BAYESIAN ANALYSIS FOR SPARSE FUNCTIONAL DATA

By

SHANGYUAN YE

Bachelors of Arts in Mathematics
University of Texas at Dallas
Richardson, Texas
May, 2014

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2018

BAYESIAN ANALYSIS FOR SPARSE FUNCTIONAL DATA

Dissertation Approved:

Dr. Ye Liang

Dissertation Adviser

Dr. Joshua Habiger

Dr. Lan Zhu

Dr. Carla Goad

Dr. Dursun Delen

Outside Committee Member

ACKNOWLEDGMENTS

First of all, I want to express my deepest appreciation and gratitude to my dissertation advisor Dr. Ye Liang, who brought me into the exciting world of functional data analysis. I can not complete my Ph.D. degree without his continuous encouragement and patiently guidance during the past few years.

I am grateful to my dissertation committee members; Dr. Joshua Habiger, Dr. Lan Zhu, Dr. Carla Goad and Dr. Dursun Delen for their valuable suggestions and comments. Special thanks are due to my previous advisor, Dr. Ibrahim Ahmad. His broad knowledge and great passion in probability and statistics set an example to me for my future research.

I would also like to acknowledge the support I received from all the professors, graduate students, staffs in the Statistics department, and all my friends at OSU. I specially wish to mention my roommate Xijia Han for his understanding and full help on taking care my dog.

I am indebted my aunt and her family for the many favors they did for me during the time of my studies in U.S.. Most of all, I wish to express my sincere thanks to my parents, who continued provide love and support throughout my whole life.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: SHANGYUAN YE

Date of Degree: MAY, 2018

Title of Study: BAYESIAN ANALYSIS FOR SPARSE FUNCTIONAL DATA

Major Field: STATISTICS

Abstract: This dissertation mainly presents a novel Bayesian method for sparse functional data. Specifically, two models are proposed, one of which models all individual functions with a common smoothness and the other groups individual functions with heterogeneous smoothness. The proposed method utilizes the mixed effects model representation of the penalized splines for both the mean function and the individual functions. Given noninformative or weakly informative priors, Bayesian inference on the proposed models are developed and computations are done by using Markov Chain Monte Carlo (MCMC) methods. It has been shown that the proposed Bayesian methods perform well on irregularly spaced sparse functional data, where a traditional mixed effects model may often fail. This dissertation also includes a small section on orthogonal series functional estimation for density functions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

1.1 Introduction

Longitudinal data arise frequently in medical, social and economic applications. They usually involve a collection of measurements at different time points for a number of subjects. The measurements are typically considered as dependent observations over time for the same subject. Often, the objective of a longitudinal analysis is to describe the relationship between the response variable and the time or other covariates. The time course is often too complicated to model parametrically and therefore, nonparametric methods for longitudinal data, also called the sparse functional data analysis, has received an increasing amount of attention recently.

In general, functional data analysis (FDA) (Ramsay & Silverman 2005) refers to the statistical analysis for random functions. That is, random curves are sample units in the analysis. Since data are only observed at a finite number of time points, the traditional FDA usually starts with data smoothing to estimate individual curves for each subject. The traditional FDA focuses on data which are repeatedly and regularly observed across all individuals (Rice & Silverman 1991, Besse & Ramsay 1986), which are called the dense functional data. However, for longitudinal data that arise in medical studies, each individual could only be observed at a small number of time points and furthermore, these time points could be irregularly spaced. The traditional

FDA is generally not applicable in such cases. Recently, there is an increasing number of studies aiming to develop functional data analysis methods for longitudinal data (Berk 2012, James et al. 2005, James 2010, Yao et al. 2005, Zhang & Wang 2016, Thompson & Rosen 2008, Wu & Zhang 2006).

There are two commonly used classes of methods for smoothing sparse functional data. The first class of methods assume that individual curves in a given population share the same covariance function. Then the problem of smoothing n individual univariate functions can be equivalent to the problem of smoothing a single bivariate function. Individual curves can then be predicted by this covariance function. Functional principle component analysis is among the first-line approaches of this class of methods (Besse & Ramsay 1986, Yao et al. 2005, Peng & Paul 2009), while other example of this class include Fan & Gijbels (1996), Xiao et al. (2017), Cai & Yuan (2010). The second class of methods, called the functional mixed effects models, assume mixed effects models which allow for strength borrowing among individuals. Brumback & Rice (1998) first proposed a penalized smoothing spline mixed effects model. Later developments include mixed effects smoothing splines (Berk 2012), semiparametric mixed effects models (Durban et al. 2005), and various methods employing B-splines (James et al. 2000, Thompson & Rosen 2008, Wu & Zhang 2006).

Most of the above approaches require multiple steps of estimation (e.g. functional principle analysis) and an additional step for model selection which selects the smoothness. The inference, following the parameter estimation, is conditioning on the selected optimal model, with additional assumptions and procedures. The separated steps and additional assumptions may lead to compromised functional estimates and inference. Furthermore, generalizing these methods to more complicated models, such as the additive model, may become problematic.

