ABSTRACT

LI, XIUQI. Bayesian Classification and Change Point Detection for Functional Data. (Under the direction of Subhashis Ghosal).

In the first part of the dissertation, we propose a Bayesian approach to estimating parameters in multiclass functional models. Unordered multinomial probit, ordered multinomial probit and multinomial logistic models are considered. We use finite random series priors based on a suitable basis such as B-splines in these three multinomial models, and classify the functional data using the Bayes rule. We average over models based on the marginal likelihood estimated from Markov Chain Monte Carlo (MCMC) output. Posterior contraction rates for the three multinomial models are computed. We also consider Bayesian linear and quadratic discriminant analyses on the multivariate data obtained by applying a functional principal component technique on the original functional data. A simulation study is conducted to compare these methods on different types of data. We also apply these methods to phoneme dataset.

In the second part, we propose a Bayesian method to detect change points for functional data. We extract the features of a sequence of functional data by the discrete wavelet transform (DWT), and treat each sequence of feature independently. We believe that there is potentially a change in each feature at possibly different time points. The functional data evolves through such changes throughout the sequence of observations. The change point for this sequence of functional data is the cumulative effect of changes in all features. We assign the features with priors which incorporate the characteristic of the wavelet coefficients. Then we compute the posterior distribution of change point for each sequence of feature, and define a matrix where each entry is a measure of similarity between two functional data in this sequence. We compute the ratio of the mean
similarities between groups and within groups for all possible partitions, and the change point is where the ratio reaches the minimum. We state a posterior consistency result for estimation and model selection. We implement this method on a set of simulated data and a dataset on climate change. We demonstrate that this method can be inherently extended to multiple change points detection.
Bayesian Classification and Change Point Detection for Functional Data

by

Xiuqi Li

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

Operations Research

Raleigh, North Carolina

2018

APPROVED BY:

Sujit Ghosh

Arnab Maity

Negash Medhin

Subhashis Ghoshal
Chair of Advisory Committee
DEDICATION

To my parents.
BIOGRAPHY

Xiuqi was born and raised in Chengdu, Sichuan, China. She obtained her Bachelor’s degree in Transportation Engineering from Southwest Jiaotong University in 2013. In the August of the same year, she was enrolled in the Master program of Operations Research at North Carolina State University. After passing the qualifying exam during the summer in 2014, she switched to the Ph.D. program in Operations Research. Under the direction of Dr. Subhashis Ghoshal, Xiuqi defended her dissertation on August 9, 2018.
ACKNOWLEDGEMENTS

I would like to express my deepest and sincerest gratitude to my advisor Dr. Subhashis Ghoshal for his great guidance, patience, and expertise. This dissertation would not have been possible without his advice and support.

I am very grateful to Dr. Sujit Ghosh, Dr. Arnab Maity, and Dr. Negash Medhin for their serving on my committee, with additional thanks to Dr. Sujit Ghosh for his serving as the graduate school representative. I appreciate their valuable comments and suggestions for my research and presentation.

I am also very grateful to the faculty and staff of Operations Research Program and Statistics Department. Especially, I would like to thank Dr. Negash Medhin for his help to find funding for me. I would like to thank Ms. Linda Smith and Ms. Lanakila Alexander for their help. I also want to thank Dr. Ana–Maria Staicu for her time and valuable advice.

Special thanks to all my friends who always support and encourage me.

Last but not the least, I am deeply grateful to my parents for their unconditional love and support. They always stand beside me whenever I face difficulties in life.
# TABLE OF CONTENTS

**LIST OF TABLES** .......................................................... vii

**LIST OF FIGURES** ........................................................ viii

Chapter 1 Introduction .................................................. 1

Chapter 2 Bayesian Classification of Multiclass Functional Data .. 5
  2.1 Introduction .......................................................... 5
  2.2 Model ................................................................. 8
    2.2.1 Ordered Multinomial Probit Model ......................... 8
    2.2.2 Unordered Multinomial Probit Model ..................... 8
    2.2.3 Multinomial Logistic Model ............................... 9
  2.3 Finite Random Series Prior ..................................... 10
    2.3.1 Ordered Multinomial Probit Model ....................... 11
    2.3.2 Unordered Multinomial Probit Model ................... 13
    2.3.3 Multinomial Logistic Model ............................. 15
  2.4 Marginal Likelihood and Model Averaging ..................... 16
    2.4.1 Ordered Multinomial Probit Model ....................... 18
    2.4.2 Unordered Multinomial Probit Model ................... 19
    2.4.3 Multinomial Logistic Model ............................. 21
  2.5 Posterior Contraction Rate ..................................... 22
    2.5.1 Ordered Multinomial Probit Model ....................... 26
    2.5.2 Unordered Multinomial Probit Model ................... 30
    2.5.3 Multinomial Logistic Model ............................. 32
  2.6 Discriminant Analysis .......................................... 32
    2.6.1 Linear Discriminant Analysis ........................... 33
    2.6.2 Quadratic Discriminant Analysis ....................... 34
  2.7 Simulation ......................................................... 35
    2.7.1 Data Generation ............................................ 35
    2.7.2 Basis Functions ........................................... 37
    2.7.3 Results .................................................... 38
  2.8 Application on Phoneme Data .................................. 39
    2.8.1 Three Classes .............................................. 41
    2.8.2 Five Classes .............................................. 44

Chapter 3 Bayesian Change Point Detection for Functional Data .... 49
  3.1 Introduction ....................................................... 49
  3.2 Model ............................................................. 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Prior Distributions</td>
<td>53</td>
</tr>
<tr>
<td>3.4</td>
<td>Posterior Probabilities of Change Point</td>
<td>54</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Initial State</td>
<td>55</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Non-initial State</td>
<td>56</td>
</tr>
<tr>
<td>3.5</td>
<td>Change Point Detection</td>
<td>58</td>
</tr>
<tr>
<td>3.6</td>
<td>Posterior Consistency</td>
<td>59</td>
</tr>
<tr>
<td>3.7</td>
<td>Simulation</td>
<td>70</td>
</tr>
<tr>
<td>3.8</td>
<td>Application on Climate Data</td>
<td>71</td>
</tr>
</tbody>
</table>

**References**                                              | 79   |
## LIST OF TABLES

| Table 2.1 | Averaged misclassification rates for simulated data | 38 |
| Table 2.2 | Averaged misclassification rates for 3-class phoneme data | 42 |
| Table 2.3 | Estimate and standard error of the posterior mean for the 3-class ordered multinomial model ($J = 6$) | 42 |
| Table 2.4 | Estimate and standard error of the posterior mean for the 3-class multinomial logistic model ($J = 10$) | 43 |
| Table 2.5 | Estimate and standard error of the posterior mean for the 3-class unordered multinomial model ($J = 14$) | 43 |
| Table 2.6 | Averaged misclassification rates for 5-class phoneme data | 44 |
| Table 2.7 | Estimate and standard error of the posterior mean for the 5-class unordered multinomial model ($J = 14$) | 46 |
| Table 2.8 | Estimate and standard error of the posterior mean for the 5-class multinomial logistic model ($J = 14$) | 47 |
| Table 2.9 | Estimate and standard error of the posterior mean for the 5-class ordered multinomial model ($J = 13$) | 48 |
| Table 3.1 | Compatible models | 64 |
| Table 3.2 | Change point detection results comparison | 71 |
# LIST OF FIGURES

| Figure 2.1 | Coefficient functions for the multinomial models | 36 |
| Figure 2.2 | Curves for each of the five phoneme classes | 40 |
| Figure 2.3 | $\gamma_2$ sampled from Metropolis-Hastings when $J=5-7$ and $13-15$ | 42 |
| Figure 2.4 | $\gamma_2, \gamma_3,$ and $\gamma_4$ sampled from Metropolis-Hastings when $J=5, 10, 15$ | 45 |
| Figure 3.1 | Land-surface average temperature curves between 1753–2016 | 72 |
| Figure 3.2 | Land-surface average temperature curves for every 66 years | 74 |
| Figure 3.3 | $C(k)$ for different $k$ between 1753–2016 | 75 |
| Figure 3.4 | The plot on the left is the land-surface average temperature curves between 1753–1913. The plot on the right is the land-surface average temperature curves between 1914–2016. | 75 |
| Figure 3.5 | $C(k)$ for different $k$ between 1753–1913 | 76 |
| Figure 3.6 | The plot on the left is the land-surface average temperature curves between 1753–1838. The plot on the right is the land-surface average temperature curves between 1839–1913. | 76 |
| Figure 3.7 | $C(k)$ for different $k$ between 1914–2016 | 77 |
| Figure 3.8 | The plot on the left is the land-surface average temperature curves between 1914–1968. The plot on the right is the land-surface average temperature curves between 1969–2016. | 77 |
| Figure 3.9 | Subgroups between 1753–2016 | 78 |
Chapter 1

Introduction

Functional data analysis (FDA) deals with the analysis of data occurring in the form of functions. Wang et al. (2016) gave an overview of FDA including functional principal component analysis, functional linear regression, clustering and classification of functional data. FDA is increasingly drawing attention in many areas, such as biomedicine, environmental studies, and economics (Ullah and Finch, 2013). Mallor, Moler, and Urmeneta (2017) proposed a model based on functional principal component analysis to predict household electricity consumption. Wagner-Muns et al. (2018) proposed a method that uses functional principal components analysis to forecast traffic volume.

Classification of functional data, especially when the data units can come from more than two categories, is a fundamental problem of interest. Generalized linear models are often used to classify the functional data (Müller and Stadtmüller, 2005; James, 2002). The linear discriminant analysis is also used for functional data classification (James and Hastie, 2001). Preda, Saporta, and Lévéder (2007) proposed the partial least squares regression on functional data for linear discriminant analysis. Rossi and Villa (2006) adapted support vector machines to functional data classification. Li and Yu (2008) pro-
posed a functional segment discriminant analysis (FSDA), which combines the classical linear discriminant analysis and support vector machines. Wavelets approaches are also applied to classify and cluster functional data (Ray and Mallick, 2006; Antoniadis et al., 2013; Chang, Chen, and Ogden, 2014; Suarez and Ghosal, 2016). There are also non-parametric approaches for functional data classification (Biau, Bunea, and Wegkamp, 2005; Ferraty and Vieu, 2003). Mosler and Mozharovskyi (2017) introduced a nonparametric procedure that transformed the original functional data into the DD-plot through a two-step transformation, and hence the classification operated on a low-dimensional space. Delaigle and Hall (2013) introduced methods to classify the functional data not observed on the same interval. Lee, Shin, and Lee (2016) proposed a logistic regression for functional data classification that is robust to misdiagnosis or label noise. Epifanio (2008) proposed several shape descriptors for classifying functional data. Alonso, Casado, and Romo (2012) proposed a method based on the distances from the observation or its derivatives to group representative functions for discriminating functional data. Nguyen, McLachlan, and Wood (2016) developed a method to classify bivariate functional data by combining the current techniques in spatial spline regression with finite mixture models and mixed-effects models. Chamroukhi, Glotin, and Samé (2013) proposed a mixture discriminant analysis approach based on hidden process regression to handle the problem of complex-shaped classes of curves, where each class is potentially composed of several sub-classes.

However, there are only a few approaches proposed in the context of Bayesian classification for functional data. Wang, Ray, and Mallick (2007) developed a Bayesian hierarchical model which combines the adaptive wavelet-based function estimation and the logistic classification. Zhu, Vannucci, and Cox (2010) proposed a Bayesian hierarchical model that takes into account random batch effects and selects effective functions

In Chapter 2, we propose a Bayesian approach to estimating parameters in multi-class functional models. Unordered multinomial probit, ordered multinomial probit and multinomial logistic models are considered. We use finite random series priors based on a suitable basis such as B-splines in these three multinomial models, and classify the functional data using the Bayes rule. We average over models based on the marginal likelihood estimated from Markov Chain Monte Carlo (MCMC) output. Posterior contraction rates for the three multinomial models are computed. We also consider Bayesian linear and quadratic discriminant analyses on the multivariate data obtained by applying a functional principal component technique on the original functional data. A simulation study is conducted to compare these methods on different types of data. We also apply these methods to a phoneme dataset.

Another important aspect in functional data analysis is change point detection. Berkes et al. (2009) developed a method that works with the difference of mean functions projected on the principal components of the data. Zhang et al. (2011) developed a self-normalization (SN) based test to identify potential change points in the dependence structure of functional observations. Aston and Kirch (2012) also proposed a method to detect change points for dependent functional observations, and they were particularly interested in the case where the change point is an epidemic change (a change occurs and then the observations return to baseline at a later time). Sharipov, Tewes, and Wendler (2016) developed a new test for structural changes in functional data that based
on Hilbert space theory and critical values are deduced from bootstrap iterations. Aue, Rice, and Sönmez (2018) proposed a method to uncover structural breaks in functional data that does not rely on dimension reduction techniques.

In Chapter 3, we propose a Bayesian method to detect change points for functional data. We extract the features of a sequence of functional data by the discrete wavelet transform (DWT) and treat each sequence of feature independently. We believe there is potentially a change in each feature at possibly different time points. The functional data evolves through such changes throughout the sequence of observations. The change point for this sequence of functional data is the cumulative effect of changes in all features. We assign the features with priors which incorporate the characteristic of the wavelet coefficients. Then we compute the posterior distribution of change point for each sequence of feature, and define a matrix where each entry is a measure of similarity between two functional data in this sequence. We compute the ratio of the mean similarities between groups and within groups for all possible partitions, and the change point is where the ratio reaches the minimum. This method can be inherently extended to multiple change points detection.
Chapter 2

Bayesian Classification of Multiclass Functional Data

2.1 Introduction

Classification of functional data, especially when the data units can come from more than two categories, is a fundamental problem of interest. Generalized linear models are often used to classify the functional data (Müller and Stadtmüller, 2005; James, 2002). The linear discriminant analysis is also used for functional data classification (James and Hastie, 2001). Preda, Saporta, and Lévéder (2007) proposed the partial least squares regression on functional data for linear discriminant analysis. Rossi and Villa (2006) adapted support vector machines to functional data classification. Li and Yu (2008) proposed a functional segment discriminant analysis (FSDA), which combines the classical linear discriminant analysis and support vector machines. Wavelets approaches are also applied to classify and cluster functional data (Ray and Mallick, 2006; Antoniadis et al., 2013; Chang, Chen, and Ogden, 2014; Suarez and Ghosal, 2016). There are also nonparametric approaches

In this chapter, we consider a response $Y$ taking values $k = 1, \ldots, K$, with functional covariate $\{X(t), t \in [0, 1]\}$. The main problem is to estimate the probability $P(Y = k|X)$, which can be conveniently modeled by a function of $\int \beta(t)X(t)dt$

$$P(Y = k|X) = H_k \left( \int \beta(t)X(t)dt \right),$$

where $H_k$ is a cumulative distribution function, and $\beta(\cdot)$ is an unknown (possibly vector of) coefficient function(s). Unordered multinomial probit, ordered multinomial probit and multinomial logistic models are considered in this chapter which correspond to different choices of $H_k, k = 1, \ldots, K$. For an ordered multinomial probit model, there are additional order restrictions. Finite random series priors (Shen and Ghosal, 2015) are applied to the three multinomial models. We compare these methods with Bayesian linear and quadratic discriminant analyses applied on the data reduced to multivariate form by a functional principal component technique. Following a Bayesian approach, the posterior distribution
of the parameters are obtained using the training data, and then the classification rules are applied to the test data using the posterior probability of class membership.

The primary goal of a basis expansion method is to reduce a more complex problem to a simpler problem which has either a known solution or is likely to be easier to solve. A prior on function through finite random series is a standard tool in nonparametric Bayesian inference, but in the context of functional data, the technique has not been utilized to its fullest potential, especially regarding the study of theoretical property of Bayesian methods. Only one paper (Shen and Ghosal, 2015) has one example of functional linear regression treated using finite random series priors. We take that idea but develop it in the context of functional data classification. Characterizing contraction rate is a major goal of this chapter. For this, we need to estimate the complexity of the model and the prior concentration. Even though, the model reduces to the finite dimensional setting from the computational point of view, the effect of the residual bias in the approximation of function must be properly addressed. Hence the treatment substantially different from that of a parametric problem. In particular, the dimension of the basis must be adapted with the smoothness and the sample size by using a prior on it.

The chapter is organized as follow. In Section 2.2, the three functional multinomial models are described. Section 2.3 gives the description of applying the finite random series prior to these models. The marginal likelihood estimation is described in Section 2.4. In Section 2.5, the posterior contraction rates of the three functional multinomial models with finite random series prior are computed. Section 2.6 describes the Bayesian discriminant analysis of functional data, which is used to compare with the proposed models. In Section 2.7, a simulation study is conducted on various types of data. In Section 2.8, the three multinomial models and Bayesian discriminant analysis are tested on phoneme dataset.
2.2 Model

2.2.1 Ordered Multinomial Probit Model

Let $X_i(t)$, $i = 1, \ldots, n$, $t \in [0, 1]$, be the observed functional data associated with a categorical variable $Y_i$ taking possible values $1, \ldots, K$. We assume that $(X_i, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d) observations.

