business and Industry*, 21:165–173, 2005.

- [7] L. Bottolo and S. Richardson. Evolutionary stochastic search (paper downloaded from web:). [http://www.bgsx.org.uk/publications/Bottolo\\_Richardson\\_ESS.pdf](http://www.bgsx.org.uk/publications/Bottolo_Richardson_ESS.pdf), 2008.
- [8] P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, 60:627–641, 1998.
- [9] P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, 64:519–536, 2002.
- [10] B. Cai and D. B. Dunson. Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *Journal of the American Statistical Association*, 102:1158–1171, 2007.
- [11] H. Cardot. Nonparametric regression for functional responses with application to conditional functional principle component analysis. Available online: <http://www.lsp.ups-tlse.fr/Recherche/Publications/2005/car01.pdf>, 2005.
- [12] H. Cardot. Conditional functional principal components analysis. *Scandinavian Journal of Statistics*, 34:317–335, 2007.
- [13] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45:11–22, 1999.
- [14] H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92:24–41, 2005.

- [15] J. Chiou, H. Müller, and J. Wang. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B*, 65:405–423, 2003.
- [16] H. Chip, E. I. George, and R. E. McCulloch. The practical implementation of bayesian model selection (with discussion). *Model Selection*, 38:66–134, 2001.
- [17] M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.
- [18] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.
- [19] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer-Verlag, 2006.
- [20] D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo*. Chapman & Hall/CRC, New York, 2006.
- [21] E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 89:881–889, 1990.
- [22] E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistics Sinica*, 7:339–373, 2002.

- [23] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28:500–531, 2000.
- [24] G. Goswami and J. S. Liu. On learning strategies for evolutionary monte carlo. *Statistics and Computing*, 17:23–38, 2007.
- [25] P. M. Grambsch, B. L. Randall, R. M. Bostick, J. D. Potter, and T. A. Louis. Modeling the labeling index distribution: An application of functional data analysis. *Journal of the American Statistical Association*, 90:813–821, 1995.
- [26] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.
- [27] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society, Series B*, 68:109–126, 2006.
- [28] P. Hall, H. Müller, and J. Wang. Properties of principle component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34:1493–1517, 2006.
- [29] P. Hall and C. Vial. Assessing extrema of empirical principal component functions. *The Annals of Statistics*, 34:1518–1544, 2006.
- [30] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

- [31] T. Hastie and R. J. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757–796, 2003.
- [32] G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, 64(3):411–432, 2002.
- [33] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.
- [34] H. Kuo. *Gaussian Measures in Banach Spaces*. Springer-Verlag, New York, 1975.
- [35] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, Boston, 2005.
- [36] J. S. Lee. *Aspects of Functional Data Inference and Its applications*. Ph.D. Thesis, Houston, TX, 2006.
- [37] S. Lee, W. Zhang, and Song X. Estimating the covariance function with functional data. *British Journal of Mathematical and Statistical Psychology*, 55:247–261, 2002.
- [38] Y. Li and T. Hsing. On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98:1782–1804, 2007.
- [39] F. Liang and W. H. Wong. Evolutionary monte carlo: applications to  $c_p$  model sampling and change point problem. *Statistica Sinica*, 10:317–342, 2000.

- [40] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2001.
- [41] N. Malfait and J. O. Ramsay. The historical functional linear model. *The Canadian Journal of Statistics*, 31:115–128, 2003.
- [42] I. W. McKeague. A statistical model for signature verification. *Journal of the American Statistical Association*, 100:231–241, 2005.
- [43] L. Meier, S. Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.
- [44] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087–1092, 1953.
- [45] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, New York, 2002.
- [46] J. S. Morris, C. Arroyo, B. A. Coull, M. R. Louise, R. Herrick, and S.L Gortmaker. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, 101:1352–1364, 2006.
- [47] J. S. Morris, P. J. Brown, R. C. Herrick, K. A. Baggerly, and K. R. Coombes. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64:479–489, 2008.

- [48] J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68:179–199, 2006.
- [49] J. S. Morris, M. Vannucci, P. J. Brown, and R. J. Carroll. Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98:573–583, 2003.
- [50] R. J. Muirhead. *Aspects of Multivariate Statisticial Theory*. John Wiley and Sons, Inc, 1982.
- [51] H. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805, 2005.
- [52] D. J. Nott and R. Kohn. Adaptive sampling for bayesian variable selection. *Biometrika*, 92(4):747–763, 2005.
- [53] D. J. Nott and D. Leonte. Sampling schmemes for bayesian variable selection in generalized linear models. *Journal of Computational & Graphical Statistics*, 13(2):362–382, 2004.
- [54] F.A. Ocaña, A. M. Aguilera, and M. J. Valderrama. Functional principal components analysis by choice of norm. *Journal of Multivariate Analysis*, 71:262–276, 1999.
- [55] T. Park. A penalized likelihood approach to rotation of principal components. *Journal of Computational and Graphical Statistics*, 14:867–888, 2005.

- [56] G. D. Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer, New York, 2006.
- [57] M. F. Ramanujam, N. and Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum. Spectroscopic diagnosis of cervical intraepithelial neoplasia(cin) in vivo using laser induced fluorescence spectra at multiple excitation wavelengths. *Lasers Surg. Med.*, 19:63–67, 1996.
- [58] J. O. Ramsay. When the data are functions. *Psychometrika*, 47:379–396, 1982.
- [59] J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B*, 53:539–572, 1991.
- [60] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 1997.
- [61] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, New York, 2002.
- [62] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, Section Edition*. Springer, New York, 2005.
- [63] S. I. Resnick. *A Probability Path*. Birkhäuser, Boston, 2001.
- [64] J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, 53:233–243, 1991.

- [65] R. Richards-Kortum and N. Ramanujam. Optical spectroscopy and imaging for detection of disease. <http://www.eng.uct.ac.cy/biaolab/Education/tutorials/>.
- [66] R. T. Rockafellar. *Convex Analysis*. Academic Press, Princeton University Press, 1970.
- [67] M. J. Schervish. *Theory of Statistics*. Springer-Verlag, New York, 1995.
- [68] B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24:1–24, 1996.
- [69] V. Siskin. Second moments of inverse wishart-matrix elements. *Biometrika*, 59:690–691, 1972.
- [70] J. G. Staniswalis and J. J. Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93:1403–1418, 1998.
- [71] W. K. Thompson and R. Ori. A bayesian model for sparse functional data. *Biometrics*, 64:54–63, 2008.
- [72] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
- [73] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762, 1994.
- [74] L. N. Trefethen. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

- [75] S. Walker. New approaches to bayesian consistency. *The Annals of Statistics*, 32:2028–2043, 2004.
- [76] S. G. Walker, Lijoi A., and Prünster I. On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35:738–746, 2007.
- [77] Y. Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005.
- [78] F. Yao. Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *Journal of Multivariate Analysis*, 98:40–56, 2007.
- [79] F. Yao, Müller H., Clifford A. J., S. R. Dueker, J. Follett, Y. Lin, B. A. Buchholz, and J.S. Vogel. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59:676–685, 2003.
- [80] F. Yao and T. C. M. Lee. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society, Series B*, 68:3–25, 2006.
- [81] K. Yosida. *Functional Analysis, Sixth Edition*. Springer-Verlag, New York, 1980.
- [82] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

- [83] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.
- [84] X. Zhou and N. A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc., New York, 2002.
- [85] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.
- [86] M. H. Zweig and G. Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577, 1993.