In this dissertation, two Bayesian functional mixed-effects models are proposed to remedy the above shortcomings. The two proposed models in Chapter 3 and 4 model both the mean function and individual functions by the semiparametric mixed model representation introduced in Ruppert et al. (2003). We consider a Bayesian framework where noninformative or weakly informative conjugate priors are used. For the first purposed model (Section 3.2), all individual functions are assumed to have a common smoothness, but different from the smoothness of the mean function. The model can be considered as the Bayesian counterpart of Durban et al. (2005). For the second purposed model (Section 4.2), we generalize the first model by allowing grouped smoothnesses for individual functions, that is, those individual functions may have different smoothnesses. Both models are fitted through Markov chain Monte Carlo (MCMC) methods described in Section 3.3 and 4.3. Section 3.4 and 4.4 present the results of simulation studies, in which we investigate the performance of the two proposed estimators. Section 3.5 and 4.5 illustrate our models on a publicly available CD4 dataset.

Finally, on an independent track, we propose an orthogonal series density estimator for complex surveys, where samples are neither independent nor identically distributed. In Section 5.3, statistical properties of the proposed estimator is proved. In Section 5.4, two data driven estimators are proposed based on the proposed oracle estimator. Section 5.5 reports the setting and results of a simulation study that compares the performance of our proposed estimators with the standard orthogonal series density estimator. A real survey data example is provided for an illustration in Section 5.6.

CHAPTER II

Literature Review

2.1 **Functional Data Analysis**

2.1.1 **Functional Data**

Functional data analysis (Ramsay & Silverman 2005) refers to the statistical analysis of random functions. A random function consists of a series of univariate or multivariate measurements over a continuum. Commonly used continuums include the time, the spatial location and other metrics. Hereafter, it will be assumed that the continuum is time, denoted by t. Functional data often refer to noisy realizations (observations) at discrete time points of a underlying function.

Longitudinal data also measure individual variables repeatedly over time. Traditionally, the distinction between the longitudinal data and the functional data rests upon the number of observed time points. Longitudinal data are typically observed at a much sparser set of time points than functional data, with substantial amount of missing data. Thus, longitudinal data also refer to sparse functional data.

Formally, let N be the total number of subjects and m_i be the number of repeat measurements for the i^{th} subject curve. Also assume that the $m_i's$ are all of the same order $m_i = O(N^{\eta})$ for some $\eta \geq 0$. Data with $\eta = 0$, *i.e.*, $m_i = O(1)$, are called sparse functional data; data satisfying $\eta \geq \eta_0$, where η_0 is a transition point to be specified, are called dense functional data; and data with $\eta \in (0, \eta_0)$ are called moderately dense functional data.

Sparse and dense functional data are analyzed with different methodologies. For dense functional data, one can smooth each curve separately, and then further estimation and inference can be obtained based on the pre-smoothed curves. For sparse functional data, the pre-smoothing approach is not applicable, and often either methods based on the covariance estimation or methods utilizing mixed-effects models (Cai & Yuan 2010, Xiao et al. 2017, Berk 2012, Yao et al. 2005) can be used to pool data together to borrow strength from individual curves.

2.1.2 Data Smoothing

In functional data analysis, we assume that data are noisy realizations of smooth underlying functions. The analysis of functional data usually starts with smoothing, which can be understood as the process of estimating the true function by using samples at discrete points.

Generally, functional data analyses are nonparametric. Some commonly used smoothing methods include local weighting (Benhenni & Degras 2014), basis function methods, and nonparametric Bayesian methods (Kaufman & Sain 2010). In this dissertation, we focus on basis function methods.

Formally, consider a single observation y_i taken at time point t_i that is modeled by

$$y_i = f(t_i) + \epsilon_i \tag{2.1}$$

where $f(t)$ is the function of interest and ϵ_i is an error term. Under the smoothness assumption, the infinitely dimensional function $f(t)$ can be parameterized by projecting it onto some known basis functions $\{B_j(t)\}$. Commonly used bases include polynomials, splines, wavelets, and Fourier bases. Ramsay & Silverman (2005) suggest using splines and wavelets for aperiodic functional data, and the similarities between these two bases under a semiparametric regression are discussed in Wand & Ormerod (2011). In this dissertation, we focus on penalized splines, while examples of functional data analysis using wavelets are Zhao et al. (2012), Antoniadis et al. (2013), and Giacofci et al. (2013).

Splines

Splines are piecewise polynomials formed by placing knots that divide the time course into a number of disjoint regions within which a separate polynomial is fit, while constraints on the spline ensure that there is a continuity between two adjacent regions (de Boor 1978). The number of the knots controls the smoothness of the underlying curve and local curve behavior is accommodated through selection of knot location.

Two most commonly used splines are the truncated power basis and basis splines (B-splines). For these two types of splines, a total of K knots have been placed at time locations $\tau_1, ..., \tau_K$. In terms of knot location, the heuristic method locates the knots at either equally spaced or equally spaced quantiles time points. The truncated power basis is given by

$$\phi(t_i) = [1, t_i, ..., t_i^p, (t_i - \tau_1)_+^p, ..., (t_i - \tau_K)_+^p]^T$$

where $[x]_+ = \max(0, x)$ and p is the degree of the polynomial used in each region. The basis matrix $\boldsymbol{Z}$ can be formed by evaluating the set of basis functions at each t_i. The main drawback of the truncated power basis is that the basis matrix $\boldsymbol{Z}$ can be ill-conditioned, i.e. the largest singular value divides the smallest singular value of the basis matrix is too large. For this reason, the truncated power basis tends to only occur in the context of penalized splines (Ruppert 2002, Ruppert et al. 2003), which allow a finer degree of control over the smoothness of the fit. Details for penalized splines will be discussed in a later section.