Following Albert and Chib (1993), we consider the model described implicitly as follows: there exists a latent variable $W_i$ distributed as $N(\int \beta(t)X_i(t)dt, 1)$, for $i = 1, \ldots, n$, and that $Y_i = k$ if $\gamma_{k-1} < W_i \leq \gamma_k$, where $k = 1, \ldots, K$. The latent variables $W_i$, $i = 1, \ldots, n$, are independent. The coefficient function $\beta(\cdot)$ is unknown. The cut-points $\gamma_k$ are also unknown except that $\gamma_0 = -\infty$ and $\gamma_K = \infty$. To ensure identifiability, we set $\gamma_1 = 0$. Under the assumed model, the probability of choosing a category $k$ is given by

$$P(Y_i = k|X_i) = \Phi\left(\gamma_k - \int \beta(t)X_i(t)dt\right) - \Phi\left(\gamma_{k-1} - \int \beta(t)X_i(t)dt\right), \quad (2.2)$$

where $\Phi$ stands for the distribution function of the standard normal distribution.

2.2.2 Unordered Multinomial Probit Model

Let $X_i(t)$, $i = 1, \ldots, n$, be the same as in the Section 2.2.1, and also same for Section 2.2.3.

The unordered multinomial probit model can be described by the following data augmentation method. As in Albert and Chib (1993), let $W_i' = (W_i', \ldots, W_i')^T$, $i = 1, \ldots, n$, be latent variable, such that $W_i'$ follows a linear model

$$W_i' = \int \beta(t)X_i(t)dt + \epsilon_i', \quad (2.3)$$
where $\varepsilon_{il}' \sim N(0, 1)$, $i = 1, \ldots, n$, $l = 1, \ldots, K$, are i.i.d. standard normal random variables. Consider the latent variables $W_i = (W_{i1}, \ldots, W_{iK-1})^T$, $W_{il} = W_{il}' - W_{iK}'$,

$$W_{il} = \int \beta_i'(t)X_i(t)dt - \int \beta_{K}'(t)X_i(t)dt + \varepsilon_{il},$$  

(2.4)

where $\varepsilon_{il} = \varepsilon_{il}' - \varepsilon_{iK}'$, and $l = 1, \ldots, K - 1$. Let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK-1})^T$. Then $\varepsilon_i$ follows $N(0, \Sigma)$, where $\Sigma$ is a $(K - 1) \times (K - 1)$ matrix with 2 at diagonal entries and 1 at all off-diagonal entries.

The probability of choosing the $k$th ($k = 1, \ldots, K - 1$) alternative is given by

$$P(Y_i = k|X_i) = P(W_{ik} > W_{il}, \text{ for all } l \neq k, \text{ and } W_{ik} > 0),$$  

(2.5)

and the probability of choosing alternative $K$ is given by

$$P(Y_i = K|X_i) = P(W_{il} < 0 \text{ for all } l = 1, \ldots, K - 1).$$  

(2.6)

### 2.2.3 Multinomial Logistic Model

In this model, the probability of choosing category $k$ is given by

$$P(Y_i = k|X_i) = \frac{\exp[\int \beta_k(t)X_i(t)dt]}{\sum_{l=1}^{K} \exp[\int \beta_l(t)X_i(t)dt]}.$$  

(2.7)

To ensure model identification, set $\beta_K(t) = 0$. Then the probability of choosing category $k$ ($k = 1, \ldots, K - 1$) is given by

$$P(Y_i = k|X_i) = \frac{\exp[\int \beta_k(t)X_i(t)dt]}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t)X_i(t)dt]}.$$  

(2.8)
and the probability of choosing category $K$ is given by

$$P(Y_i = K | X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[\int \beta_l(t)X_i(t)dt]}.$$  \hspace{1cm} (2.9)

### 2.3 Finite Random Series Prior

The functional coefficient $\beta(t)$ (or $\beta_1(t), \ldots, \beta_K(t)$ for unordered multinomial probit and multinomial logistic models) is given a prior which is a finite linear combination of a certain chosen basis functions: $\beta(t) = \sum_{j=1}^{J} \theta_j \psi_j(t)$, where $\{\psi_1(t), \ldots, \psi_J(t)\}$ is a basis, for example, formed by B-splines, Fourier functions, or wavelets. A prior is put on the unknown coefficients ($\theta_1, \ldots, \theta_J$). The number of basis function $J$ is also unknown and should be given a hyperprior. Instead of sampling across the different dimensions using reversible jump MCMC (Green, 1995) which has computational difficulty for complicated models, we can implement MCMC for a given $J$ value, and repeat it for relevant $J$ values. Thus, we can compute the marginal likelihood $m(Y|J)$ for potentially interesting values of $J$, and obtain the posterior probability of $J$, which are discussed in Section 2.4.

The advantage of using a finite random series prior is that the inner product between the functional coefficient and the functional data $\int \beta(t)X_i(t)dt$ is reduced to a simple linear combination

$$\int \beta(t)X_i(t)dt = \int \sum_{j=1}^{J} \theta_j \psi_j(t)X_i(t)dt = \sum_{j=1}^{J} \theta_j Z_{ij},$$  \hspace{1cm} (2.10)

where $Z_{ij} = \int \psi_j(t)X_i(t)dt$ is known, and can be computed by Simpson’s rule.
2.3.1 Ordered Multinomial Probit Model

Using a finite random series $\beta(t) = \sum_{j=1}^{J} \theta_j \psi_j(t)$, the model in (2.2) can be rewritten as

$$
P(Y_i = k | X_i) = \Phi\left( \gamma_k - \sum_{j=1}^{J} \theta_j Z_{ij} \right) - \Phi\left( \gamma_{k-1} - \sum_{j=1}^{J} \theta_j Z_{ij} \right),$$  \hspace{1cm} (2.11)

where $Z_{ij} = \int \psi_j(t) X_i(t) dt$.

Define $\theta = (\theta_1, \ldots, \theta_J)^T$, and $Z_i = (Z_{i1}, \ldots, Z_{iJ})^T$. Then (2.11) can be written compactly as

$$
P(Y_i = k | X_i) = \Phi(\gamma_k - Z_i^T \theta) - \Phi(\gamma_{k-1} - Z_i^T \theta).$$ \hspace{1cm} (2.12)

Clearly the unobserved latent variable $W_i$ follows $N_J(Z_i^T \theta, 1)$, where $N_J$ stands for the $J$-variate normal distribution. Assign a conjugate prior $\theta \sim N_J(\theta_0, B_0)$, where $\theta_0$ is $J \times 1$ mean vector, and $B_0$ is a $J \times J$ covariance matrix. Then the posterior distribution of $\theta$ is given by

$$
\theta|Y, W \sim N_J(\theta_n, B_n), \quad B_n = (B_0^{-1} + ZT Z)^{-1}, \quad \theta_n = B_n(B_0^{-1} \theta_0 + ZT W),
$$ \hspace{1cm} (2.13)

where $Z = (Z_1^T, \ldots, Z_n^T)^T$, and $W = (W_1, \ldots, W_n)^T$.

We follow the scheme introduced by Albert and Chib (1993). The posterior distribution of $W_i$ is given by

$$
W_i|(\theta, \gamma, Y_i = k) \sim TN(Z_i^T \theta, 1, \gamma_{k-1}, \gamma_k),
$$ \hspace{1cm} (2.14)

where $TN(Z_i^T \theta, 1, \gamma_{k-1}, \gamma_k)$ is the truncation of the (univariate) normal distribution with
mean $Z_i^T \theta$, and variance 1 to the interval $(\gamma_{k-1}, \gamma_k)$.

Albert and Chib (1993) assigned a diffuse prior on the cut-points. However, model averaging needs a proper prior. A normal prior is not appropriate due to the order restriction on $\gamma_1, \ldots, \gamma_K$. Albert and Chib (1997) proposed a transformation of $\gamma = (\gamma_1, \ldots, \gamma_K)$ which avoids the order restriction.

$$\alpha_1 = \log \gamma_2, \quad \alpha_j = \log(\gamma_{j+1} - \gamma_j), \quad 2 \leq j \leq K - 2.$$ (2.15)

Note that $\gamma_1 = 0$ and by the inverse map

$$\gamma_j = \sum_{l=1}^{j-1} e^{\alpha_l}, \quad 2 \leq j \leq K - 1.$$ (2.16)

Then $\gamma$ can be reparameterized by $\alpha = (\alpha_1, \ldots, \alpha_{K-2})$. Assign a multivariate normal prior with mean $\alpha_0$, and covariance $A_0$ on $\alpha$. To sample $\gamma$, apply the following steps of Metropolis-Hastings algorithm.

1. Sample $\alpha'$ from a proposal distribution $q(\alpha', \alpha|Y, \theta, W)$. Here we allow the proposal density to depend on the data and the two remaining blocks for the convenience of computing the marginal likelihood in the future.

2. Move to $\alpha'$ from the current $\alpha$ with probability

$$\rho(\alpha, \alpha'|Y, \theta, W) = \min \left\{ \frac{f(Y|\alpha', \theta, W)\pi(\alpha', \theta)}{f(Y|\alpha, \theta, W)\pi(\alpha, \theta)} \frac{q(\alpha', \alpha|Y, \theta, W)}{q(\alpha, \alpha'|Y, \theta, W)}, 1 \right\}. \quad (2.17)$$

3. Compute $\gamma$ by the inverse map (2.16).

To implement the MCMC sampling, first draw $\gamma$ by the above steps. Then sample from the posterior distributions (2.14) and (2.13).
The values of $\gamma$ sampled from the Metropolis-Hastings algorithm converges quickly. We demonstrate it on the real data in Section 2.8 by plotting the sampling values of $\gamma$.

2.3.2 Unordered Multinomial Probit Model

Let $\beta'_l(t) = \sum_{j=1}^J \theta'_{lj} \psi_j(t)$, where $l = 1, \ldots, K$. Then (2.4) can be rewritten as

$$W_{il} = \sum_{j=1}^J \theta'_{lj} Z_{ij} - \sum_{j=1}^J \theta'_{Kj} Z_{ij} + \epsilon_{il} = \sum_{l=1}^J (\theta'_{jl} - \theta'_{lK}) Z_{ij} + \epsilon_{il}, \quad (2.18)$$

where $Z_{ij} = \int \psi_j(t) X_i(t) dt$.

Let $\theta_{ij} = \theta'_{ij} - \theta'_{Kj}$, where $j = 1, \ldots, J$. Define $\theta_l = (\theta_{l1}, \ldots, \theta_{lj})^T$, and $Z_i = (Z_{i1}, \ldots, Z_{ij})^T$. Then (2.18) is given by

$$W_{il} = Z_i^T \theta_l + \epsilon_{il}, \quad (2.19)$$

where $i = 1, \ldots, n$, $l = 1, \ldots, K - 1$.

Define a $J \times (K - 1)$ matrix $\Theta = (\theta_1, \ldots, \theta_{K-1})$. Then we have $W_i = Z_i^T \Theta + \epsilon_i$, where $W_i = (W_{i1}, \ldots, W_{iK-1})^T$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK-1})^T$, and $\epsilon_i \sim N(0, \Sigma)$.

In the model described in Section 2.2, $\Sigma$ is known with 2 on diagonal entries and 1 on all off-diagonal entries. The only parameter needs to be estimated is $\Theta$. In order to draw the matrix $\Theta$ using the Gibbs sampling, we can stack the data in a matrix form which is given by

$$W = Z \Theta + \epsilon, \quad (2.20)$$

where $W = (W_1^T, \ldots, W_n^T)^T$ is an $n \times (K - 1)$ matrix, $Z = (Z_1^T, \ldots, Z_n^T)^T$ is an $n \times J$
matrix, and \( \varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_n^T)^T \) is an \( n \times (K - 1) \) matrix.

This results in a matrix normal distribution. The density function of matrix normal distribution \( \text{MN}_{n \times p}(M, U, V) \) is

\[
(2\pi)^{-np/2}|V|^{-n/2}|U|^{-p/2} \exp \left( -\frac{1}{2} \text{tr}[V^{-1}(X - M)^TU^{-1}(X - M)] \right),
\]

where \( M \) is an \( n \times p \) mean matrix, \( U \) is an \( n \times n \) row variance matrix, \( V \) is a \( p \times p \) column variance matrix, \( \text{tr} \) stands for the trace of a matrix, and \( |U| \) and \( |V| \) denote the determinants of \( U \) and \( V \) respectively.

Thus \( W|\Theta \sim \text{MN}_{n \times (K-1)}(Z\Theta, I_n, \Sigma) \). Here the row variance-covariance matrix \( I_n \) is an identity matrix of rank \( n \), since \( W_1, \ldots, W_n \) are independent. We consider the matrix normal prior \( \Theta \sim \text{MN}_{J \times (K-1)}(U_0, V_0, \Sigma) \). By a standard conjugacy calculation, the posterior is given by

\[
\Theta|Y, W \sim \text{MN}_{J \times (K-1)}(U_n, V_n, \Sigma),
\]

\[
V_n = (Z^T Z + V_0^{-1})^{-1}, \quad U_n = V_n( Z^T W + V_0^{-1} U_0 ).
\]

To draw a sample of \( W \), we use the method introduced by McCulloch and Rossi (1994). Let \( W_{i,-l} \) denote \( (W_{i1}, \ldots, W_{i,l-1}, W_{i,l+1}, \ldots, W_{i(K-1)})^T \), \( Z_{i,-l} \) denote the \( i \)th row of \( Z \), the vector \( \Theta_{l,-l} \) denote the \( l \)th column of \( \Theta \), the matrix \( \Theta_{l,-l} \) denote \( \Theta \) without the \( l \)th column, the scalar \( \Sigma_{l,l} \) denote the \((l,l)\)th entry of \( \Sigma \), \( \Sigma_{l,-l} \) denote \( \Sigma \) without the \( l \)th row and the \( l \)th column, \( \Sigma_{-l,l} \) denote the \( l \)th column of \( \Sigma \) without the \( l \)th entry, and \( \Sigma_{-l,-l} \) denote the \( l \)th row of \( \Sigma \) without the \( l \)th entry. We draw \( W_{il} \) from the conditional truncation of the normal distribution with the mean \( m_{il} \) and variance \( \tau_{il}^2 \) to the interval
(a, b) described below:

\[ W_{il}(W_{i, -, l}, \Theta, Y_i) \sim TN(m_{il}, \tau_{il}^2, a, b), \]

\[ m_{il} = Z_i \cdot \Theta_{il} + \Sigma_{l,l}^{-1} \Sigma_{l, -, l}^{-1} (W_{i, -, l} - Z_i \cdot \Theta_{il}), \]

\[ \tau_{il}^2 = \Sigma_{l,l} - \Sigma_{l, -, l} \Sigma_{l, -, l}^{-1} \Sigma_{l,l}, \]

\[ (a, b) = \begin{cases} 
  (\max\{W_{i, -, l}, 0\}, \infty), & \text{if } Y_i = l, \ l = 1, \ldots, K - 1, \\
  (-\infty, \max\{W_{i, -, l}\}), & \text{if } Y_i \neq l, \ l = 1, \ldots, K - 1, \\
  (-\infty, 0), & \text{if } Y_i = K,
\end{cases} \quad (2.23) \]

\[ i = 1, \ldots, n, \ l = 1, \ldots, K - 1. \]

To implement the Gibbs sampling, we draw samples from (2.22) and (2.23).

### 2.3.3 Multinomial Logistic Model

Let \( \beta_k(t) = \sum_{j=1}^{J} \theta_{kj} \psi_j(t) \). Then (2.8) and (2.9) can be rewritten as

\[ P(Y_i = k|X_i) = \frac{\exp\left[\sum_{j=1}^{J} \theta_{kj} Z_{ij}\right]}{1 + \sum_{l=1}^{K-1} \exp\left[\sum_{j=1}^{J} \theta_{lj} Z_{ij}\right]}, \ k = 1, \ldots, K - 1, \quad (2.24) \]

\[ P(Y_i = K|X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\left[\sum_{j=1}^{J} \theta_{lj} Z_{ij}\right]}, \quad (2.25) \]

where \( Z_{ij} = \int \psi_j(t) X_i(t) dt \).

Define \( \theta_k = (\theta_{k1}, \ldots, \theta_{kJ})^T, k = 1, \ldots, K - 1, \) and \( Z_i = (Z_{i1}, \ldots, Z_{iJ})^T \). Then (3.24) and (3.25) are given by

\[ P(Y_i = k|X_i) = \frac{\exp\left[Z_i^T \theta_k\right]}{1 + \sum_{l=1}^{K-1} \exp\left[Z_i^T \theta_l\right]}, \ k = 1, \ldots, K - 1, \quad (2.26) \]
\[ P(Y_i = K|X_i) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp[Z_i^T \theta_l]} \]  

(2.27)

For each \( \theta_k, k = 1, \ldots, K - 1 \), we assign a multivariate normal prior \( N_J(\mu_k, \Sigma_k) \), and apply Metropolis-Hastings algorithm to sample \( \theta_k \).

1. Sample \( \theta_k' \) from the proposal distribution \( q(\theta_k', \theta_k|Y, \theta_{-k}) \).

2. Move to \( \theta_k' \) from the current \( \theta_k \) with probability

\[
\rho(\theta_k, \theta_k'|Y, \theta_{-k}) = \min\left\{ \frac{f(Y|\theta_k', \theta_{-k}) \pi(\theta_k', \theta_{-k}) \, q(\theta_k', \theta_k|Y, \theta_{-k})}{f(Y|\theta_k, \theta_{-k}) \pi(\theta_k, \theta_{-k}) \, q(\theta_k, \theta_k'|Y, \theta_{-k})}, 1 \right\},
\]

(2.28)

where \( \theta_{-k} \) denotes all the blocks except the \( k \)th one.