B-Splines are originally defined in terms of divided differences of truncated power basis. In recent literature, B-splines are constructed by using the Cox-de Boor algorithm (de Boor 1972) which defines the ith basis function of degree p as

$$B_{i,0}(t) = \begin{cases} 1, & \text{if } \tau_i \le t < \tau_{t+1} \\ 0, & \text{otherwise} \end{cases}, \quad B_{i,p}(t) = \frac{t - \tau_i}{\tau_{i+p} - \tau_i} B_{i,p-1}(t) + \frac{\tau_{i+p+1} - t}{\tau_{i+p+1} - \tau_{i+1}} B_{i+1,p-1}(t).$$

In this way, B-splines of higher orders are recursively defined in terms of those of lower orders. Compared with equivalent order truncated power basis, the basis matrix for B-splines is sparser and well-conditioned. Because of these attractive computational advantages, B-splines are more popular than other splines representations.

Smoothing Splines

With the truncated power basis or B-splines, the smoothness of the fit is controlled by the number of knots K. Thus, fitting a spline model requires a separate model selection for the knot sequence (Dung & Tjahjowidodo 2017).

An alternative approach is the *smoothing spline* (Gu 2013). The idea of smoothing spline comes from estimating the unknown function f in (2.1) by minimizing the penalized least square score

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \mathcal{J}(f), \tag{2.2}$$

where the first term discourages the lack of fit of f to the data (bias), the second term penalizes the roughness of f (variance), and the λ is called the smoothing parameter, which controls the trade-off between the bias and variance. In other words, the least square part in (2.2) treats the data as independent samples, and any dependence between these data, presented through the functional relationship, is captured by the penalty $\mathcal{J}(f)$. One popular penalization has the form $J(f) = \int \ddot{f}^2(t)dt$, where $\ddot{f} = d^2f/dt^2$ represents the second order derivative of the underlying function f. This is also called the *cubic smoothing spline*, because under this smoothness constraint, the (2.2) has an explicit, finite-dimensional, unique minimizer which is a natural cubic spline with knots at the unique values of $t_i, i = 1, ..., m$ (Hastie et al. 2009, Green & Silverman 1994).

Thus, since the solution is a natural cubic spline, we can write it as

$$f(t) = \sum_{j=1}^{m} u_j Z_j(t),$$

where $Z_j(t)$ are a m-dimensional set of basis functions for representing this family of natural cubic splines. Then (2.2) reduces to

$$\text{PSS} = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{u})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{u}) + \lambda \boldsymbol{u}^T\boldsymbol{D}\boldsymbol{u}, \tag{2.3}$$

where $\{\boldsymbol{Z}\}_{ij} = Z_j(t_i)$ and $\{\boldsymbol{D}\}_{jk} = \int Z_j''(t)Z_k''(t)dt$. Setting (2.3) equal to 0, taking derivative and solving, we see the solution of the above penalized least square is

$$\hat{\boldsymbol{u}} = (\boldsymbol{Z}^T\boldsymbol{Z} + \lambda\boldsymbol{D})^{-1}\boldsymbol{Z}\boldsymbol{y}, \tag{2.4}$$

and the fitted spline is given by

$$\hat{f}(t) = \sum_{j=1}^{m} \hat{u}_j Z_j(t).$$

The model selection is now reduced from determining both the number and location of the knots to optimizing the smoothing parameter λ. When $\lambda = 0$, the model can be any function that interpolates the data; when λ tends to infinity, (2.3) tends to a linear least square fit.

Penalized Splines

One drawback of smoothing splines is that placing a knot at each time point can incur a significant computational cost when the number of time points is large, due to the size of the resulting incidence matrix. For this reason, this dissertation focuses on the *penalized splines*, which combines the knot selection of regression splines with the penalty approach of smoothing splines.

Penalized splines often appear in the literatures of semiparametric regression, e.g. Ruppert et al. (2003). The semiparametric model for a regression function f can be written as

$$f(t) = \beta_0 + \beta_1 t + \sum_{k=1}^{K} u_k Z_k(t) \text{ for some } K < m, \tag{2.5}$$

where $Z_k(t)$ are spline basis functions. In this way, a moderate number of knots, which is less than the number of observation time points, can be used to minimize the computational cost while retaining the penalization to parameter estimation so that a fine degree of control over the resulting smoothness is still possible. Thus, the penalized least square score (2.3) for smoothing splines and its solution (2.4) can be also applied onto penalized splines. Wand (2003) suggests the number of knots $K = \min(m/4, 35)$, where m is number of possible time points.

The most commonly used penalty matrix $\boldsymbol{D}$ is defined as the $(K+2) \times (K+2)$

matrix

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \boldsymbol{I}_{K\times K}, \end{bmatrix}$$

and the corresponding estimator (2.4) becomes a generalized ridge regression with the parametric parameters β_0 and β_1 remain unpenalized (Wand 2003). Wand & Ormerod (2009) suggest to use the B-spline basis and the above penalization setup, called O'Sullivan penalized splines. The Algorithm 1 in Wand & Ormerod (2011) describes the construction of default $\boldsymbol{Z}$ matrices for penalized splines, and the R Software function "ZOSull.r" will be used to construct O'Sullivan spline basis matrices for this dissertation, which can be found in the web-supplement of Wand & Ormerod (2009).

2.1.3 Model Selection

Whether smoothing or penalized splines are used, it is necessary to determine some optimal level of smoothing by adjusting the smoothing parameter, which involves trading off between goodness-of-fit and model complexity. Popular model selection methods for splines include cross-validation, generalized cross-validation, and information criteria (e.g. AIC, BIC, AICc).