### 2.4 Marginal Likelihood and Model Averaging

In Section 2.3, we described the MCMC sampling technique for a given \( J \) value, which we need to repeat it for all possible \( J \) values. In the actual computation, however, it is impossible to consider all values of \( J \). With a given prior on \( J \), for example, geometric or Poisson distribution, the posterior probabilities for very small or very large values of \( J \) decay to zero very quickly. Thus, we do not need to consider these \( J \) values. Let \( J_1, \ldots, J_S \) denote the values of \( J \) we need to consider. If we can get the marginal likelihood \( m(Y|J_s) \), then we can compute the posterior probability of \( J_s \) using Bayes’s rule

\[
P(J_s|Y) = \frac{m(Y|J_s)p(J_s)}{\sum_{l=1}^{S} m(Y|J_l)p(J_l)},
\]

(2.29)

where \( p(J_s), s = 1, \ldots, S \), is the prior probability for \( J = J_s \).
For each given $J_s$, we have a misclassification rate $r_s$, which is defined as the ratio of the number of falsely classified data to the total number of data. Then we can obtain the average misclassification rate $\bar{r}$ for each multinomial model:

$$\bar{r} = \frac{1}{S} \sum_{s=1}^{S} P(J_s | Y) \cdot r_s.$$  \hspace{1cm} (2.30)

We call it the model averaging method.

The marginal likelihood can be written as the normalizing constant of the posterior density

$$m(Y | J_s) = \frac{f(Y | J_s, B) \pi(B | J_s)}{\pi(B | Y, J_s)},$$  \hspace{1cm} (2.31)

where $B$ is a convenient value of the parameter in the context of the support of the posterior distribution such as the posterior mean, because (2.31) holds for any $B$. The numerator is the product of the likelihood and the prior. The denominator is the posterior density of $B$. For a given $B^*$, the posterior density $\pi(B^* | Y, J_m)$ can be estimated from the Gibbs output (Chib, 1995) and the Metropolis-Hasting output (Chib and Jeliazkov, 2001). Then the estimated marginal likelihood in the logarithm scale is

$$\log \hat{m}(Y | J_s) = \log f(Y | J_s, B^*) + \log \pi(B^* | J_s) - \log \hat{\pi}(B^* | Y, J_s).$$  \hspace{1cm} (2.32)

The following sections give the details for $\pi(B^* | Y, J_s)$ estimation.
2.4.1 Ordered Multinomial Probit Model

There are two parameter blocks in this model, \( \theta \) and \( \alpha \), where \( \alpha \) is the transformation of \( \gamma \) as in (2.15). Given \( \theta^* = G^{-1} \sum_{g=1}^{G} \theta^{(g)} \), and \( \alpha^* = G^{-1} \sum_{g=1}^{G} \alpha^{(g)} \), where \( \{\theta^{(g)}, \alpha^{(g)}\}_{g=1}^{G} \) are from the MCMC output, the joint posterior density can be written as

\[
\pi(\theta^*, \alpha^* | Y, J_s) = \pi(\alpha^* | Y, J_s) \pi(\theta^* | Y, J_s, \alpha^*),
\]

(2.33)

where

\[
\pi(\theta^* | Y, J_s, \alpha^*) = \int \pi(\theta^* | Y, J_s, \alpha^*, W) \pi(W | Y, J_s, \alpha^*) dW.
\]

(2.34)

The Monte Carlo estimate of \( \pi(\theta^* | Y, J_s, \alpha^*) \) is

\[
\hat{\pi}(\theta^* | Y, J_s, \alpha^*) = M^{-1} \sum_{m=1}^{M} \pi(\theta^* | Y, J_s, \alpha^*, W^{(m)}),
\]

(2.35)

where \( \{W^{(m)}\}_{m=1}^{M} \) are sampled from distribution \([W | Y, J_s, \alpha^*]\). The draws of \( W \) from the Gibbs sampler are from the distribution \([W | Y, J_s]\), so \( \pi(\theta^* | Y, J_s, \alpha^*, W) \) cannot be averaged directly by the Gibbs sampling output. Additional sampling for \( W \) is needed.

We sample \( \{\theta^{(m)}\} \) from the density \( \pi(\theta | Y, J_s, \alpha^*, W) \), and given that \( \theta^{(m)} \), we sample \( \{W^{(m)}\} \) from \( \pi(W | Y, J_s, \theta, \alpha^*) \).

The explicit distribution of \( \alpha^* \) given \( (Y, J_s) \) is unknown, and hence the draws of \( \alpha \) are obtained from a Metropolis-Hastings sampling. By the local reversibility condition (see Chib and Jeliazkov (2001) for details), the posterior density of \( \alpha \) can be written as

\[
\pi(\alpha^* | Y, J_s) = \frac{E_1 \{\rho(\alpha, \alpha^* | Y, J_s, \theta, W)q(\alpha, \alpha^* | Y, J_s, \theta, W)\}}{E_2 \{\rho(\alpha^*, \alpha | Y, J_s, \theta, W)\}},
\]

(2.36)
where \( \rho(\alpha, \alpha^*|Y, J_s, \theta, W) \) is defined in (2.17), \( q(\alpha, \alpha^*|Y, J_s, \theta, W) \) is the proposal density, the expectation \( E_1 \) is with respect to the distribution \( \pi(\theta, \alpha, W|Y, J_s) \), and \( E_2 \) is with respect to the distribution \( \pi(\theta, W|Y, J_s, \alpha^*) \times q(\alpha^*, \alpha|Y, J_s, \theta, W) \).

Then an estimate of \( \pi(\alpha^*|Y, J_s) \) is given by

\[
G^{-1} \sum_{g=1}^{G} \frac{\rho(\alpha^{(g)}, \alpha^*|Y, J_s, \theta^{(g)}, W^{(g)}) q(\alpha^{(g)}, \alpha^*|Y, J_s, \theta^{(g)}, W^{(g)})}{M^{-1} \sum_{m=1}^{M} \rho(\alpha^*, \alpha^{(m)}|Y, J_s, \theta^{(m)}, W^{(m)})},
\]

(2.37)

where \( \{\theta^{(g)}, \alpha^{(g)}, W^{(g)}\}_{g=1}^{G} \) are obtained from the MCMC output. \( \{\theta^{(m)}, W^{(m)}\} \) are obtained from \( \pi(\theta|Y, J_s, \alpha^*, W) \) and \( \pi(W|Y, J_s, \theta, \alpha^*) \), and then given \( \{\theta^{(m)}, W^{(m)}\} \), sample \( \alpha^{(m)} \) from \( q(\alpha^*, \alpha|Y, J_s, \theta^{(m)}, W^{(m)}) \).

### 2.4.2 Unordered Multinomial Probit Model

The only unknown parameter is \( \Theta \). For \( \Theta^* = G^{-1} \sum_{g=1}^{G} \Theta^{(g)} \), where \( \{\Theta^{(g)}\} \) are from the Gibbs sampling output, the posterior density of \( \Theta \) at \( \Theta^* \) can be written as

\[
\pi(\Theta^*|Y, J_s) = \int \pi(\Theta^*|Y, J_s, W) \pi(W|Y, J_s) dW.
\]

(2.38)

Then the Monte Carlo estimate of \( \pi(\Theta^*|Y, J_s) \) is

\[
\hat{\pi}(\Theta^*|Y, J_s) = \sum_{g=1}^{G} \pi(\Theta^*|Y, J_s, W^{(g)}),
\]

(2.39)

where \( \{W^{(g)}\}_{g=1}^{G} \) are from the Gibbs sampling output.

For the unordered multinomial probit model, we also need to estimate the likelihood at some convenient values in the support of the posterior distribution. From Section 2.3.2,
\( \Theta = (\theta_1, \ldots, \theta_{K-1}) \), where \( \theta_l = \theta'_l - \theta'_K \), \( l = 1, \ldots, K-1 \). Then (2.6) can be rewritten as

\[
P(Y = K) = \frac{1}{(2\pi)^{(K-1)/2}|\Sigma|^{1/2}} \int_{-\infty}^{-\infty} \cdots \int_{-\infty}^{-\infty} \exp \left( -\frac{1}{2} U^T \Sigma^{-1} U \right) dU, \tag{2.40}\]

where \( \Theta_{:, l} \) denotes the \( l \)th column of \( \Theta \).

For \( l \neq K \), let \( \Theta^l = (\theta_1 - \theta_l, \ldots, \theta_{l-1} - \theta_l, \theta_{l+1} - \theta_l, \ldots, \theta_{K-1} - \theta_l, -\theta_l) \), then

\[
P(Y = l) = \frac{1}{(2\pi)^{(K-1)/2}|\Sigma|^{1/2}} \int_{-\infty}^{-\infty} \cdots \int_{-\infty}^{-\infty} \exp \left( -\frac{1}{2} U^T \Sigma^{-1} U \right) dU. \tag{2.41}\]

Due to the exchangeable correlation structure of \( \Sigma \), (2.41) can be reduced to a one dimensional integral (Dunnett, 1989) given by

\[
P(Y = l) = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \left\{ \prod_{k=1}^{K-1} \Phi(-u\sqrt{2} - Z^T \Theta^l_{:, k}) + \prod_{k=1}^{K-1} \Phi(u\sqrt{2} - Z^T \Theta^l_{:, k}) \right\} e^{-u^2} du. \tag{2.42}\]

The expression in (2.40) can also be reduced to the same form as in (2.42). Then (2.42) can be approximated by Gaussian quadrature as follow

\[
P(Y = l) \approx \frac{1}{2} \sum_{q=1}^{Q} w_q \left\{ \prod_{k=1}^{K-1} \Phi(-\sqrt{2x_q} - Z^T \Theta^l_{:, k}) + \prod_{k=1}^{K-1} \Phi(\sqrt{2x_q} - Z^T \Theta^l_{:, k}) \right\}, \tag{2.43}\]

where \( w_q \) and \( x_q \) are the weights and roots of the Laguerre polynomial of order \( Q \).

Thus, the likelihood of this unordered multinomial probit model can be approximated using (2.43).
2.4.3 Multinomial Logistic Model

There are $K - 1$ unknown parameters: $\theta_1, \ldots, \theta_{K-1}$. Given $\theta^*_k = G^{-1} \sum_{g=1}^G \theta^{(g)}_k$, $k = 1, \ldots, K - 1$, where $\{\theta^{(g)}_k\}_{g=1}^G$ are from the Metropolis-Hastings sampling output, the joint posterior density can be written as

$$\pi(\theta^*_1, \ldots, \theta^*_{K-1}|Y, J_s) = \prod_{i=1}^{K-1} \pi(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1}).$$  \hspace{1cm} (2.44)

By the local reversibility, each full conditional density can be written as

$$\pi(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1})$$

$$= \frac{E_1\{\rho(\theta_i, \theta^*_i|Y, J_s, \Psi^i_{i-1}, \Psi^{i+1})q(\theta_i, \theta^*_i|Y, J_s, \Psi^*_i, \Psi^{i+1})\}}{E_2\{\rho(\theta^*_i, \theta_i|Y, J_s, \Psi^*_{i-1}, \Psi^{i+1})\}},$$ \hspace{1cm} (2.45)

where $\Psi^*_{i-1} = (\theta_1, \ldots, \theta_{i-1})$, $\Psi^{i+1} = (\theta_{i+1}, \ldots, \theta_{K-1})$, $\rho(\theta_i, \theta^*_i|Y, J_s, \Psi^*_i, \Psi^{i+1})$ is defined in (2.28), $q(\theta_i, \theta^*_i|Y, J_s, \Psi^*_i, \Psi^{i+1})$ is the proposal density, $E_1$ is the expectaion with respect to the distribution $\pi(\theta_i, \Psi^{i+1}|Y, J_s, \Psi^*_i)$, and $E_2$ is that with respect to $\pi(\Psi^{i+1}|Y, J_s, \Psi^*_i, \theta^*_i) \times q(\theta^*_i, \theta_i|Y, J_s, \Psi^*_i, \Psi^{i+1})$.

Then an estimate of $\pi(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1})$ is given by

$$\hat{\pi}(\theta_i|Y, J_s, \theta^*_1, \ldots, \theta^*_{i-1})$$

$$= \frac{G^{-1} \sum_{g=1}^G \rho(\theta^{(g)}_i, \theta^*_i|Y, J_s, \Psi^*_i, \Psi^{i+1,(g)})q(\theta^{(g)}_i, \theta^*_i|Y, J_s, \Psi^*_i, \Psi^{i+1,(g)})}{M^{-1} \sum_{m=1}^M \rho(\theta^*_m, \theta^{(m)}_i|Y, J_s, \Psi^*_i, \Psi^{i+1,(m)})},$$ \hspace{1cm} (2.46)

where $\{\theta^{(g)}_i, \Psi^{i+1,(g)}\}_{g=1}^G$ are obtained from $\pi(\theta_i, \Psi^{i+1}|Y, J_s, \Psi^*_i)$. $\{\Psi^{i+1,(m)}\}$ are obtained from $\pi(\Psi^{i+1}|Y, J_s, \Psi^*_i, \theta^*_i)$, and then for each $\{\Psi^{i+1,(m)}\}$, sample $\theta^*_i$ from $q(\theta^*_i, \theta_i|Y, J_s, \Psi^{i+1,(m)})$. 

21
2.5 Posterior Contraction Rate

For classification problem, the most important object to study is the misclassification rate. By examining convergence to the true distribution, it follows that the Bayes procedure has misclassification rate close to that of the oracle procedure which uses the true values of the regression functions and other parameters (if any), e.g., cut-points in the ordered multinominal probit model. In the Bayesian nonparametric setting, Hellinger convergence is established by applying the general theory (Ghosal and van der Vaart, 2017). Thus, in this section, we only consider the contraction rate of the posterior distribution with respect to a metric on the probability of categories, which is equivalent with Hellinger distance on the joint distribution. The posterior contraction rates of the three multinominal models with finite random series prior can be obtained using calculation similar to those in Shen and Ghosal (2015) on posterior contraction rates for finite random series.

We use \( \lesssim \) to denote an inequality up to a constant multiple, \( f \asymp g \) for \( f \lesssim g \lesssim f \). For a vector \( \theta \in \mathbb{R}^d, ||\theta||_p = \left\{ \sum_{i=1}^d |\theta_i|^p \right\}^{1/p} \), where \( 1 \leq p < \infty \), and \( ||\theta||_\infty = \max_{1 \leq i \leq d} |\theta_i| \). Similarly, for a function \( f \) with respect to a measure \( G \), we define \( ||f||_{p,G} = \left\{ \int |f(x)|^p dG \right\}^{1/p} \), where \( 1 \leq p < \infty \), and \( ||f||_{\infty,G} = \sup_x |f(x)| \). Let \( \mathcal{N}(\epsilon, T, d) \) denote the \( \epsilon \)-covering number of a set \( T \) for a metric \( d \). Let \( h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu \) be the squared Hellinger distance, \( K(p, q) = \int p \log(p/q) d\mu, V(p, q) = \int p \log^2(p/q) d\mu \) be the Kullback-Leibler (KL) divergences.

Suppose that \((X_i, Y_i), i = 1, \ldots, n\), are the independent observations. Let \( p \) denote the joint probability of \((X, Y)\), where \( Y \) takes values \( 1, \ldots, K \), and \( p_0 \) denote the true joint probability. Let \((X^{(n)}, Y^{(n)})\) be the vector of \( n \) observations following the probability \( p^{(n)} \). Let \( \pi_k(X) = P(Y = k|X) \) be the probability of the \( k \)th category conditioned on \( X \), and
$\pi_{0k}$ be the true probability of the $k$th category conditioned on $X$. Define the probability vector $\pi = (\pi_1, \ldots, \pi_K)^T$, where $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$, and $\pi_0 = (\pi_{01}, \ldots, \pi_{0K})^T$, where $\pi_{0K} = 1 - \sum_{k=1}^{K-1} \pi_{0k}$. Assume that the distribution of $X$ is $G$, and $\nu$ denotes the counting measure on $\{1, \ldots, K\}$. For these multinomial models, the KL divergences $K(p_0, p)$, and $V(p_0, p)$ can be reduced to

$$K(p_0, p) = \int \int p_0(x, y) \log \frac{p_0(x, y)}{p(x, y)} \, d\nu(y) \, dx$$

$$= \int \int \pi_0(y|x) \log \frac{\pi_0(y|x)}{\pi(y|x)} \, d\nu(y) \, dG(x)$$

$$= \mathbb{E}_X \left\{ \sum_{k=1}^{K} \pi_{0k}(X) \log \frac{\pi_{0k}(X)}{\pi_k(X)} \right\}$$

$$= K(\pi_0, \pi), \text{ say} \quad (2.47)$$

$$V(p_0, p) = \int \int p_0(x, y) \log^2 \frac{p_0(x, y)}{p(x, y)} \, d\nu(y) \, dx$$

$$= \int \int \pi_0(y|x) \log^2 \frac{\pi_0(y|x)}{\pi(y|x)} \, d\nu(y) \, dG(x)$$

$$= \mathbb{E}_X \left\{ \sum_{k=1}^{K} \pi_{0k}(X) \log^2 \frac{\pi_{0k}(X)}{\pi_k(X)} \right\}$$

$$= V(\pi_0, \pi), \text{ say.} \quad (2.48)$$

Similarly, the squared Hellinger distance $h^2(p_1, p_2)$ can be reduced to

$$h^2(p_1, p_2) = \int \int (\sqrt{\pi_1(y|x)} - \sqrt{\pi_2(y|x)}) \, d\nu(y) \, dG(x)$$

$$= \mathbb{E}_X \left\{ \sum_{k=1}^{K} \left( \sqrt{\pi_{1k}(X)} - \sqrt{\pi_{2k}(X)} \right)^2 \right\}$$

$$= h^2(\pi_1, \pi_2), \text{ say.} \quad (2.49)$$
We define a metric by
\[ d(\pi, \pi_0) = \sqrt{\sum_{k=1}^{K} E_X |\pi_k(X) - \pi_{0k}(X)|^2}. \] (2.50)

Then we have the following slightly simplification of a general posterior contraction theorem suitable in our context.