In this dissertation, however, we will focus on the mixed model representation of penalized splines introduced by Ruppert et al. (2003). Consider the spline model

(2.5), let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ and } \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}$$

be the coefficients of the polynomial functions and basis functions, respectively. Corresponding to these vectors, define

$$\boldsymbol{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \text{ and } \boldsymbol{Z} = \begin{bmatrix} Z_1(t_1) & \cdots & Z_K(t_1) \\ \vdots & \ddots & \vdots \\ Z_1(t_m) & \cdots & Z_K(t_m) \end{bmatrix}$$

Treating $\boldsymbol{u}$ as a set of random coefficient with $\text{Cov}(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I}$, and assuming the measurement errors ϵ_{ij}s are independent and identically distributed (i.i.d) as $N(0, \sigma_\epsilon^2)$, Ruppert et al. (2003) shows that the estimator (2.4) is equivalent to the best linear unbiased predictor (BLUP) of the above model, with $\lambda = \sigma_\epsilon^2/\sigma_u^2$.

2.2 Bayesian Inference

Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ be the observed data and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_D)$ be unknown parameters of interest. Bayesian inference focuses on the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{\int p(\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta}} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{2.6}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function and $p(\boldsymbol{\theta})$ is the prior distribution. However, the denominator in Equation (2.6) is intractable for most models. Instead, we need resort to approximation algorithms like Markov Chain Monte Carlo (MCMC) or mean field variational Bayes (MFVB). The above described representation of the penalized spline allows smoothing by a mixed effects model, and also leads to a Bayesian estimator, which will be discussed in later chapters.

2.2.1 Markov Chain Monte Carlo

The Markov Chain Monte Carlo (MCMC) method is among the central components of Bayesian computation (Geyer 1992, Robert & Casella 2004). The fundamental idea is to generate a collection of samples from the posterior distribution by constructing a Markov chain and use them to estimate expectations.

A Markov chain, denoted by $\{\boldsymbol{\theta}^{(t)}\},\ t = 0, 1, \cdots$, is a stochastic process satisfying

$$P(\boldsymbol{\theta}^{(t+1)} \in A | \boldsymbol{\theta}^{(0)}, \cdots, \boldsymbol{\theta}^{(t)}) = P(\boldsymbol{\theta}^{(t+1)} \in A | \boldsymbol{\theta}^{(t)}),$$

that is, the conditional distribution of $\boldsymbol{\theta}^{(t+1)}$ only depends on $\boldsymbol{\theta}^{(t)}$. If a Markov chain is irreducible, aperiodic, and positive recurrent, then the chain is *ergodic*, and the draws from the steady-state of the Markov chain can be used as simulated samples from the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{y})$. Specifically, given a set of samples $\{\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(M)}\} \sim p(\boldsymbol{\theta}|\boldsymbol{y})$ drawn from a ergodic Markov chain, we can estimate the expectation of any function $f(\boldsymbol{\theta})$ with respect to the posterior by

$$\mathrm{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[f(\boldsymbol{\theta})] \approx \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{\theta}^{(m)})$$

The Ergodic property of Markov chain ensures the above Monte Carlo estimator converges to the true value almost surely.

The two most popular methods of constructing a Markov chain to sample from the posterior distribution are the Metropolis-Hastings (MH) Algorithm (Hastings 1970), and the Gibbs Sampling (Geman & Geman 1984, Gelfand & Smith 1990). The MH algorithm simulates sample from the posterior distribution by using of the full joint density function and a (independent) proposal distribution. The main steps of MH algorithm have three steps: (1) Generate a candidate sample from the proposal distribution; (2) Compute the acceptance probability based upon the proposal distribution and the full joint distribution; (3) Accept or reject the candidate sample based on the acceptance probability. Given a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$ the procedure of the MH Algorithm states above can be summarized by the pseudo-code in Algorithm 1.

Algorithm 1 Metropolis-Hastings Algorithm

```
Initialize $\boldsymbol{\theta}^{(0)}$ with $p(\boldsymbol{\theta}^{(0)}|\boldsymbol{y}) > 0$
for iteration $i = 1, 2, \cdots, M$ do
  $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{i})$
  $R(\boldsymbol{\theta}', \boldsymbol{\theta}^{(i)}) = \frac{p(\boldsymbol{\theta}'|\boldsymbol{y})q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^{(i)}|\boldsymbol{y})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})}$
  $u \sim \text{Uniform}(0, 1)$
  if $u < \min(1, R(\boldsymbol{\theta}', \boldsymbol{\theta}^{(i)}))$ then
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$
  else
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
  end if
end for
```

Often, if $\boldsymbol{\theta}$ has many dimensions, it is sometimes more practical to implement the Metropolis-Hastings algorithm one variable at a time (Tierney 1994).

The choice of the proposal distribution, $q(\cdot|\cdot)$, is very important. It should be chosen so that candidate values explore the relevant subsets of the support of the stationary distribution in a reasonable number of iterations, and candidate values are not accepted nor rejected too frequently. If the proposal distribution is too diffused relative to the target distribution, then too many candidate values will be rejected so that the chain will often remain the same for many consecutive iterations. On the other hand, if the proposal distribution is too focused relative to the target distribution, then too many candidate values will be accepted so that the chain slowly explores the space and is highly autocorrelated. The goal is to find a suitable proposal distribution so that the parameter space is fully explored with the correct probability measure within a reasonable number of iterations. The ideal acceptance rate is often set to be 0.44 when the dimension of $\boldsymbol{\theta}$ is larger than one and 0.234 when the dimension of $\boldsymbol{\theta}$ is one (Gelman et al. 2013).