**Theorem 1.** Assume that \( \pi_0 \) is bounded away from zero. Let \( \epsilon_n \geq \bar{\epsilon}_n \) be two sequences of positive numbers satisfying \( \epsilon_n \to 0 \) and \( n\bar{\epsilon}_n^2 \to \infty \). Let \( \mathcal{X}_0 \) be such that \( P(X \in \mathcal{X}_0) = 1 \) and \( \pi_k(x), k = 1, \ldots, K \) for \( x \in \mathcal{X}_0 \) is bounded away from 0. Let \( \mathcal{W}_n \) be a subset of the parameter space such that the following conditions hold for some positive constants \( a_2 \) and \( a_1 > a_2 + 2 \):

\[
\log \mathcal{N}(\epsilon_n, \mathcal{W}_n, h) \lesssim n\epsilon_n^2, \tag{2.51}
\]
\[
\Pi(\pi \not\in \mathcal{W}_n) \leq \exp\{-a_1 n\bar{\epsilon}_n^2\}, \tag{2.52}
\]
\[
- \log \Pi \left( \sum_{k=1}^{K} \|\pi_k - \pi_{0k}\|_{\infty, \mathcal{X}_0} \leq \epsilon_n^2 \right) \leq a_2 n\epsilon_n^2, \tag{2.53}
\]

where \( \|\pi_k - \pi_{0k}\|_{\infty, \mathcal{X}_0} = \sup_{x \in \mathcal{X}_0} |\pi_k(x) - \pi_{0k}(x)| \). Then for every \( M_n \to \infty \), we have \( \Pi \left( d(\pi, \pi_0) \geq M_n \epsilon_n |X^{(n)}, Y^{(n)}| \right) \to 0 \) in probability.

The proof follows from Theorem 4 of Ghosal and van der Vaart (2007a), by observing that

\[ h^2(\pi, \pi_0) = E_X \sum_{k=1}^{K} \frac{|\pi_k(X) - \pi_{0k}(X)|^2}{\left| \sqrt{\pi_k(X)} + \sqrt{\pi_{0k}(X)} \right|^2} \geq E_X \sum_{k=1}^{K} |\pi_k(X) - \pi_{0k}(X)|^2, \] (2.54)
and by expanding in Taylor’s expansion

$$\max\{K(\pi_0, \pi), V(\pi_0, \pi)\} \lesssim \sum_{k=1}^{K} \|\pi_k - \pi_{0k}\|_{\infty, X_0}^2, \quad (2.55)$$

Let $\Pi$ be a generic notation for priors on the number $J$ of basis functions. As in Shen and Ghosal (2015), the priors on $J$ and the coefficients of the basis functions $\theta = (\theta_1, \ldots, \theta_J)^T$ need to satisfy the conditions (A1) and (A2). For the ordered multinomial probit model, we add condition (A3).

(A1) For some $c_1, c_2 > 0$, $0 \leq t_2 \leq t_1 \leq 1$, $\exp\{-c_1 J \log t_1^J\} \leq \Pi(J = j) \leq \exp\{-c_2 J \log t_2^J\}$.

(A2) Given $J$, $\Pi(\|\theta - \theta_0\|_2 \leq \epsilon) \geq \exp\{-c_3 J \log(1/\epsilon)\}$ for every $\|\theta_0\|_\infty \leq H$, where $c_3$ is some positive constant, $H$ is chosen sufficiently large, and $\epsilon > 0$ is sufficiently small. Also, assume that $\Pi(\theta \notin [-M, M]^J) \leq J \exp\{-CM t_3\}$ for some constant $C$, $t_3 > 0$,

(A3) Given $K$ categories, $\Pi(\|\gamma - \gamma_0\|_2 \leq \epsilon) \geq \exp\{-c_4 K \log(1/\epsilon)\}$, where $c_4$ is some positive constant.

Geometric distribution with $t_1 = t_2 = 0$, and Poisson distribution with $t_1 = t_2 = 1$ on $J$ satisfy (A1). The multivariate normal distribution on $\theta$ and $\gamma$ satisfy (A2) and (A3) respectively.

To obtain the posterior contraction rate, we need to verify the conditions (2.51)–(2.53), and we also need additional assumptions on the basis. We use $\theta^T \psi(t)$ to approximate $\beta(t)$, where $\theta = (\theta_1, \ldots, \theta_J)^T$, and $\psi(t) = (\psi_1(t), \ldots, \psi_J(t))^T$. Let $\beta_0(t)$ be the true value, and $r = 2$ or $\infty$. Assume that there exist a $\theta_0 \in \mathbb{R}^J$, $\|\theta_0\|_\infty \leq H$ and $K_0 \geq 0$ such
that

\[ \|\beta_0(\cdot) - \theta_0^T \psi(\cdot)\|_r \lesssim J^{-\alpha}, \quad (2.56) \]
\[ \|\theta_1^T \psi(\cdot) - \theta_2^T \psi(\cdot)\|_r \lesssim J^{K_0}\|\theta_1 - \theta_2\|_2, \quad \theta_1, \theta_2 \in \mathbb{R}^J. \quad (2.57) \]

Remark 2 of Shen and Ghosal (2015) gave examples of bases satisfying relations (2.56) and (2.57). For B-splines, the relations hold when \( K_0 = 1/2 \) with \( r = 2 \), and \( K_0 = 1 \) with \( r = \infty \).

**Remark 1.** Parameter estimation plays a secondary role here. The problem of estimating model parameters is interesting in its own right but is not necessary for good classifications. Cai and Hall (2006) and Yuan and Cai (2010) showed that the parameter function estimation and the prediction from an estimator of the parameter function have different characteristics.

### 2.5.1 Ordered Multinomial Probit Model

Let \( \gamma = (\gamma_1, \ldots, \gamma_K)^T \) be the vector of the threshold points, and \( \gamma_0 = (\gamma_{01}, \ldots, \gamma_{0K})^T \) be the vector of the true values of the threshold points. Let \( \beta(t) \) be the parameter function on \([0, 1] \), and \( \beta_0(t) \) be the true parameter function on \([0, 1] \). Let

\[ \pi_k(X) = \Phi\left( \gamma_k - \int \beta(t)X(t)dt \right) - \Phi\left( \gamma_{k-1} - \int \beta(t)X(t)dt \right), \quad (2.58) \]

and

\[ \pi_{0k}(X) = \Phi\left( \gamma_{0k} - \int \beta_0(t)X(t)dt \right) - \Phi\left( \gamma_{0k-1} - \int \beta_0(t)X(t)dt \right). \quad (2.59) \]
Theorem 2. Assume that $\|X\|_1 = \int |X(t)| dt$ is a bounded random variable, the priors satisfy the conditions (A1), (A2) and (A3), and that the basis $\psi(t)$ satisfies (2.56) and (2.57) with $r = \infty$. Then the posterior contraction rate of the ordered multinomial probit model is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+1-t_2}/2$ relative to $d(\pi, \pi_0)$. More explicitly, for every $M_n \to \infty$, $\Pi(\beta \colon \rho(\beta, \beta_0) \geq M_n \epsilon_n |X(n), Y(n)| \to 0$ in probability, where $\rho(\beta, \beta_0) = E_X \int (\beta(t) - \beta_0(t)) X(t) dt |$, and $\Pi(\gamma : \max_j |\gamma_j - \gamma_0_j| \geq M_n \epsilon_n |X(n), Y(n)| \to 0$ in probability.

Proof. For any $x \in \mathcal{X}_0 = \{ \int |X(t)| dt \leq M \}$, say, by the Lipschitz continuity of $\Phi$, we have

$$|\pi_k(x) - \pi_{0k}(x)| \lesssim \max_k |\gamma_k - \gamma_{0k}| + \left| \int (\beta(t) - \beta_0(t)) x(t) dt \right|$$

$$\lesssim \| \gamma - \gamma_0 \|_\infty + \| \beta(\cdot) - \beta_0(\cdot) \|_\infty \int |x(t)| dt \quad (2.60)$$

$$\lesssim \| \gamma - \gamma_0 \|_\infty + \| \beta(\cdot) - \beta_0(\cdot) \|_\infty.$$

Observe that with the finite random series prior, the $L_\infty$-distance between $\beta(\cdot)$ and $\beta_0(\cdot)$ is bounded by

$$\| \beta(\cdot) - \beta_0(\cdot) \|_\infty = \| \theta^T \psi(\cdot) - \theta_0^T \psi(\cdot) + \theta_0^T \psi(\cdot) - \beta_0(\cdot) \|_\infty$$

$$\leq \| \theta^T \psi(\cdot) - \theta_0^T \psi(\cdot) \|_\infty + \| \theta_0^T \psi(\cdot) - \beta_0(\cdot) \|_\infty. \quad (2.61)$$
Then we have
\[
\Pi \left( \sum_{k=1}^{K} \| \pi_k - \pi_{0k} \|_{2, \infty, \chi_0}^2 \leq \bar{\epsilon}_n^2 \right)
\]
\[
\geq \Pi(\| \gamma - \gamma_0 \| \leq \bar{\epsilon}_n / \sqrt{2}) \Pi(\| \beta(\cdot) - \beta_0(\cdot) \|_{\infty} \leq \bar{\epsilon}_n / \sqrt{2})
\]
\[
\geq \Pi(\| \gamma - \gamma_0 \| \leq \bar{\epsilon}_n / \sqrt{2}) \Pi(\| \theta - \theta_0 \| \leq \bar{\epsilon}_n / (2 \sqrt{2} \bar{J}_n K_0))
\]
\[
\geq \exp\left\{ -K \log \left( \sqrt{2} / \bar{\epsilon}_n \right) \right\} \exp\left\{ -\bar{J}_n \log \left( 2 \sqrt{2} \bar{J}_n K_0 / \bar{\epsilon}_n \right) \right\}.
\] (2.62)

To satisfy the relation (2.53), we need \( \bar{J}_n^{-\alpha} \lesssim \bar{\epsilon}_n \) and
\[
K \log \left( \sqrt{2} / \bar{\epsilon}_n \right) + \bar{J}_n \log \left( 2 \sqrt{2} \bar{J}_n K_0 / \bar{\epsilon}_n \right) \lesssim n \bar{\epsilon}_n^2.
\] (2.63)

Thus (2.63) leads to the conditions that \( J_n \log n \lesssim n \bar{\epsilon}_n^2 \). Then we obtain the preliminary contraction rate \( \bar{\epsilon}_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)} \), for \( \bar{J}_n \asymp (n / \log n)^{1/(2\alpha+1)} \).

Using (2.60), we obtain
\[
\log \mathcal{N}(\epsilon_n, \mathcal{W}_n, h) \lesssim \log \mathcal{N}(\epsilon_n, \mathcal{W}_n, \| \cdot \|_{\infty}) \lesssim n \bar{\epsilon}_n^2.
\] (2.64)

According to Theorem 2 of Shen and Ghosal (2015), to satisfy (2.64), we need
\[
J_n \{(K_0 + 1) \log J_n + \log M_n + C_0 \log n \} \leq n \bar{\epsilon}_n^2,
\] (2.65)

for some positive constant \( C_0 \). To satisfy (2.52), we need
\[
b n \bar{\epsilon}_n^2 \leq J_n \log^{t_2} J_n, \quad \log J_n + n \bar{\epsilon}_n^2 \leq M_n^{t_3},
\] (2.66)

for some \( b > 0 \). For \( M_n = n^{1/t_3} \), (2.66) implies that \( J_n \log^{t_2} n \gtrsim n \bar{\epsilon}_n^2 \). Thus \( J_n \asymp
\[ n^{1/(2\alpha+1)(\log n)^{2\alpha/(2\alpha+1) - t_2}}. \] Relation (5.65) implies that \( J_n \log n \lesssim n\epsilon_n^2 \). As a result, the posterior contraction rate is \( \epsilon_n \asymp n^{-\alpha/(2\alpha+1)(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}} \) relative to \( d(\pi, \pi_0) \).

Further, by Jensen’s inequality, we have

\[
E_X|\pi_k(X) - \pi_{0k}(X)|^2 \geq \left\{ E_X|\pi_k(X) - \pi_{0k}(X)| \right\}^2. \tag{2.67}
\]

If \( k = 1 \), by the mean value theorem and the uniform positivity of \( \Phi \) on compact interval, then

\[
E_X|\pi_1(X) - \pi_{01}(k)| = E_X \left| \Phi(- \int \beta(t)X(t)dt) - \Phi(- \int \beta_0(t)X(t)dt) \right| \\
\gtrsim E_X \left| \int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt \right|. \tag{2.68}
\]

Hence if \( E_X|\pi_1(X) - \pi_{01}(X)|^2 \) is small, then \( E_X \left| \int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt \right| \) is also small. If \( k = 2 \), we have

\[
E_X|\pi_2(X) - \pi_{02}(k)| = E_X \left| \Phi(\gamma_2 - \int \beta(t)X(t)dt - \Phi(\gamma_02 - \int \beta_0(t)X(t)dt) \\
- \Phi(- \int \beta(t)X(t)dt) + \Phi(- \int \beta_0(t)X(t)dt) \right| \\
\gtrsim E_X \left| \Phi(\gamma_2 - \int \beta(t)X(t)dt - \Phi(\gamma_02 - \int \beta_0(t)X(t)dt) \\
- E_X \left| \Phi(- \int \beta(t)X(t)dt) - \Phi(- \int \beta_0(t)X(t)dt) \right| \right|. \tag{2.69}
\]

From (2.68), we know that \( E_X \left| \Phi(- \int \beta(t)X(t)dt) - \Phi(- \int \beta_0(t)X(t)dt) \right| \) is small, and if \( E_X|\pi_2(X) - \pi_{02}(X)|^2 \) is small, then

\[
E_X \left| \Phi(\gamma_2 - \int \beta(t)X(t)dt - \Phi(\gamma_02 - \int \beta_0(t)X(t)dt) \right| \tag{2.70}
\]

29
is also small. By the mean value theorem and the uniform positivity of \( \Phi \) on compact interval, we have

\[
\mathbb{E}_X \left| \Phi(\gamma_2 - \int \beta(t)X(t)dt) - \Phi(\gamma_{02} - \int \beta_0(t)X(t)dt) \right| \\
\gtrsim \mathbb{E}_X \left| \gamma_2 - \gamma_{02} - \int \beta(t)X(t)dt + \int \beta_0(t)X(t)dt \right| \\
\gtrsim |\gamma_2 - \gamma_{02}| - \mathbb{E}_X \left| \int \beta(t)X(t)dt - \int \beta_0(t)X(t)dt \right|. 
\]

(2.71)

Hence \( |\gamma_2 - \gamma_{02}| \) is small. Similarly, we can prove that \( |\gamma_k - \gamma_0| \) is small for any \( k \).

2.5.2 Unordered Multinomial Probit Model

Note that by (2.42)

\[
\pi_k(X) = \frac{1}{\sqrt{\pi}} \int_0^\infty \left\{ \prod_{l=1}^{K-1} \Phi(-z\sqrt{2} - \int \beta_l(t)X(t)dt) \\
+ \prod_{l=1}^{K-1} \Phi(z\sqrt{2} - \int \beta_l(t)X(t)dt) \right\} e^{-z^2} dz
\]

(2.72)

**Theorem 3.** Assume that \( \|X\|_1 = \int |X(t)| \, dt \) is a bounded random variable, the priors satisfy the conditions (A1) and (A2), and that the basis \( \psi(t) \) satisfies (2.56) and (2.57) with \( r = \infty \). Then the posterior contraction rate of the unordered multinomial probit model is \( \epsilon_n \lesssim n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2} \) relative to \( d(\pi, \pi_0) \).

**Proof.** For some \( M > 0 \), \( P(\mathcal{X}_0) = 1 \) for \( \mathcal{X}_0 = \{ \int |X(t)| \, dt \leq M \} \). For any \( x \in \mathcal{X}_0 \), by the
Lipschitz continuity of the function $\Phi$, we have

$$|\pi_k(x) - \pi_0k(x)| \lesssim \left| \int \beta_k(t)x(t)dt - \int \beta_0k(t)x(t)dt \right|$$

$$\lesssim \int |\beta_k(t) - \beta_0k(t)||x(t)|| dt$$

$$\lesssim \|\beta_k(\cdot) - \beta_0k(\cdot)\|_\infty. \quad (2.73)$$

The $L_\infty$-distance between $\beta_k(\cdot)$ and $\beta_0k(\cdot)$ is bounded by

$$\|\beta_k(\cdot) - \beta_0k(\cdot)\|_\infty = \|\theta_k^T \psi(\cdot) - \theta_0k^T \psi(\cdot) + \theta_0k^T \psi(\cdot) - \beta_0k(\cdot)\|_\infty$$

$$\leq \|\theta_k^T \psi(\cdot) - \theta_0k^T \psi(\cdot)\|_\infty + \|\theta_0k^T \psi(\cdot) - \beta_0k(\cdot)\|_\infty. \quad (2.74)$$

Then we have

$$\Pi\left(\sum_{k=1}^K \|\pi_k - \pi_0k\|_{\infty,x_0}^2 \leq \bar{\epsilon}_n^2\right) \geq \Pi\left(\sum_{k=1}^K \|\beta_k(\cdot) - \beta_0k(\cdot)\|_\infty^2 \leq \bar{\epsilon}_n^2\right)$$

$$\geq \Pi(\|\theta_k - \theta_0k\| \leq \bar{\epsilon}_n/(2\sqrt{\bar{J}_n J_n K_0}))$$

$$\gtrsim \exp\{-\bar{J}_n \log\left(2\sqrt{\bar{K} \bar{J}_n K_0} / \bar{\epsilon}_n\right)\}. \quad (2.75)$$

To satisfy the relation (2.53), we need $\bar{J}_n^{-\alpha} \lesssim \bar{\epsilon}_n$ and

$$\bar{J}_n \log\left(2\sqrt{\bar{K} \bar{J}_n K_0} / \bar{\epsilon}_n\right) \lesssim n\bar{\epsilon}_n^2. \quad (2.76)$$

Thus (2.76) leads to the conditions that $\bar{J}_n \log n \lesssim n\bar{\epsilon}_n^2$. Then we obtain the preliminary contraction rate $\bar{\epsilon}_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)}$, for $\bar{J}_n \asymp (n/\log n)^{1/(2\alpha+1)}$.

Following the same arguments as (2.64)–(2.66), the posterior contraction rate is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_2)/2}$ relative to $d(\pi, \pi_0)$. \qed
2.5.3 Multinomial Logistic Model

Let $\beta_k(t), \ k = 1, \ldots, K - 1,$ be the coefficient functions on $[0, 1],$ and $\beta_{0k}(t), \ k = 1, \ldots, K - 1,$ be the true coefficient functions on $[0, 1].$

**Theorem 4.** Assume that $\|X\|_1 = \int |X(t)| \, dt$ is a bounded random variable, the priors satisfy the conditions (A1) and (A2), and that the basis $\psi(t)$ satisfies (2.56) and (2.57) with $r = \infty.$ Then the posterior contraction rate of the multinomial logistic model is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)} + (1-t^2)/2$ relative to $d(\pi, \pi_0).$

**Proof.** The proof is similar to that of Theorem 3. \qed

2.6 Discriminant Analysis

As a comparison to those multinomial models, we use Bayesian discriminant analysis to classify the functional data. Instead of modeling the class probability directly, the discriminant analysis uses Bayes’s rule to compute the marginal likelihood of $Y_i$ (Gelman et al., 2013). The classical discriminant analysis applies only to multivariate data. For functional data, we can use certain orthogonal linear functions to determine the classification probabilities:

$$ (f_{i1}, \ldots, f_{im})^T = \left( \int \beta_1(t)X_i(t) \, dt, \ldots, \int \beta_m(t)X_i(t) \, dt \right)^T \quad (2.77) $$

Ideally these $\beta_1(t), \ldots, \beta_m(t)$ are unknown, but putting a prior on these functions with identifiability restrictions is complicated. We instead consider $\beta_1(t), \ldots, \beta_m(t)$ to be known as the first $m$ principal components (Ramsay and Silverman, 2005), but let
the means and the covariance matrices be unknown. Then discriminant analysis can be applied to the \( m \) principal components.

### 2.6.1 Linear Discriminant Analysis

Linear discriminant analysis assumes that for each of the \( K \) category, the set of linear function \((f_1, \ldots, f_m)\) follows a normal distribution with the same covariance matrix: \((f_{il1}, \ldots, f_{ilm})^T \sim N(\mu_l, \Sigma)\), where \( \mu_l \) is the population mean of category \( l \), \( l = 1, \ldots, K \), \( i = 1, \ldots, n_l \), and \( n_l \) is the number of data in category \( l \). Then the probability of choosing category \( k \) is given by

\[
P(Y_i = k | X_i) = \frac{p_k \cdot \phi(f_{ik1}, \ldots, f_{ikm}; \mu_k, \Sigma)}{\sum_{l=1}^{K} p_l \cdot \phi(f_{il1}, \ldots, f_{ilm}; \mu_l, \Sigma)},
\]

where \( \phi(f_1, \ldots, f_m; \mu, \Sigma) \) is the multivariate normal density function with mean \( \mu \) and covariance \( \Sigma \), and \( p_l, l = 1, \ldots, K \), are the probability of choosing category \( l \).

The variables \( f_{il1}, \ldots, f_{ilm} \) are the \( m \) principal components of \( X_i(t) \) in category \( l \), where \( l = 1, \ldots, K \). Define \( f_l = (f_{il1}, \ldots, f_{ilm})^T \), where \( i = 1, \ldots, n_l \), and \( \sum_{l=1}^{K} n_l = n \). To estimate the mean \( \mu_l \) for each category \( l \), and the common covariance \( \Sigma \) among all categories, we use the conjugate normal-inverse-Wishart prior with hyperparameters (Gelman et al., 2013) for \((\mu_l, \Sigma)\)

\[
\Sigma \sim IW_{\nu_0}(\Lambda_0^{-1}), \quad \mu_l | \Sigma \sim N(\mu_{l0}, \Sigma / \kappa_0).
\]

Then the posterior distribution of \((\mu_l, \Sigma)\) can be obtained in the following order

\[
\Sigma | Y \sim IW_{\nu_n}(\Lambda_n^{-1}), \quad \mu_l | \Sigma, Y \sim N(\mu_{ln}, \Sigma / \kappa_n),
\]
where \( \nu_n = \nu_0 + n \), \( \bar{f}_l = \sum_{i=1}^{n_l} f_{il}/n_l \), \( S = \sum_{l=1}^{K} \sum_{i=1}^{n_l} (f_{il} - \bar{f}_l)(f_{il} - \bar{f}_l)^T \),

\[
\Lambda_n = \Lambda_0 + S + \sum_{l=1}^{K} \frac{\kappa_0 n_l}{\kappa_0 + n_l} (\bar{f}_l - \mu_{l0})(\bar{f}_l - \mu_{l0})^T,
\]

and

\[
k_n = \kappa_0 + n, \quad \mu_{ln} = \frac{\kappa_0 \mu_{l0} + n_l \bar{f}_l}{\kappa_0 + n_l}, \quad l = 1, \ldots, K.
\]

### 2.6.2 Quadratic Discriminant Analysis

Quadratic discriminant analysis is defined in a similar way, except that it has a different covariance matrix for each category. The probability of choosing category \( k \) is given by

\[
P(Y_i = k|X_i) = \frac{p_k \cdot \phi(f_{ik1}, \ldots, f_{ikm}; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} p_l \cdot \phi(f_{il1}, \ldots, f_{ilm}; \mu_l, \Sigma_l)}.
\]

To estimate the mean \( \mu_l \) and the covariance \( \Sigma_l \) for each category \( l \), where \( l = 1, \ldots, K \), we use the conjugate normal-inverse-Wishart prior with hyperparameters for \((\mu_l, \Sigma_l)\)

\[
\Sigma_l \sim IW_{\nu_l0}(\Lambda_{l0}^{-1}), \quad \mu_l|\Sigma_l \sim N(\mu_{l0}, \Sigma_l/\kappa_{l0}),
\]

for \( l = 1, \ldots, K \). Then the posterior distribution of \((\mu_l, \Sigma_l)\) can be obtained in the following order

\[
\Sigma_l|Y \sim IW_{\nu_{ln}}(\Lambda_{ln}^{-1}), \quad \mu_l|\Sigma_l, Y \sim N(\mu_{ln}, \Sigma_l/\kappa_{ln}),
\]

34
where $\nu_t = \nu_0 + n_t$, $\tilde{f}_t = \sum_{i=1}^{n_t} f_{it}/n_t$, $S_t = \sum_{i=1}^{n_t} (f_{it} - \tilde{f}_t)(f_{it} - \tilde{f}_t)^T$, 

$$
\Lambda_t = \Lambda_0 + S_t + \frac{\kappa_{i0} n_t}{\kappa_{i0} + n_t} (\tilde{f}_t - \mu_{i0})(\tilde{f}_t - \mu_{i0})^T,
$$

(2.86)

and

$$
\kappa_{i0} = \kappa_{i0} + n_t, \quad \mu_{i0} = \frac{\kappa_{i0} \mu_{i0} + n_t \tilde{f}_t}{\kappa_{i0} + n_t}, \quad l = 1, \ldots, K.
$$

(2.87)

2.7 Simulation

2.7.1 Data Generation

The simulated data are generated following different data generating process. All of the simulated data have three categories. In all cases considered below, we generate the functional data from a Gaussian process at discrete time points 0, 0.01, \ldots, 0.99, 1, with the mean function $\sin t$ and variation kernel $100 \exp\{ -100(t_i - t_j)^2 \}$, where $t_i$ and $t_j$ were the discrete time point 0, 0.01, \ldots, 0.99, 1.

For the ordered multinomial probit data, the coefficient function $\beta(t)$ is plotted in Figure 1 (a), and the four threshold points are chosen to be $-\infty$, 0, 8, $\infty$. The four cut-off points construct three intervals. If the inner product of a functional data and the coefficient function plus a standard normal variable falls in the $k$th interval $(\gamma_{k-1}, \gamma_k)$, then the functional data attributes to the category $k$.

For unordered multinomial probit data, the coefficient functions $\beta_1(t), \beta_2(t), \beta_3(t)$ are plotted in Figure 1 (b)–(d). The inner product of a functional data and the three coefficient functions are added with standard normal variables, respectively. We sample from these three normal variables, and obtain the corresponding probabilities. Then the
(a) $\beta(t)$ for the ordered multinomial probit model  
(b) $\beta_1(t)$ for the unordered multinomial probit model  
(c) $\beta_2(t)$ for the unordered multinomial probit model  
(d) $\beta_3(t)$ for the unordered multinomial probit model  
(e) $\beta_1(t)$ for the multinomial logistic model  
(f) $\beta_2(t)$ for the multinomial logistic model

Figure 2.1: Coefficient functions for the multinomial models
functional data belongs to the category with the largest sampled value.

For the multinomial logistic data, the coefficient functions $\beta_1(t), \beta_2(t)$ are plotted in Figure 1 (e)–(f), and the third coefficient function $\beta_3(t)$ can be assumed to be zero everywhere. We compute the probability of a functional data falling into each category. Then the data attributes to the category with the largest probability.

For data satisfying the assumption of the linear discriminant analysis, we generate them from three Gaussian processes with different mean functions $\sin t + 2\cos t$, $\sin t$, and $\sin t - 3\cos t$, but the same variation kernel $\exp\{-30(t_i - t_j)^2\}$.

For data satisfying the assumption of the quadratic discriminant analysis, we generate them from three Gaussian processes with different mean functions and different variation kernels. The mean functions are $\sin t + 2\cos t$, $\sin t$, and $\sin t - 3\cos t$, and the variation kernels are $\exp\{-2\sin^2(\pi(t_i - t_j))\}$, $\exp\{-30(t_i - t_j)^2\}$, and $\exp\{-|t_i - t_j|\}$, respectively.

In this simulation study, we generate total 900 (300 for each category) functional data for each type of dataset. We constructe the training data with 720 (240 for each category) of them and the testing data with the remaining 180 (60 for each category) of them.

2.7.2 Basis Functions

For models using the finite random series prior, we consider the B-spline basis. The B-spline basis functions on interval $[0, 1]$ can be created using the R package fda. In this simulation study, we put a geometric prior with $p = 0.5$ on $J$. We only consider the possible number of B-spline basis functions to be $J = 5, \ldots, 15$, since the probability outside this range is too small. Those B-spline basis functions are generated at the same discrete time points as the functional data, that is $0, 0.01, \ldots, 0.99, 1$. 
Table 2.1: Averaged misclassification rates for simulated data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OMP Model</th>
<th>UMP Model</th>
<th>MLO Model</th>
<th>LDA</th>
<th>QDA</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMP</td>
<td>7.69%</td>
<td>30.56%</td>
<td>28.33%</td>
<td>38.89%</td>
<td>48.89%</td>
<td>15.00%</td>
</tr>
<tr>
<td>UMP</td>
<td>38.96%</td>
<td>7.22%</td>
<td>7.78%</td>
<td>21.11%</td>
<td>21.11%</td>
<td>10.56%</td>
</tr>
<tr>
<td>MLO</td>
<td>49.44%</td>
<td>4.75%</td>
<td>3.89%</td>
<td>32.22%</td>
<td>36.11%</td>
<td>7.78%</td>
</tr>
<tr>
<td>LDA</td>
<td>26.32%</td>
<td>25.69%</td>
<td>26.11%</td>
<td>5.00%</td>
<td>5.00%</td>
<td>7.78%</td>
</tr>
<tr>
<td>QDA</td>
<td>24.28%</td>
<td>21.95%</td>
<td>21.67%</td>
<td>10.56%</td>
<td>9.44%</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

2.7.3 Results

Under the chosen models, we apply Baysian estimation methods described in Section 3 on the training data. In this study, 5000 MCMC iterations are obtained, and the first 1000 of them are discarded as burn-in. We use the last 4000 MCMC output of the parameter $B$ to classify the 180 transformed testing data, where $B = (\theta, \gamma_2, \gamma_3)$ for the ordered multinomial probit model, $B = \Theta$ for the unordered multinomial probit model, $B = (\theta_1, \theta_2)$ for the logistic model, $B = (\mu_1, \mu_2, \mu_3, \Sigma)$ for the linear discriminant analysis model, and $B = (\mu_1, \mu_1, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3)$ for the quadratic discriminant analysis model. A transformed testing data $z_i$ or $f_i$ is in category $k$ if $\sum_{g=1}^{4000} \mathbb{1}(Y_i = k|z_i \text{ or } f_i, B^{(g)}) > \sum_{g=1}^{4000} \mathbb{1}(Y_i = l|z_i \text{ or } f_i, B^{(g)})$, where $l \neq k$. Then we use the techniques described in Section 4 to average the results from the multinomial models. As a comparison with the Bayesian method, the linear support vector machine (SVM) is also applied to the principal components of these training data, and made predictions on the testing data. To apply SVM, we use the R package e1071. Table 2.1 shows the averaged misclassification rates for each data type under different models.
2.8  Application on Phoneme Data

Ferraty and Vieu (2006) extracted a speech recognition data originally introduced by a collaboration between Andreas Buja, Werner Stuetzle and Martin Maechler, and illustrated in the paper by Hastie, Buja, and Tibshirani (1995). The dataset has five speech frames corresponding to the following five phonemes:

1. “sh” as in “she”;
2. “iy” as in “she”;
3. “dcl” as in “dark”;
4. “aa” as in “dark”;
5. “ao” as in “water”.

The original data is from a continuous speech of 50 male speakers, and 4509 speech frames of 32 msec duration were selected, approximately 2 examples of each phoneme from each speaker. For each speech frame, Ferraty and Vieu (2006) extracted 400 samples at a 16kHz sampling rate, and each frame represents one of the above five phonemes. They retained the first 150 frequencies from each subject. Thus, the data consist of 2000 log-periodograms of length 150 with known phoneme classes. This dataset can be found in the R package fds, and can also be found at https://www.math.univ-toulouse.fr/staph/npfda/. The data are the discretized log-periodograms. As an illustration, we smooth them by basis expansion. we use 10 B-splines bases on the interval [0, 150]. The following plots in Figure 2.2 show all the curves in each phoneme class. These curves are the relationship between amplitude and frequency.
Figure 2.2: Curves for each of the five phoneme classes
2.8.1 Three Classes

From Figure 2.2, we can see that class “aa” and “ao” have a very similar pattern which can cause difficulty in classification. To simplify the problem, in the beginning, we only classify the first three classes (“sh”, “iy”, and “dcl”). For computational efficiency, we only use 900 of them from three categories. We split the data into training and testing set by randomly sampling from each class, and keeping the same percentage of samples of each class as the complete set. The size of the testing data is 20% of the total data size. That is we have 240 data for each class in the training set, and 60 data for each class in the testing set. We put a geometric prior with $p = 0.5$ on $J$, and it is enough for us to consider the number of B-spline basis functions to be $J = 5, \ldots, 15$. We obtain 5000 MCMC iterations and discard the first 1000 of them as burn-in.