Gibbs sampling, on the other hand, simplifies a higher dimensional sampling to a much lower dimensional sampling. This procedure is attractive because many compositional models are designed such that the conditional distributions are easy to sample from. The Gibbs sampler is given in Algorithm 2

Algorithm 2 Gibbs Sampler

```
Initialize $\boldsymbol{\theta}^{(0)} \sim \pi(\boldsymbol{\theta})$
for iteration $i = 1, 2, \cdots, M$ do
    $\theta_1^{(i)} \sim p(\theta_1|\theta_2^{(i-1)}, \cdots, \theta_D^{(i-1)})$
    $\vdots$
    $\theta_k^{(i)} \sim p(\theta_k|\theta_1^{(i)}, \cdots, \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, \cdots, \theta_D^{(i-1)})$
    $\vdots$
    $\theta_D^{(i)} \sim p(\theta_D|\theta_1^{(i)}, \cdots, \theta_{D-1}^{(i)})$
end for
```

In some situations where the full conditional distribution is not easy to sample from, a Metropolis-Hastings update can be used. Let $q(\theta'|\boldsymbol{\theta})$ be a proposal density, where θ' is the proposal and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{k-1}, \theta_k, \theta_{k+1}, \cdots, \theta_D)$ is the current value before the update of θ_k. Then the Hasting ratio can be defined as

$$R_k(\theta', \theta_k) = \frac{p(\theta'|\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_D)q(\theta_k|\theta_1, \cdots, \theta_{k-1}, \theta', \theta_{k+1}, \cdots, \theta_D)}{p(\theta_k|\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_D)q(\theta'|\theta_1, \cdots, \theta_{k-1}, \theta_k, \theta_{k+1}, \cdots, \theta_D)}.$$

MCMC algorithms based on such a combination of Gibbs sampler and Metropolis-Hastings sampler are called Metropolis within Gibbs or hybrid Metropolis-Hastings algorithm (Chib & Greenberg 1995). More details can be found in Robert & Casella (2004) and Givens & Hoeting (2005).

CHAPTER III

Bayesian Functional Mixed-effects Model

3.1 **Introduction**

Longitudinal data arise frequently in medical, social and economic applications. They usually involve a collection of measurements at different time points for a number of subjects. The measurements are typically considered as dependent observations over time for the same subject. Often, the objective of a longitudinal analysis is to describe the relationship between the response variable and time or other covariates. The time course is often too complicated to model parametrically and therefore, nonparametric methods for longitudinal data, also called the sparse functional data analysis, has received an increasing amount of attention recently.

In general, functional data analysis (FDA) (Ramsay & Silverman 2005) refers to the statistical analysis for random functions. That is, random curves are sample units in the analysis. Since data are only observed at a finite number of time points, the traditional FDA usually starts with data smoothing to estimate individual curves for each subject. The traditional FDA focuses on data which are repeatedly and regularly observed across all individuals (Rice & Silverman 1991, Besse & Ramsay 1986), which are called the dense functional data. However, for longitudinal data that arise in medical studies, each individual could only be observed at a small number of time points and furthermore, these time points could be irregularly spaced. The traditional FDA is generally not applicable in such cases. We primarily focus on developing a Bayesian semiparametric method for smoothing sparse functional data. The proposed method is still useful for smoothing functional data that are not sparse.

For dense functional data, one may start with smoothing each individual curve separately, based on which the estimation and inference about the population are further developed. This approach is also said to be a "direct method". However, using direct method on sparse functional data usually does not perform well in practice. On the one hand, an individual that is only observed at a small number of time points does not provide enough information to get a reliable functional estimate; on the other hand, irregularly spaced time points may be clustered and only contain functional information on local areas. Moreover, when making inference about the population, assigning equal weights to individual curves with different number of time points leads to inefficient subsequent analysis. Therefore, in sparse functional data analysis, the superpopulation of individual curves needs to be modeled, so that individuals can borrow information from each other for the estimation and inference.

There are two commonly used classes of methods for smoothing sparse functional data. The first class of methods assume that individual curves in a given population share the same covariance function. Then the problem of smoothing n individual univariate functions can be equivalent to the problem of smoothing a single bivariate function. Individual curves can then be predicted by this covariance function conditioning on the observations. Functional principle component analysis is among the first-line approaches of this class of methods (Besse & Ramsay 1986, Yao et al. 2005, Peng & Paul 2009), while other example of this class include but not limit to Fan & Gijbels (1996), Xiao et al. (2017), Cai & Yuan (2010). The second class of methods, called the functional mixed effects models, assume mixed effects models which allow for strength borrowing among individuals. Brumback & Rice (1998) first proposed a penalized smoothing spline mixed effects model. Later developments include mixed effects smoothing splines (Berk 2012), semiparametric mixed effects models (Durban et al. 2005), and various methods employing B-splines (James et al. 2000, Thompson & Rosen 2008, Wu & Zhang 2006).