According to Table 2.2, the unordered multinomial probit model is the best model for the phoneme data. The misclassification rate of the unordered multinomial model is only 0.56%. For this data, the categories are not naturally ordered, and hence ordered multinomial probit model is not natural for this problem, but we include it in the analysis for comparison. Figure 2.3 displays the cut-point $\gamma_2$ sampled by Metropolis-Hastings under different $J$, and we can tell that $\gamma_2$ converges around 500 iterations. Tables 2.3, 2.4, and 2.5 show the estimate and standard error of the posterior mean of the phoneme data under ordered multinomial probit model, multinomial logistic model, and unordered multinomial probit model, when $J = 6$, $J = 10$, and $J = 14$, respectively. We choose these $J$ values because under these values the model has the largest posterior probability $P(J|Y)$. Although ordered multinomial probit model is not intuitive in this context, its performance is not too inferior.
Figure 2.3: $\gamma_2$ sampled from Metropolis-Hastings when $J=5–7$ and 13–15

Table 2.2: Averaged misclassification rates for 3-class phoneme data

<table>
<thead>
<tr>
<th></th>
<th>OMP Model</th>
<th>UMP Model</th>
<th>MLO Model</th>
<th>LDA</th>
<th>QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged rate</td>
<td>9.84%</td>
<td>0.56%</td>
<td>5.56%</td>
<td>7.78%</td>
<td>5.00%</td>
</tr>
</tbody>
</table>

Table 2.3: Estimate and standard error of the posterior mean for the 3-class ordered multinomial model ($J = 6$)

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_2$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>3.87 (52.12, −9.60, −8.89, −0.19, −4.91, 2.85)</td>
<td>(0.34, 0.11, 0.13, 0.08, 0.10, 0.10)</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4: Estimate and standard error of the posterior mean for the 3-class multinomial logistic model \((J = 10)\)

<table>
<thead>
<tr>
<th>(\theta_2)</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((13.10, 18.25, 6.04, -15.29, 15.52, 1.30, -5.81, 4.65, -28.24, -16.91))</td>
<td>((0.94, 1.08, 0.64, 0.63, 0.75, 0.66, 0.37, 0.48, 0.70, 1.03))</td>
</tr>
<tr>
<td>(\theta_3)</td>
<td>Estimate</td>
<td>Standard error</td>
</tr>
<tr>
<td></td>
<td>((39.42, 34.30, -3.47, 5.26, -7.36, 0.38, -17.99, -4.43, -11.23, 2.44))</td>
<td>((1.21, 1.44, 0.42, 0.31, 1.08, 0.27, 0.53, 0.33, 0.93, 0.32))</td>
</tr>
</tbody>
</table>

Table 2.5: Estimate and standard error of the posterior mean for the 3-class unordered multinomial model \((J = 14)\)

\[
\begin{bmatrix}
-15.40 & 49.92 \\
35.78 & 79.79 \\
45.94 & 32.11 \\
4.97 & -0.76 \\
-23.23 & -12.58 \\
-15.09 & -14.88 \\
23.43 & -21.71 \\
-11.87 & 1.67 \\
-0.96 & -5.06 \\
-0.27 & -9.82 \\
1.58 & -7.46 \\
-12.43 & -14.68 \\
-28.97 & -7.74 \\
-28.49 & -3.38
\end{bmatrix}
\begin{bmatrix}
0.93 & 0.85 \\
0.52 & 0.81 \\
0.58 & 0.75 \\
0.66 & 0.73 \\
0.60 & 0.71 \\
0.62 & 0.67 \\
0.68 & 0.73 \\
0.74 & 0.80 \\
0.63 & 0.64 \\
0.63 & 0.69 \\
0.70 & 0.78 \\
0.57 & 0.60 \\
0.64 & 0.69 \\
0.53 & 0.57
\end{bmatrix}
\]
2.8.2 Five Classes

Our multinomial models perform well on classifying the first three classes. Now we apply our models to the entire dataset. Similarly, we split the data into training and testing set by randomly sampling from each class, and keeping the same percentage of samples of each class as the complete set. We have 320 data for each class in the training set, and 80 data for each class in the testing set. Same as Section 2.8.1, we put a geometric prior with $p = 0.5$ on $J$, and consider the number of B-spline basis functions to be $J = 5, \ldots, 15$. We obtain 5000 MCMC iterations and discard the first 1000 of them as burn-in.

The misclassification rates of all models increase as shown in Table 2.6. The misclassification rate of the unordered multinomial model is 10.75%, which is still good comparing to the result of Li and Yu (2008) who applied the functional segment discriminant analysis on phoneme data having a misclassification rate of 18.5%. Figure 2.4 shows the cut-point $\gamma_2$, $\gamma_3$, and $\gamma_4$ sampled by Metropolis-Hastings when $J = 5, 10, 15$, and they all converged in 500 iterations. Tables 2.7, 2.8, and 2.9 show the estimate and standard error of the posterior mean of the phoneme data under unordered multinomial probit model, multinomial logistic model, and ordered multinomial probit model, when $J = 14$, $J = 14$, and $J = 13$, respectively. We choose these $J$ values because under these values the model has the largest posterior probability $P(J|Y)$.

<table>
<thead>
<tr>
<th>Model</th>
<th>OMP</th>
<th>UMP</th>
<th>MLO</th>
<th>LDA</th>
<th>QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22.00%</td>
<td>10.75%</td>
<td>14.75%</td>
<td>24.00%</td>
<td>16.25%</td>
</tr>
</tbody>
</table>
Figure 2.4: $\gamma_2$, $\gamma_3$, and $\gamma_4$ sampled from Metropolis-Hastings when $J=5, 10, 15$
Table 2.7: Estimate and standard error of the posterior mean for the 5-class unordered multinomial model ($J = 14$)

<table>
<thead>
<tr>
<th>estimate</th>
<th>standard error</th>
</tr>
</thead>
</table>
| $\begin{bmatrix}
-7.80 & 54.08 & -2.84 & 3.05 \\
46.31 & 85.16 & 16.92 & 29.51 \\
38.43 & 28.16 & 26.26 & 41.23 \\
-14.55 & -16.50 & 15.61 & 35.23 \\
-1.41 & -3.17 & -1.61 & -23.32 \\
-11.41 & 2.50 & -15.08 & -14.98 \\
-4.33 & -4.78 & -6.60 & -5.70 \\
-0.50 & -13.51 & -4.62 & -6.74 \\
4.66 & -6.25 & -6.19 & -5.78 \\
-10.02 & -6.25 & -13.68 & -13.68 \\
-23.61 & -4.93 & -7.44 & -17.29
\end{bmatrix}$ | $\begin{bmatrix}
1.05 & 1.05 & 1.05 & 1.03 \\
0.87 & 0.83 & 0.81 & 0.88 \\
0.66 & 0.66 & 0.74 & 0.72 \\
0.97 & 1.00 & 1.16 & 1.19 \\
0.96 & 0.93 & 0.92 & 0.90 \\
1.00 & 1.04 & 1.08 & 1.134 \\
0.98 & 1.11 & 1.10 & 1.13 \\
0.95 & 1.14 & 1.02 & 0.98 \\
0.96 & 0.99 & 1.08 & 1.09 \\
0.85 & 0.98 & 0.74 & 0.80 \\
0.70 & 0.73 & 0.94 & 0.92 \\
0.81 & 0.86 & 1.09 & 1.06 \\
0.74 & 0.69 & 0.94 & 0.99 \\
1.02 & 1.10 & 1.14 & 1.11
\end{bmatrix}$ |
Table 2.8: Estimate and standard error of the posterior mean for the 5-class multinomial logistic model ($J = 14$)

<table>
<thead>
<tr>
<th>$\theta_2$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>1.81, 10.18, 3.24, -1.87, -9.06, 2.82, 14.20, -2.48, -4.15, -2.08, -1.85, -2.95, -4.14, -3.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard error</td>
<td>0.18, 0.57, 0.49, 0.22, 0.33, 0.20, 0.66, 0.12, 0.31, 0.13, 0.33, 0.45, 0.58, 0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_3$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>4.01, 25.83, 8.13, 0.84, -9.93, 1.94, 0.37, -4.14, -0.73, -2.80, -9.29, -2.49, 0.27, 2.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard error</td>
<td>0.64, 1.09, 0.50, 0.29, 0.20, 0.17, 0.18, 0.21, 0.32, 0.22, 0.28, 0.22, 0.17, 0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_4$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-8.04, 6.63, -1.35, 12.21, 12.29, -2.26, -1.04, -0.18, -3.28, -6.06, -10.81, -5.54, -0.87, 1.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard error</td>
<td>0.35, 0.27, 0.16, 0.44, 0.73, 0.34, 0.19, 0.30, 0.29, 0.28, 0.65, 0.24, 0.22, 0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-2.94, 0.69, 7.03, 17.97, 5.65, -5.70, -11.68, -3.17, 2.37, -2.87, -5.38, -8.40, -5.84, 7.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard error</td>
<td>0.23, 0.17, 0.59, 0.78, 0.22, 0.27, 0.40, 0.26, 0.26, 0.40, 0.38, 0.28, 0.25, 0.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.9: Estimate and standard error of the posterior mean for the 5-class ordered multinomial model ($J = 13$)

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_2$</th>
<th>$\gamma_3$</th>
<th>$\gamma_4$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>2.384</td>
<td>4.95</td>
<td>7.61</td>
<td>(18.73, 7.80, 13.29, 20.12, -2.48, -22.15, 0.69, -6.66, 4.31, -7.65, 1.72, -4.89, -5.85)</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>(0.11, 0.11, 0.10, 0.08, 0.09, 0.04, 0.05, 0.05, 0.06, 0.08, 0.09, 0.08)</td>
</tr>
</tbody>
</table>
Chapter 3

Bayesian Change Point Detection for Functional Data

3.1 Introduction

Change point detection has always been an important aspect in data analysis. In recent years, there are increasing interests in developing methods to detect change point for functional data. Berkes et al. (2009) developed a method that works with the difference of mean functions projected on the principal components of the data. Zhang et al. (2011) developed a self-normalization (SN) based test to identify potential change points in the dependence structure of functional observations. Aston and Kirch (2012) also proposed a method to detect change points for dependent functional observations, and they were particularly interested in the case where the change point is an epidemic change (a change occurs and then the observations return to baseline at a later time). Sharipov, Tewes, and Wendler (2016) developed a new test for structural changes in functional data that based on Hilbert space theory and critical values are deduced from bootstrap iterations. Aue,
Rice, and Sönmez (2018) proposed a method to uncover structural breaks in functional
data that does not rely on dimension reduction techniques.

In this chapter, we propose a Bayesian method to detect change points for functional
data. We extract the features of a sequence of functional data by the discrete wavelet
transform (DWT), and treat each sequence of feature independently. We believe there is
potentially a change in each feature at possibly different time points. The functional data
evolves through such changes throughout the sequence of observations. The change point
for this sequence of functional data is the cumulative effect of changes in all features.
Such gradual evolutionary model for changes seems to be very appropriate for functional
data, since functions have many aspects and it is hard to imagine that all those change
at once. When such a cumulative effect becomes substantial to make the following func-
tional observations significantly dissimilar with the previous ones—in that the variation
across the two groups relative to the variation within the two groups is the maximum at
that time point. We assign the features with priors which incorporate the characteristic
of the wavelet coefficients. Then we compute the posterior distribution of change point
for each sequence of feature, and define a matrix where each entry is a measure of sim-
ilarity between two functional data in this sequence. We compute the ratio of the mean
similarities between groups and within groups for all possible partitions, and the change
point is where the ratio reaches the minimum. Once we have detected one change point,
we can successively apply the procedure to subgroups divided by the change point. We
can continue finding the change points in subgroups until a stopping criterion has been
met. For example, we can stop if a certain number of change points have been detected,
or there is no significant difference in the previous and following observations. Thus, this
method can be inherently extended to multiple change points detection.
3.2 Model

We follow the formulation of Suarez and Ghosal (2016) for the structure of functional observations, who applied their model in the context of clustering. We extend their approach to change point detection for functional data, which can be regarded as a special case of clustering with the constraint that for each characteristic, there are at most two clusters and they are linearly ordered. Suppose that the functional observations arise from true signals \( f_i(t), t \in [0,1], i = 1, \ldots, n \), corrupted by some noise process, where \( n \) denotes the sample size. We observe the functional data at some discrete time points. Then the model can be represented as

\[
Y_i(T_l) = f_i(T_l) + \varepsilon_{il},
\]

where \( \varepsilon_{il} \) is assumed to follow a normal distribution with mean 0 and variance \( \sigma^2 \), and is independent across \( i \) and \( l \). Let \( Y_i = (Y_i(T_1), \ldots, Y_i(T_m))^T \) be the \( i \)th observation at points \( T_1, \ldots, T_m \), where \( T_l \in [0,1] \), for \( l = 1, \ldots, m \). Similarly, let \( f_i = (f_i(T_1), \ldots, f_i(T_m))^T \), and \( \varepsilon_i = (\varepsilon_i(T_1), \ldots, \varepsilon_i(T_m))^T \). For functional data, the discrete wavelet transform (DWT) is one of the most common feature extraction technique. To implement the DWT, \( m \) needs to be a power of 2, and \( T_1, \ldots, T_m \) need to be equidistant. For \( m \) that is not a power of 2, we can first smooth to obtain a function, and then take a power of 2 number of discrete points from that function. In terms of the orthonormal basis \( \{\phi_0\} \cup \{\psi_{jk} : j = 0, \ldots, J-1, k = 0, \ldots, 2^j - 1\} \), we can define the following DWT operator (Antoniadis et al., 2013):

\[
W : \mathbb{R}^m \to \mathbb{R}^m, \quad f \to (\alpha_0, \beta_0, \ldots, \beta_{J-1}),
\]

51
with $\beta_j = (\beta_{j,0}, \ldots, \beta_{j,2^j-1})$. Applying the DWT operator on $Y_i = f_i + \varepsilon_i$, then we have

$$WY_i = Wf_i + W\varepsilon_i,$$

(3.3)

where $W\varepsilon_i \overset{d}{=} \varepsilon_i$ by the orthogonality of $W$. Let $\alpha_0$ denote the scaling coefficient at the level 0, and $\beta_{jk}$ be the wavelet coefficients at the multiresolution level $j, k$. As a result, (3.3) can be rewritten as

$$a_0^{(i)} = a_0^{(i)} + e_0^{(i)}, \quad b_{jk}^{(i)} = \beta_{jk}^{(i)} + e_{jk}^{(i)},$$

(3.4)

where $e_0^{(i)}$ and $e_{jk}^{(i)}$ follow a normal distribution with mean 0 and variance $\sigma^2$, for $k = 0, \ldots, 2^j - 1, j = 0, \ldots, J - 1$.

When the functional data are (essentially) observed continuously in time, we also consider the following infinite Gaussian white noise model

$$dY_i(t) = f_i(t)dt + \sigma dB_i(t),$$

(3.5)

where $B_i(\cdot)$ are independent Brownian motions on $[0, 1]$. Let

$$a_0^{(i)} = \int_0^1 \phi_0(t)dY_i(t), \quad a_0^{(i)} = \int_0^1 \phi_0(t)f_i(t)dt,$$

$$b_{jk}^{(i)} = \int_0^1 \psi_{jk}(t)dY_i(t), \quad \beta_{jk}^{(i)} = \int_0^1 \psi_{jk}(t)f_i(t)dt,$$

(3.6)

$$e_0^{(i)} = \sigma \int_0^1 \phi_0(t)dB_i(t), \quad e_{jk}^{(i)} = \sigma \int_0^1 \psi_{jk}(t)dB_i(t).$$

Then $e_0^{(i)}$ and $e_{jk}^{(i)}$ follow the normal distribution with mean 0 and variance $\sigma^2$, for $k = 0, \ldots, 2^j - 1, j = 1, 2, \ldots$, independent of each other, for each $i = 1, \ldots, n$. 

52
To detect the change point of this sequence of functional data, we first find the change in each component, that is, we detect the change for each feature $\beta_{jk}$, and we decide the overall change point from them.

In the section on posterior consistency, we state the results only in terms of the infinite model. However, in practice, we can only work with the finite model. By letting $\beta_{jk}^{(i)} = 0$ for all $j > J$, the infinite model can be related to the finite one with a random $J$. If the coefficients are obtained following the schema of Abramovich, Sapatinas, and Silverman (1998), then $J$ will have a limiting Poisson distribution by Proposition 1 of Suarez and Ghosal (2016). Under this schema, the total number of nonzero coefficients also has a limiting Poisson distribution.

### 3.3 Prior Distributions

For each $\beta_{jk}^{(i)}$, we define the following probabilities:

$$P(\beta_{jk}^{(i)} \neq 0) = \pi_j, \quad P(\beta_{jk}^{(i)} = 0) = 1 - \pi_j.$$  \hspace{1cm} (3.7)

As the wavelet coefficients of a signal function are sparse, Abramovich, Sapatinas, and Silverman (1998) proposed the following priors incorporating this characteristic feature of wavelet coefficients:

$$\beta_{jk}^{(i)} \overset{\text{ind}}{\sim} \pi_j N(0, c_j^2 \sigma^2) + (1 - \pi_j)\delta_0,$$  \hspace{1cm} (3.8)

where $\delta_0$ is a point mass at 0, and the hyperparameters in (3.8) are given by

$$c_j^2 = \nu_1 2^{-\gamma j}, \quad \pi_j = \min(1, \nu_2 2^{-\gamma j}), \quad j = 0, \ldots, J - 1.$$  \hspace{1cm} (3.9)
and \( \nu_1, \nu_2, \gamma_1 \geq 0 \), and \( 0 \leq \gamma_2 \leq 1 \). A vague prior is placed on \( \alpha_0 \).

Let \( \psi \) be a mother wavelet function of regularity \( r \). Consider constants \( s, p \) and \( q \) such that \( \max(0, 1/p - 1/2) < s < r \), \( 1 \leq p, q \leq \infty \). If either

\[
s + \frac{1}{2} - \frac{\gamma_2}{p} - \frac{\gamma_1}{2} < 0, \quad (3.10)
\]

or

\[
s + \frac{1}{2} - \frac{\gamma_2}{p} - \frac{\gamma_1}{2} = 0, \text{ and } 0 \leq \gamma_2 < 1, 1 \leq p < \infty, q = \infty, \quad (3.11)
\]

then \( f \in B^s_{p,q} \) almost surely, where \( B^s_{p,q} \) denotes Besov space of index \((p, q)\) and smoothness \( s \) (Abramovich, Sapatinas, and Silverman, 1998).

The prior on \( \sigma \) is given by

\[
\sigma^2 \sim IG(\theta, \lambda), \quad (3.12)
\]

where IG stands for the inverse gamma distribution. Let \( g \) denote the density function of the inverse-gamma distribution.

### 3.4 Posterior Probabilities of Change Point

For any \( j, k \), let \( \tau_{jk} \) denote the change point, and let \( \tau_{jk} \) take possible values \( 1, \ldots, n \). Let \( \rho_i \) denote the prior probability of changing at point \( i \), where \( \rho_i > 0 \), \( \sum_{i=1}^n \rho_i = 1 \), and \( i = 1, \ldots, n \). Then the posterior probability of \( \tau_{jk} = i \) is

\[
P(\tau_{jk} = i | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) = \frac{P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = i) \rho_i}{\sum_{i=1}^N P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = l) \rho_l}. \quad (3.13)
\]
The main problem is to compute the marginal likelihood \( P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = i) \).

When \( \tau_{jk} = 1 \), it is the initial state meaning no change. For \( \tau_{jk} = 2, \ldots, n \), the marginal likelihood is derived from four scenarios: change from zero to zero (which is no change), change from zero to non-zero, change from non-zero to zero, and change from non-zero to non-zero.

### 3.4.1 Initial State

When \( \tau_{jk} = 1 \), this is the initial state. If the initial state is zero, then the marginal likelihood is given by

\[
(1 - \pi_j) \int \left\{ \prod_{i=1}^{n} \phi(b_{jk}^{(i)}, 0, \sigma^2) \right\} g(\sigma^2; \theta, \lambda) d\sigma.
\] (3.14)

If the initial state is non-zero, then we have

\[
\pi_j \int \int \left\{ \prod_{i=1}^{n} \phi(b_{jk}^{(i)}, \xi, \sigma^2) \right\} \phi(\xi; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi d\sigma.
\] (3.15)

Thus, the marginal likelihood of the initial state is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = 1) = (1 - \pi_j)(2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \left[ \frac{\sum_{i=1}^{n} (a_0^{(i)})^2}{2} + \lambda \right]^{(n/2 + \theta)} \\
+ \pi_j(2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} (c_j^2 n + 1)^{-1/2} \left[ \frac{\sum_{i=1}^{n} (a_0^{(i)})^2}{2} - \frac{c_j^2}{c_j^2 n + 1} \left( \frac{\sum_{i=1}^{n} a_0^{(i)})^2}{2} + \lambda \right)^{(n/2 + \theta)}. \]
\] (3.16)
3.4.2 Non-initial State

For $\tau_{jk} = i$, where $i = 2, \ldots, n$, if changing from zero to zero, that is, no change, then the marginal likelihood is

$$(1 - \pi_j)^2 \int \left\{ \prod_{i=1}^{n} \phi(b_{jk}^{(i)}; 0, \sigma^2) \right\} g(\sigma^2; \theta, \lambda) d\sigma. \quad (3.17)$$

If changing from zero to non-zero at $\tau_{jk} = i$, then the marginal likelihood is

$$(1 - \pi_j) \pi_j \int \int \left\{ \prod_{l=1}^{i-1} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} \left\{ \prod_{l=i}^{n} \phi(b_{jk}^{(l)}; \xi, \sigma^2) \right\} \phi(\xi; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi d\sigma. \quad (3.18)$$

If changing from non-zero to zero at $\tau_{jk} = i$, then the marginal likelihood is

$$\pi_j (1 - \pi_j) \int \int \left\{ \prod_{l=1}^{i-1} \phi(b_{jk}^{(l)}; \xi, \sigma^2) \right\} \left\{ \prod_{l=i}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} \phi(\xi; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi d\sigma. \quad (3.19)$$

If changing from non-zero to non-zero at $\tau_{jk} = i$, then the marginal likelihood is

$$\pi_j^2 \int \int \left\{ \prod_{l=1}^{i-1} \phi(b_{jk}^{(l)}; \xi_1, \sigma^2) \right\} \left\{ \prod_{l=i}^{n} \phi(b_{jk}^{(l)}; \xi_2, \sigma^2) \right\}$$

$$\times \phi(\xi_1; 0, c_j^2 \sigma^2) \phi(\xi_2; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi_1 d\xi_2 d\sigma. \quad (3.20)$$
Thus, we have

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = i) \\
= (1 - \pi_j)^2 (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \left[ \frac{\sum_{l=1}^{n} (b_{jk}^{(l)})^2}{2} + \lambda \right]^{(n/2 + \theta)} \\
+ \pi_j (1 - \pi_j) (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} [c_j^2(i - 1) + 1]^{-1/2} \frac{\Gamma(n/2 + \theta)}{[B_{i-1} + \lambda]^{(n/2 + \theta)}} \\
+ (1 - \pi_j) \pi_j (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} [c_j^2(n - i + 1) + 1]^{-1/2} \frac{\Gamma(n/2 + \theta)}{[B_{n-i+1} + \lambda]^{(n/2 + \theta)}} \\
+ \pi_j^2 (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} [c_j^2(i - 1) + 1]^{-1/2} \frac{\Gamma(n/2 + \theta)}{[B_{i-1} + B_{n-i+1} - \sum_{l=1}^{n} (b_{jk}^{(l)})^2/2 + \lambda]^{(n/2 + \theta)}},
\]

where

\[
B_{i-1} = \frac{\sum_{l=1}^{n} (b_{jk}^{(l)})^2}{2} - \frac{c_j^2}{c_j^2(i - 1) + 1} \frac{(\sum_{l=1}^{i-1} b_{jk}^{(l)})^2}{2},
\]

and

\[
\tilde{B}_{n-i+1} = \frac{\sum_{l=1}^{n} (b_{jk}^{(l)})^2}{2} - \frac{c_j^2}{c_j^2(n - i + 1) + 1} \frac{(\sum_{l=i}^{n} b_{jk}^{(l)})^2}{2}.
\]

Similarly, we can compute the marginal likelihood \(P(a_0^{(1)}, \ldots, a_0^{(n)} | \tau_0 = i)\), where \(\tau_0\) denotes the change point in \(a_0^{(1)}, \ldots, a_0^{(n)}\), and obtain the posterior probability of \(\tau_0 = i\) through Bayes’s rule.
3.5 Change Point Detection

The change point of a sequence of functional data is the accumulative effect of all features where the contrast is the largest before and after. Since its the special case of clustering, following Suarez and Ghosal (2016), to quantify the similarity between two functional data, we need to consider the following similarity matrix. Suppose that there are $J$ levels and $i < i'$. Then the similarity between $i$th and $i'$th functional data is

$$S(i, i') = \frac{1}{2J} \left[ \mathbb{1}(\alpha_0^{(i)} = \alpha_0^{(i')}) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \mathbb{1}(\beta_{jk}^{(i)} = \beta_{jk}^{(i')}) \right].$$  \hfill (3.24)$$

For any $k$ that divides the data into two groups, we compute the ratio of the mean similarity between group and the mean similarity within group. We denote the ratio by $C(k)$. The change point is where this ratio is the minimum. Here we assume that $3 \leq k \leq n - 1$, which means there are at least two data points in each group. Then

$$\arg\min_k C(k) = \frac{\sum_{1 \leq i \leq k-1, k \leq j \leq N} S_{ij} / [(k-1)(n-k+1)]}{\{\sum_{1 \leq i \leq j \leq k-1} S_{ij} + \sum_{k \leq i \leq j \leq N} S_{ij} \}/[(k-1)^2 + (n-k+1)^2]}.$$  \hfill (3.25)$$

Since we cannot obtain the true value of $\alpha_0$ and $\beta_{jk}$, we take posterior expectation of (3.24) given the data. Then we have

$$E(S(i, i')|a_0^{(1)}, \ldots, a_0^{(n)}, b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) = \frac{1}{2J} \left[ P(\alpha_0^{(i)} = \alpha_0^{(i')}|a_0^{(1)}, \ldots, a_0^{(n)}) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')}|b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \right].$$  \hfill (3.26)$$

$P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')}|b_{jk}^{(1)}, \ldots, b_{jk}^{(n)})$ can be obtained from the expression for the posterior proba-
probability of the change point:

\[ P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')} | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) = P(\tau_{jk} \leq i \text{ or } \tau_{jk} \geq i' + 1 | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) + \sum_{t=i+1}^{i'} P(\beta_{jk} = 0, \tau_{jk} = t | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \]

\[ = \sum_{t=1}^{i} P(\tau_{jk} = t | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) + \sum_{t=i+1}^{i} P(\tau_{jk} = t | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) + \sum_{t=i+1}^{i'} P(\beta_{jk} = 0, \tau_{jk} = t | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \]

(3.27)

where \( P(\tau_{jk} = t | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \) are obtained from (3.13). Similarly, we can obtain \( P(\alpha_0^{(i)} = \alpha_0^{(i')} | a_0^{(1)}, \ldots, a_0^{(n)}) \).

3.6 Posterior Consistency

In this section, we state a posterior consistency result for the infinite model. With some minor notational modification, the result also holds for the finite model with a fixed depth \( J \).

We study consistency in our model when \( \sigma^2 \to 0 \). This is equivalent to averaging \( r \) i.i.d. replications of the observations with \( r \to \infty \), and replacing \( \sigma^2 \) by \( \sigma^2 / r \) with a known \( \sigma^2 \). To simplify notation, we assume that \( \alpha_0^{(i)} = 0 \) for \( i = 1, \ldots, n \). Let \( f = (f_1, \ldots, f_n) \).

Then the square of the norm on \( f \) is defined by

\[ \| f \|^2 = \sum_{i=1}^{n} \| f_i \|^2 = \sum_{i=1}^{n} \sum_{j=0}^{2^J-1} \sum_{k=0}^{2^J-1} |\beta_{jk}^{(i)}|^2. \]

(3.28)

59
We define the square of Sobolev norm on \( f \) as
\[
\|f\|_{\mathcal{H}^s}^2 = \sum_{i=1}^{n} \sum_{j=0}^{\infty} 2^{2js} \|\beta_j^{(i)}\|_2^2.
\] (3.29)

We denote this space by \( \mathcal{H}^s_n \), where \( s \) is the number of weak derivatives of the function in \( L_2([0,1]) \). Let \( D_r \) be the set of all observations.

**Theorem 5.** Let \( \gamma_1 > 2s + 1 \), and \( f_0 \in \mathcal{H}^s_n \) be the vector of true functions. Then the posterior is consistent, i.e., for any \( \epsilon > 0 \), \( \Pi(\|f - f_0\| < \epsilon | D_r) \to 1 \) in probability as \( r \to \infty \).

**Proof.** Let \( \Pi \) be a prior on \( \mathcal{H}^s_n \). Schwartz’s theorem (Schwartz, 1965) gives the strong consistency of the posterior distribution under approximate condition. According to Example 6.20 in Ghosal and van der Vaart (2017), if the Kullback-Leibler property holds for the prior, then the posterior distribution is consistent in the weak topology. Thus, for prior in space \( \mathcal{H}^s_n \), we need \( f_0 \) in the Kullback–Leibler support of \( \Pi \). The Kullback-Leibler divergence is defined as \( \mathcal{H}(f_0, f) = \sum_{i=1}^{n} \int f_{i,0} \log(f_{i,0}/f_i) d\mu \), where \( \mu \) is a dominating measure on the space of \( f \). In other words, we want \( \Pi(\mathcal{H}(f_0, f) < \epsilon) > 0 \) for all \( \epsilon > 0 \). The prior setting in Section 3.3 reduces \( \mathcal{H}(f_0, f) \) to the Kullback–Leibler divergence between two Gaussian distributions that is the Kullback–Leibler divergence between \( \beta_jk \) and \( \beta_{jk,0} \), and thus \( \Pi(\mathcal{H}(f_0, f) < \epsilon) \) is bounded by
\[
\Pi \left( \sum_{i=1}^{n} \sum_{j=0}^{\infty} \sum_{k=0}^{2j-1} |\beta_j^{(i)} - \beta_{jk,0}^{(i)}|^2 < \epsilon^2 \right),
\] (3.30)

where \( \beta_{jk,0} \) are the wavelet coefficients of the true function \( f_{i,0} \).

First, we consider a bounded subset \( \mathcal{H}^s_n(B) = \{ f \in \mathcal{H}^s_n, \|f\|_{\mathcal{H}^s_n} < B \} \) of the Sobolev
space. The Lemma 1 and 2 of Lian (2011) imply that (3.30) is positive. Thus, for any $B > 0$, we have

$$\Pi (f \in \mathcal{H}_{n}^{s}(B) : \|f - f_0\| > \epsilon |D_r|) \rightarrow 0 \text{ in probability.} \quad (3.31)$$

To complete the proof, we need to show that $\lim_{B \rightarrow \infty} \sup_{r > 0} E_{f_0} \Pi(\mathcal{H}_{n}^{s}(B)^c | D_r) = 0$. By Markov’s inequality, we have

$$\Pi(\mathcal{H}_{n}^{s}(B)^c | D_r) \leq B^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=0}^{\infty} 2^{2j} \sum_{k=0}^{2^j - 1} E \left( |\beta_{jk}^{(i)}|^2 \bigg| D_r \right) \right\}. \quad (3.32)$$

The expectation can be bounded by

$$E \left( |\beta_{jk}^{(i)}|^2 \bigg| D_r \right) = \sum_{t=1}^{n} E \left( |\beta_{jk}^{(i)}|^2 \bigg| \tau_{jk} = t, D_r \right) \Pi(\tau_{jk} = t | D_r) \leq \max_{1 \leq t \leq n} E \left( |\beta_{jk}^{(i)}|^2 \bigg| \tau_{jk} = t, D_r \right). \quad (3.33)$$

For $\tau_{jk} = t$, the posterior distribution of the common value $\xi$ of $\{\beta_{jk}^{(i)} : i = 1, \ldots, t - 1\}$ given ($\beta_{jk} \neq 0, D_r$) is proportional to

$$\left\{ \prod_{l=1}^{t-1} \exp \left\{ -\frac{1}{2\sigma^2/r} (b_{jk}^{(l)} - \xi)^2 \right\} \right\} \times \exp \left\{ -\frac{1}{2c_j^2\sigma^2} \xi^2 \right\} \times \exp \left\{ -\frac{1}{2c_j^2\sigma^2} \xi^2 \right\} \propto \exp \left\{ -\frac{(t - 1)c_j^2 + 1/r}{2c_j^2\sigma^2/r} \left[ \xi^2 - 2\frac{c_j^2 \sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t - 1)c_j^2 + 1/r} \xi \right] \right\}, \quad (3.34)$$

and hence the corresponding distribution is $N \left( \frac{c_j^2 \sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t - 1)c_j^2 + 1/r}, \frac{c_j^2\sigma^2}{1 + (t - 1)c_j^2/r} \right)$. Thus, for $i < t,$
we have

\[
E \left( \left| \beta^{(i)}_{jk} \right|^2 \right| \tau_j = t, D_r) = \frac{c_j^2 \sigma^2}{1 + (t - 1)c_j^2} + \left( \frac{c_j^2}{(t - 1)c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2. \quad (3.35)
\]

Similarly, for \( i \geq t, \)

\[
E \left( \left| \beta^{(i)}_{jk} \right|^2 \right| \tau_j = t, D_r) = \frac{c_j^2 \sigma^2}{1 + (n - t + 1)c_j^2} + \left( \frac{c_j^2}{(n - t + 1)c_j^2 + 1/r} \right)^2 \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2. \quad (3.36)
\]

Note that if \( t = 1, \) we only need to consider (3.36). Both (3.35) and (3.36) can be bounded by

\[
E \left( \left| \beta^{(i)}_{jk} \right|^2 \right| \tau_j = t, D_r) \leq c_j^2 \sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2. \quad (3.37)
\]

Thus we have

\[
\Pi(\mathcal{H}_n^s(B)^c|D_r) \leq B^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=0}^{\infty} 2^{2^j s} \sum_{k=0}^{2^j - 1} \left[ c_j^2 \sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 \right] \right\}. \quad (3.38)
\]

Now we take the expectation of (3.38) with respect to \( f_0 \) to obtain

\[
E_{f_0} \Pi(\mathcal{H}_n^s(B)^c|D_r) \leq B^{-2} \left\{ \sum_{j=0}^{\infty} 2^{2^j s} \sum_{k=0}^{2^j - 1} \left[ c_j^2 \sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 \right] \right\}. \quad (3.39)
\]
Replacing the hyperparameters using (3.9), we can further bound (3.39) by

\[
B^{-2} \left\{ n(n+1)\sigma^2 \nu_1 \sum_{j=0}^{\infty} 2^{(2s+1-\gamma_1)j} + n^2 \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2j-1} \sum_{l=1}^{n} |\beta_{jk,l}^{(l)}|^2 \right\}
\]

\[
= B^{-2} \left\{ n(n+1)\sigma^2 \nu_1 \sum_{j=0}^{\infty} 2^{(2s+1-\gamma_1)j} + n^2 \|f_0\|_{\mathcal{H}_n}^2 \right\}.
\]

Under the assumption that \(\gamma_1 > 2s + 1\), we have

\[
n(n+1)\sigma^2 \nu_1 \sum_{j=0}^{\infty} 2^{(2s+1-\gamma_1)j} + n^2 \|f_0\|_{\mathcal{H}_n}^2 < \infty.
\]