On the one hand, existing functional mixed effects models, usually fitted by the maximum likelihood (ML) or restricted maximum likelihood (REML) estimator, may fail to estimate the smoothing parameter efficiently when observations are overly sparse. This is because we need to simultaneously determine the smoothness of mean curve and individual curves, and the parameters that control these smoothness are usually highly correlated or coupled. On the other hand, most of the above approaches require multiple steps of estimation (e.g. functional principle analysis) and an additional step for model selection which selects the smoothness. The inference, following the parameter estimation, is conditioning on the selected optimal model, with additional assumptions and procedures. The separated steps and additional assumptions may lead to compromised functional estimates and inference. Furthermore, generalizing these methods to more complicated models, such as the additive model, may become problematic. We proposed a Bayesian nonparametric approach which can remedy the above mentioned shortcomings. We model both the mean curve and individual effects by the mixed effects model representation of penalized splines (Ruppert et al. 2003), with the assumption that all the individual curves share the same smoothness. This method can be considered as the Bayesian counterpart of Durban et al. (2005). In this chapter, we show that our proposed Bayesian method can estimate the smoothness parameter more efficiently via simulation studies.

3.2 The Model and Prior Specification

3.2.1 The Model

Suppose we have n individual curves. The response $y_i(t)$ for the i^{th} individual, as a function of t, is assumed to be independent of other individual functions, and can be written as

$$y_i(t) = \mu(t) + b_i(t) + \epsilon_i(t),\ 0 \leq t \leq T,\ i = 1, 2, \ldots, n, \tag{3.1}$$

where $\mu(\cdot)$ is the mean function; $b_i(\cdot)$ is an individual effect function for subject i; and the error function $\epsilon_i(\cdot)$ is assumed a zero mean white-noise process with constant variance σ_ϵ^2 and independent with $\mu(t)$ and $b_i(t)$.

Under O'Sullivan's setting of penalized splines (Wand & Ormerod 2009), a function can be closely approximated by a linear function (parametric part) plus a linear combination of basis functions (nonparametric part):

$$f(x) = \beta_0 + \beta_1 x + \sum u_k Z_k(x)$$

where $\{\beta_0, \beta_1, u_1, \cdots, u_K\}$ are coefficients, $\{Z_1(\cdot), \cdots, Z_K(\cdot)\}$ are given K-dimensional B-splines basis functions spanning over the time range $[0, T]$.

Thus, model (3.1) can be expressed as

$$y_i(t) = \beta_0 + \beta_1 t + \sum_{k=1}^{K} u_k Z_k(t) + a_{i0} + a_{i1} t + \sum_{k=1}^{K} b_{ik} Z_k(t) + \epsilon_i(t). \tag{3.2}$$

where coefficients β_0, β_1 and u_k correspond to the penalized spline for the mean function, and coefficients α_{i0}, α_{i1} and v_{ik} correspond to the penalized splines for the subject-specific functions.

Suppose the ith individual is observed at time points $\boldsymbol{t}_i = [t_{i1}, \cdots, t_{im_i}]^T$, which may vary considerably from one to another, in terms of both the number and distribution of obervations. Using equation (3.2), the observed outcome y_{ij} can be modeled as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} u_k Z_k(t_{ij}) + a_{i0} + a_{i1} t_{ij} + \sum_{k=1}^{K} b_{ik} Z_k(t_{ij}) + \epsilon_{ij}. \tag{3.3}$$

Let

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{bmatrix} \text{ and } \boldsymbol{Z}_i = \begin{bmatrix} Z_1(t_{i1}) & \cdots & Z_K(t_{i1}) \\ \vdots & \ddots & \vdots \\ Z_1(t_{im_i}) & \cdots & Z_K(t_{im_i}) \end{bmatrix},$$

and then (3.3) can be expressed in the following matrix form

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u} + \boldsymbol{X}_i \boldsymbol{a}_i + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i. \tag{3.4}$$

In practice, however, the number and placement of the knots, which determine the B-spline basis, are rarely known. One way to handle this in a semiparametric model is to choose a large number of knots that overfit the model, and shrink the corresponding coefficients by penalization. Ruppert (2002) suggests that the number of knots $K = \min(N/4, 35)$ and the knots are placed at equally spaced quantiles, where N is number of all observed time points among all individuals. Following Ruppert et al. (2003), based on the mixed effects model representation of penalized splines, the coefficients in (3.4) can be modeled as

$$\boldsymbol{u} \sim N(\boldsymbol{0},\ \sigma_u^2 \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{b}_i \sim N(\boldsymbol{0},\ \sigma_v^2 \boldsymbol{I}), \tag{3.5}$$

where $N(\boldsymbol{\mu}, \Sigma)$ denotes a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ. $\mathbf{0}$ is a vector of zeros and $\boldsymbol{I}$ is the identity matrix. Ruppert et al. (2003) shows that the variance parameters σ_u^2 and σ_v^2 control the shrinkage and hence the smoothness of functions. Here, we assume a common variance σ_v^2 for all individual effects so that all individual functions share the same smoothness.

To let individual functions borrow strength from each other, we further assume that

$$\boldsymbol{a}_i \sim N(\mathbf{0},\ \Sigma), \tag{3.6}$$

where $\boldsymbol{a}_i = (a_{i0}, a_{i1})$ is the linear parametric part for each subject specific effect, and Σ is a 2 by 2 positive definite matrix. For convenience, under models (3.5) and (3.6), we rewrite the mixed effects model (3.4) in a hierarchical form:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha}_i + \boldsymbol{Z}_i\boldsymbol{v}_i + \epsilon_i \tag{3.7}$$

where

$$\boldsymbol{\alpha}_i \sim N(\boldsymbol{\beta},\ \Sigma),\ \boldsymbol{v}_i \sim N(\boldsymbol{u},\ \sigma_v^2\boldsymbol{I}),\ i = 1, \cdots, n \quad \text{and} \quad \boldsymbol{u} \sim N(\mathbf{0},\ \sigma_u^2\boldsymbol{I}). \tag{3.8}$$

Under this specification, parameter estimation can be developed using either Bayesian or non-Bayesian techniques. However, due to the fact that the two smoothing parameters σ_u^2 and σ_v^2 are highly coupled, existing frequentist estimators tend to perform poorly when functional data are very sparse.

3.2.2 Prior Specification

Under models (3.7) and (3.8), we need to specify prior distributions for parameters $\{\boldsymbol{\beta}, \Sigma, \sigma_\epsilon, \sigma_u, \sigma_v\}$. Let the prior distribution for $\boldsymbol{\beta}$ be a conjugate normal

$$\boldsymbol{\beta} \sim N(\mathbf{0},\ c\boldsymbol{I}), \tag{3.9}$$

where the multiplier c is a prespecified constant. In our case, we choose $c = 10^5$, presenting a weakly informative prior for $\boldsymbol{\beta}$.

It is usually non-trivial to choose noninformative priors for the variance parameters $\{\sigma_\epsilon, \sigma_u, \sigma_v\}$ in a hierarchical model. Following Gelman et al. (2006), we specify the following priors:

$$\sigma_\epsilon \sim \text{Half-Cauchy}(A_\epsilon),\ \sigma_u \sim \text{Half-Cauchy}(A_u),\ \sigma_v \sim \text{Half-Cauchy}(A_v), \tag{3.10}$$

where the scale parameters $\{A_\epsilon, A_u, A_v\}$ are prespecified as positive constants (in our case we choose 10^5 for each). The following result from (Wand et al. 2011) is useful for sampling from the Half-Cauchy distribution

$$\begin{gathered}\sigma \sim \text{Half-Cauchy}(A) \quad \text{if and only if} \\ \sigma^2 | a \sim \text{IG}\left(\frac{1}{2},\ \frac{1}{a}\right) \quad \text{and} \quad a \sim \text{IG}\left(\frac{1}{2},\ \frac{1}{A^2}\right),\end{gathered} \tag{3.11}$$

where $\text{IG}(A, B)$ denotes the Inverse Gamma distribution, whose density function is given by

$$p(x) = \Gamma(A)^{-1} B^A x^{-A-1} \exp(-B/x)$$

where the shape parameter $A > 0$ and rate parameter $B > 0$.

Similar to the above variance parameters, we specify the prior distribution for the covariance matrix Σ by the hierarchical Inverse Wishart distribution proposed by Huang et al. (2013):

$$\begin{aligned} \Sigma|a_1, a_2 &\sim \mathrm{IW}(\nu+1,\ 2\nu\,\mathrm{diag}\left(\frac{1}{a_1},\ \frac{1}{a_2}\right)) \\ \text{and } a_j &\sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_j^2}\right),\ j = 1 \text{ or } 2, \end{aligned} \tag{3.12}$$

where $\mathrm{IW}(A, \boldsymbol{B})$ denotes the Inverse Wishart distribution with density function given by

$$p(\Sigma) = C_{p,A}^{-1}|\boldsymbol{B}|^{A/2}|\Sigma|^{-(A+p+1)/2}\exp\left\{\frac{1}{2}\,\mathrm{tr}(\boldsymbol{B}\Sigma^{-1})\right\}$$

where p is the dimension of the random matrix Σ, $A > 0$, $\boldsymbol{B}$ is positive definite, and $C_{p,A}^{-1} = 2^{Ap/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{A+1-i}{2}\right)$. Using the Property 4 in Huang et al. (2013), we choose the hyperparameter $\nu = 2$, which implies a uniformly distributed prior for the correlation parameter in Σ.

Combining models (3.5-3.8) and priors (3.9-3.12), we obtain the following hierarchical Bayesian semiparametric model for the sparse functional data:

$$\begin{aligned} &\boldsymbol{y}_i|\boldsymbol{\alpha}_i, \boldsymbol{v}_i, \sigma_\epsilon^2 \sim N(\boldsymbol{X}_i\boldsymbol{\alpha}_i + \boldsymbol{Z}_i\boldsymbol{v}_i,\ \sigma_\epsilon^2\boldsymbol{I}_{m_i}),\ \ 1 \le i \le n \\ &\boldsymbol{\alpha}_i|\boldsymbol{\beta}, \Sigma \sim N(\boldsymbol{\beta},\ \Sigma),\ \ \boldsymbol{v}_i|\boldsymbol{u}, \sigma_v^2 \sim N(\boldsymbol{u},\ \sigma_v^2\boldsymbol{I}_K),\ \ \sigma_\epsilon^2|a_\epsilon \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_\epsilon}\right) \\ &\boldsymbol{\beta} \sim N(\boldsymbol{0},\ c\boldsymbol{I}_K),\ \ \Sigma|a_1, a_2 \sim \mathrm{IW}(\nu+1,\ 2\nu\,\mathrm{diag}\left(\frac{1}{a_1},\ \frac{1}{a_2}\right)) \\ &\boldsymbol{u}|\sigma_u \sim N(\boldsymbol{0},\ \ \sigma_u^2\boldsymbol{I}_K),\ \ \sigma_v^2|a_v \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_v}\right),\ \ \sigma_u^2|a_u \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_u}\right) \\ &a_\epsilon \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_\epsilon^2}\right),\ \ a_v \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_v^2}\right) \\ &a_u \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_u^2}\right),\ \ a_j \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_j^2}\right),\ j = 1, 2 \end{aligned} \tag{3.13}$$

where $c = 10^5$, $A_\epsilon = A_u = A_v = A_1 = A_2 = 10^5$, and $\nu = 2$.