(3.41)

Thus, (3.39) goes to 0 as \(B \to \infty\). \(\square\)

We also need to show that we find the right model. For given \(j, k\), we can define the following structures 1 to 5 denoted by \(S_{jk}^1, \ldots, S_{jk}^5\):

1. Change from nonzero to nonzero at \(\tau_{jk} = t\):
   \[
   S_{jk}^1 = \{\tau_{jk} = t, \beta_{jk}^{(1)} = \cdots = \beta_{jk}^{(t-1)} = \xi_1, \beta_{jk}^{(t)} = \cdots = \beta_{jk}^{(n)} = \xi_2, \xi_1 \neq \xi_2\};
   \]

2. Change from nonzero to zero at \(\tau_{jk} = t\):
   \[
   S_{jk}^2 = \{\tau_{jk} = t, \beta_{jk}^{(1)} = \cdots = \beta_{jk}^{(t-1)} = \xi, \beta_{jk}^{(t)} = \cdots = \beta_{jk}^{(n)} = 0, \xi \neq 0\};
   \]

3. Change from zero to nonzero at \(\tau_{jk} = t\):
   \[
   S_{jk}^3 = \{\tau_{jk} = t, \beta_{jk}^{(1)} = \cdots = \beta_{jk}^{(t-1)} = 0, \beta_{jk}^{(t)} = \cdots = \beta_{jk}^{(n)} = \xi, \xi \neq 0\};
   \]

4. No change and the value is nonzero:
   \[
   S_{jk}^4 = \{\beta_{jk}^{(1)} = \cdots = \beta_{jk}^{(n)} = \xi, \xi \neq 0\};
   \]

5. No change and the value is zero:
   \[
   S_{jk}^5 = \{\beta_{jk}^{(1)} = \cdots = \beta_{jk}^{(n)} = 0\}.
   \]
Table 3.1: Compatible models

<table>
<thead>
<tr>
<th>True Structure</th>
<th>Compatible Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure 1</td>
<td>Structure 1</td>
</tr>
<tr>
<td>Structure 2</td>
<td>Structure 1,2</td>
</tr>
<tr>
<td>Structure 3</td>
<td>Structure 1,3</td>
</tr>
<tr>
<td>Structure 4</td>
<td>Structure 1,4</td>
</tr>
<tr>
<td>Structure 5</td>
<td>Structure 1,2,3,4,5</td>
</tr>
</tbody>
</table>

We define a compatible model as the structure that not only has the same change point as the true model, but also can have $\beta_{jk}^{(i)}$ values in the neighborhood of the true value $\beta_{jk,0}^{(i)}$. For example, if the true Structure is 5, then the compatible model can be Structure 1–5, because a nonzero value $\xi$ can be small enough to be in the neighborhood of 0. If the true Structure is 1, then the only compatible model is itself, because 0 cannot be in the neighborhood of a predetermined nonzero value. Table 3.1 shows the compatible models for each true structure. Theorem 1 implies that the posterior probability of $\beta_{jk}^{(i)}$ in any neighborhood of the true value $\beta_{jk,0}^{(i)}$ tends to 1. This shows that the posterior probability of all non-compatible models together tends to 0. Hence for consistency of model selection, we only need to consider compatible models.

**Lemma 1.** Let $S_{jk,0}$ denote the true structure for given $j, k$. Then $\Pi(S_{jk} = S_{jk,0}|D_r) \to 1$ in probability as $r \to \infty$.

**Proof.** It suffices to show that the ratio of the marginal likelihood of a compatible structure other than the true structure and the true structure goes to zero in probability. In this proof, we only show the cases when the true parameter has Structures 4 or 5. The proofs for other cases follow from similar arguments.

First, we need to compute the following marginal likelihoods with a known $\sigma^2$. The
marginal likelihood for Structure 1 is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^1) \\
= \int \int \left\{ \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}; \xi_1, \sigma^2/r) \right\} \left\{ \prod_{l=t}^{n} \phi(b_{jk}^{(l)}; \xi_2, \sigma^2/r) \right\} \phi(\xi_1; 0, c_j^2 \sigma^2) \phi(\xi_2; 0, c_j^2 \sigma^2) \, d\xi_1 \, d\xi_2 \\
= (c_j^2 r (t - 1) + 1)^{-1/2} (c_j^2 r (n - t + 1) + 1)^{-1/2} (2\pi \sigma^2/r)^{-n/2} \\
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2(t - 1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 \\
+ \frac{c_j^2}{c_j^2(n - t + 1) + 1/r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}. \tag{3.42}
\]

The marginal likelihood for Structure 2 is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^2) \\
= \int \left\{ \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}; \xi, \sigma^2/r) \right\} \left\{ \prod_{l=t}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2/r) \right\} \phi(\xi; 0, c_j^2 \sigma^2) \, d\xi \\
= (c_j^2 r (t - 1) + 1)^{-1/2} (2\pi \sigma^2/r)^{-n/2} \tag{3.43} \\
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2(t - 1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}. \]
The marginal likelihood for Structure 3 is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^3) = \int \left\{ \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}; 0, \sigma^2/r) \right\} \left\{ \prod_{l=t}^{n} \phi(b_{jk}^{(l)}, \xi, \sigma^2/r) \right\} \phi(\xi; 0, c_j^2 \sigma^2) d\xi
\]

\[
= (c_j^2 r (n - t + 1) + 1)^{-1/2} (2\pi \sigma^2/r)^{-n/2} \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 n + 1/r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} b_{jk}^{(l)} \right]^2 \right\}.
\]

The marginal likelihood for Structure 4 is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^4) = \int \left\{ \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; \xi, \sigma^2/r) \right\} \phi(\xi; 0, c_j^2 \sigma^2) d\xi
\]

\[
= (c_j^2 r n + 1)^{-1/2} (2\pi \sigma^2/r)^{-n/2} \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 n + 1/r} \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} b_{jk}^{(l)} \right]^2 \right\}.
\]

The marginal likelihood for Structure 5 is

\[
P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^5) = \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2/r)
\]

\[
= (2\pi \sigma^2/r)^{-n/2} \exp \left\{ -\frac{r}{2\sigma^2} \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right\}.
\]

If the true parameter has Structure 4, and the compatible model is Structure 1, then
we have the following marginal likelihood ratio:

\[
\frac{P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^{1})}{P(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^{4})} = \left[ \frac{c_{j}^{2}r_{n} + 1}{(c_{j}^{2}r(t - 1) + 1)(c_{j}^{2}r(n - t + 1))} \right]^{1/2}
\]

\[
\times \exp \left\{ \frac{r}{2\sigma^{2}} \left[ \frac{c_{j}^{2}}{c_{j}^{2}(t) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^{2} + \frac{c_{j}^{2}}{c_{j}^{2}(n - t + 1)} \right] \right\}.
\]

The first term in the square root goes to 0 as \( r \to \infty \). Hence it suffices to show that the form inside the exponential is \( O_p(1) \). Being a special case of clustering, our situation is similar to that of Suarez and Ghosal (2016), but it seems that their argument is incomplete as they overlooked a factor \( r \). For the sake of completeness, we present the argument, which can also be used to complete the proof Lemma 1 of Suarez and Ghosal (2016).

As \( r \to \infty \), we have

\[
\frac{r}{2\sigma^{2}} \left[ \frac{c_{j}^{2}}{c_{j}^{2}(t - 1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^{2} + \frac{c_{j}^{2}}{c_{j}^{2}(n - t + 1) + 1/r} \right] = \frac{r}{2\sigma^{2}} \left[ \frac{\left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^{2}}{(t - 1) + O(1/r)} + \frac{\left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^{2}}{(n - t + 1) + O(1/r)} - \frac{\left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^{2}}{n + O(1/r)} + O(1) \right]
\]

\[
= \frac{r}{2\sigma^{2}} \left[ \frac{1}{(t - 1)} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^{2} + \frac{1}{(n - t + 1)} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^{2} - \frac{n}{n} \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^{2} \right] + O(1).
\]

(3.48)

Let \( U = \sum_{l=1}^{t-1} b_{jk}^{(l)} \) and \( V = \sum_{l=t}^{n} b_{jk}^{(l)} \). Consider a random variable \( W \) which has the
following distribution:

\[
W = \begin{cases} 
    \frac{U}{t-1}, & \text{with probability } \frac{t-1}{n}, \\
    \frac{V}{n-t+1}, & \text{with probability } \frac{n-t+1}{n}.
\end{cases} 
\]  

(3.49)

Let \( \psi(w) = w^2 \). Then by Jensen’s inequality, we have

\[
\left[ \frac{U}{t-1} \frac{t-1}{n} + \frac{V}{n-t+1} \frac{n-t+1}{n} \right]^2 
\leq \left[ \left( \frac{U}{t-1} \frac{2t-1}{n} + \left( \frac{V}{n-t+1} \frac{2n-t+1}{n} \right)^2 \right] n^{-t+1} \right].
\]

(3.50)

That is

\[
\frac{U^2}{t-1} + \frac{V^2}{n-t+1} - \frac{(U + V)^2}{n} \geq 0.
\]

(3.51)

Thus, the term in the brackets of the exponential in (3.47) is nonnegative. Hence it suffices to control its expectation and show that it remains bounded as \( r \to \infty \). Suppose that the true value is \( \beta_{jk,0}^{(i)} = \xi \). Then the expectation of (3.48) with respect to the true value is

\[
\frac{r}{2\sigma^2} \left[ \frac{(t-1)^2 \xi^2}{(t-1)} + \frac{(n-t+1)^2 \xi^2}{n} - \frac{(n-t+1)^2 \sigma^2}{n} - \frac{n \xi^2}{n} \frac{r}{n} \right] + o(1)
= \frac{r \xi^2}{2\sigma^2} \left[ (t-1) + (n-t+1) - n \right] + o(1),
\]

(3.52)

and the first term vanishes. Thus the exponential term in (3.47) is bounded in probability. Hence (3.47) goes to 0 as \( r \to \infty \).

If the true parameter has Structure 5, and the compatible model is Structure 1, then
we have the following marginal likelihood ratio:

\[
\frac{P(b^{(1)}_{jk}, \ldots, b^{(n)}_{jk} | S^1_{jk})}{P(b^{(1)}_{jk}, \ldots, b^{(n)}_{jk} | S^5_{jk})} = (c_j^2 r (t - 1) + 1)^{-1/2} (c_j^2 r (n - t + 1) + 1)^{-1/2} \times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 (t - 1) + 1/r \left( \sum_{l=1}^{t-1} b^{(l)}_{jk} \right)^2} + \frac{c_j^2}{c_j^2 (n - t + 1) + 1/r \left( \sum_{l=t}^{n} b^{(l)}_{jk} \right)^2} \right] \right\}.
\]  

(3.53)

The first two terms with the square root goes to 0 as \( r \to \infty \). Similarly, we have

\[
\frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 (t - 1) + 1/r \left( \sum_{l=1}^{t-1} b^{(l)}_{jk} \right)^2} + \frac{c_j^2}{c_j^2 (n - t + 1) + 1/r \left( \sum_{l=t}^{n} b^{(l)}_{jk} \right)^2} \right] = \frac{r}{2\sigma^2} \left[ \frac{\left( \sum_{l=1}^{t-1} b^{(l)}_{jk} \right)^2}{(t - 1)} + \frac{\left( \sum_{l=t}^{n} b^{(l)}_{jk} \right)^2}{(n - t + 1)} \right] + \mathcal{O}(1),
\]  

(3.54)

which is always nonnegative. Since the true value is \( \beta^{(i)}_{jk,0} = 0 \), then the expectation of (3.54) with respect to the true value is

\[
\frac{r}{2\sigma^2} \left[ \frac{(t - 1)\sigma^2}{(t - 1)} + \frac{(n - t + 1)\sigma^2}{(n - t + 1)} \right] + \mathcal{O}(1),
\]  

(3.55)

which is bounded. Hence we can conclude that the exponential term in (3.53) is bounded in probability, and the whole expression in (3.53) goes to 0 as \( r \to \infty \).

Similarly, we can show that the marginal likelihood ratio also goes 0 when the compatible models are Structure 2–4.
3.7 Simulation

In order to study the performance of our method, we implement it on a set of simulated data. Since we detect the change point through the features extracted by the DWT, we first generate the wavelet coefficients and then apply the inverse discrete wavelet transform to get the functional data. We generate 16 features for first data point by a uniform distribution on \([0, 0.5]\). To make the change distinguishable, we generate 16 features for last data point by a uniform distribution on \([0.5, 1]\). Suppose that we have 100 data points, and there is one change in each feature. We randomly sample 16 numbers from 1 to 100 and regard them as the change point for the 16 features. To generate the sequence of 100 data, we repeat the feature of the first data point and change it to that of the last data point after the change point. Thus, we have a sequence of data represented by the true feature values. After applying the inverse discrete wavelet transform, we get a sequence of 100 true signals. To generate features for the data with different noise, we sample from the normal distribution with the true feature values as the mean value and variance of 0.01, 0.1, and 1, respectively. Hence we get three sequences of 100 functional observations after applying the inverse discrete wavelet transform to them.

We apply our method to the observations. For the true signals, we use (3.24) to compute the similarity. The change point is the value of \(k\) where \(C(k)\) in (3.25) is the minimum. Once we detect the first change point, we divide the sequence of data into two subgroups. Furthermore, we can find the change point in these two groups. We can continue the process to divide the data into more subgroups, and stop either the plot of \(C(k)\) versus \(k\) is relatively flat which means that there is not much difference in these data, or the minimum number of data points is reached, or the the maximum step of the resulting binary tree is reached. In this study, we stop either if \(\max(C(k)) - \min(C(k)) < \)
0.1, or there are less than 10 data points in the group, or the resulting binary tree has 3 steps. We compare our results with the E-Divisive method (James and Matteson, 2014) in the R package `ecp`, which also estimates multiple change points by iteratively applying a procedure for locating a single change point. E-Divisible determines the number and locations of the change points by the scaled sample measure of divergence based on Euclidean distances (Matteson and James, 2014). We apply the E-Divisive method on the wavelet transform of the observations. Table 3.2 shows the change points for different sequences of observations detected by our method and E-Divisive method. The numbers in the parentheses denote the hierarchical order of the change points. When the variance is small (0.01), the change points our method detects are exactly the same as the true change points. With a larger variance (0.1), our method still can detect the most of the change points correctly. When the variance is large, naturally it would be difficult to detect the change points by any method.

### 3.8 Application on Climate Data

On Berkeley Earth (http://berkeleyearth.org/data/), we can find the land-surface monthly average temperature between 1753–2016. These temperatures are in degrees
Celsius and reported as anomalies relative to the average temperature from Jan. 1951 to Dec. 1980. We can construct a set of functional data by the 12 monthly average temperatures in each year. We smooth the data by the basis expansion. Thus we get 264 functional data ordered by the year. Figure 3.1 shows the plot of the 264 functional data. We believe there is a change in these functional data. Figure 3.2 displays the curves for every 66 years, and we can see the change in the pattern of curves.

Figure 3.1: Land-surface average temperature curves between 1753–2016

Figure 3.3 is the plot of $C(k)$ versus different $k$. We detect the change point of this sequence of functional data at the year 1914. Figure 3.4 shows the curves before the change point which are the years 1753–1913 and the curves after the change point which are the years 1914–2016. We can see that the patterns are very different in these two
Furthermore, we can find the change point in these two subgroups. Figure 3.5 is the plot of $C(k)$ versus $k$ between 1753–1913, and we detect the change point at year 1839. Figure 3.6 shows the curves before the change point which are the years 1753–1838 and the curves after the change point which are the years 1839–1913. Figure 3.7 is the plot of $C(k)$ versus $k$ between 1914–2016, and we detect the change point at year 1969. Figure 3.8 shows the curves before the change point which are the years 1914–1968 and the curves after the change point which are the years 1969–2016. Hence we divide the data into four subgroups. We continue the process to divide the data into more subgroups, and stop if $\max(C(k)) - \min(C(k)) < 0.1$. We generate 15 subgroups. Figure 3.9 demonstrates the hierarchical structure in the subgroups.

In Section 3.6, we assume that we have $r$ replications of the data. To align with this assumption, we may group the data by every several consecutive years, and treat each group as a non-separable block. We can represent each block by the most representative pattern. For example, we can group the climate data by every 10 years and get 26 blocks. There are multiple ways to generate the most representative patterns, hence we can get different replications of the 26 blocks. Note that when we choose the number of observations in each block, we need to make sure that there is no distinct difference in patterns in that block.
Figure 3.2: Land-surface average temperature curves for every 66 years
Figure 3.3: $C(k)$ for different $k$ between 1753–2016

Figure 3.4: The plot on the left is the land-surface average temperature curves between 1753–1913. The plot on the right is the land-surface average temperature curves between 1914–2016.
Figure 3.5: $C(k)$ for different $k$ between 1753–1913

Figure 3.6: The plot on the left is the land-surface average temperature curves between 1753–1838. The plot on the right is the land-surface average temperature curves between 1839–1913.
Figure 3.7: $C(k)$ for different $k$ between 1914–2016

Figure 3.8: The plot on the left is the land-surface average temperature curves between 1914–1968. The plot on the right is the land-surface average temperature curves between 1969–2016.
Figure 3.9: Subgroups between 1753–2016
REFERENCES