3.3 Bayesian Inference

The inference for the hierarchical Bayesian model specified by (3.13) is done by using Gibbs sampling. We outline the sampling scheme as follows.

We are interested in fitted functionals as posterior means

$$\hat{\boldsymbol{\mu}} = \mathrm{E}(\boldsymbol{X\beta} + \boldsymbol{Zu}|\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{E}(\boldsymbol{\beta}|\boldsymbol{y}) + \boldsymbol{Z}\mathrm{E}(\boldsymbol{u}|\boldsymbol{y})$$

and

$$\hat{\boldsymbol{f}}_i = \mathrm{E}(\boldsymbol{X\alpha}_i + \boldsymbol{Zv}_i|\boldsymbol{y}) = \boldsymbol{X}\mathrm{E}(\boldsymbol{\alpha}_i|\boldsymbol{y}) + \boldsymbol{Z}\mathrm{E}(\boldsymbol{v}_i|\boldsymbol{y}).$$

The posterior distributions of the above needed parameters $\{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\alpha}_i, \boldsymbol{v}_i\}$, as well as scale parameters $\{\sigma_\epsilon, \sigma_u, \sigma_v, \Sigma\}$, are not in closed form expressions. With the model specified in the Section 2.2, let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_n^T]^T$, $\boldsymbol{v} = [\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_n^T]^T$, and $\boldsymbol{C}_i = \begin{bmatrix} \boldsymbol{X}_i & \boldsymbol{Z}_i \end{bmatrix}$, we have the joint posterior distribution

$$\begin{aligned}
& p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{v}, \sigma_\epsilon, \sigma_u, \sigma_v, \Sigma, a_\epsilon, a_u, a_v, a_1, a_2|\boldsymbol{y}) \\
\propto\; & \sigma_\epsilon^{-N} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n \left(\boldsymbol{y}_i - \boldsymbol{C}_i \begin{bmatrix}\boldsymbol{\alpha}_i \\ \boldsymbol{v}_i\end{bmatrix}\right)^T \left(\boldsymbol{y}_i - \boldsymbol{C}_i \begin{bmatrix}\boldsymbol{\alpha}_i \\ \boldsymbol{v}_i\end{bmatrix}\right)\right\} \\
& |\Sigma|^{-n/2}\sigma_v^{-nK} \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\begin{bmatrix}\boldsymbol{\alpha}_i \\ \boldsymbol{v}_i\end{bmatrix} - \begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right)^T \left(\begin{bmatrix}\boldsymbol{\alpha}_i \\ \boldsymbol{v}_i\end{bmatrix} - \begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right)\right\} \\
& \exp\left\{-\frac{1}{2c}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \sigma_u^{-K} \exp\left\{-\frac{1}{2\sigma_u^2}\boldsymbol{u}^T\boldsymbol{u}\right\} a_\epsilon^{-1/2}\sigma_\epsilon^{-3} \exp\left\{-\frac{1}{a_\epsilon\sigma_\epsilon^2}\right\} \\
& a_v^{-1/2}\sigma_v^{-3} \exp\left\{-\frac{1}{a_v\sigma_v^2}\right\} a_u^{-1/2}\sigma_u^{-3} \exp\left\{-\frac{1}{a_u\sigma_u^2}\right\} \\
& |2\nu\,\mathrm{diag}(1/a_1, 1/a_2)|^{(\nu+1)/2}|\Sigma|^{(-\nu+4)/2} \exp\left\{-\frac{1}{2}\,\mathrm{tr}(2\nu\,\mathrm{diag}(1/a_1, 1/a_2)\Sigma^{-1})\right\}
\end{aligned}$$

Fortunately, the *full conditionals* under the current model specifications are shown to have the following well-known distributions:

$$
\begin{aligned}
\begin{bmatrix}\boldsymbol{\alpha}_i \\ \boldsymbol{v}_i\end{bmatrix}\Bigg|\text{rest} &\sim N\left(\left(\sigma_\epsilon^{-2}\boldsymbol{C}_i^T\boldsymbol{C}_i + \begin{bmatrix}\Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_v^{-2}\boldsymbol{I}\end{bmatrix}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{C}_i^T\boldsymbol{y}_i + \begin{bmatrix}\Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_v^{-2}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right),\right. \\
&\qquad\qquad \left.\left(\sigma_\epsilon^{-2}\boldsymbol{C}_i^T\boldsymbol{C}_i + \begin{bmatrix}\Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_v^{-2}\boldsymbol{I}\end{bmatrix}\right)^{-1}\right) \\
\boldsymbol{\beta}|\text{rest} &\sim N\left(\left(n\Sigma^{-1} + c^{-1}\boldsymbol{I}\right)^{-1}\Sigma^{-1}\sum_{i=1}^{n}\boldsymbol{\alpha}_i,\ \left(n\Sigma^{-1} + c^{-1}\boldsymbol{I}\right)^{-1}\right) \\
\boldsymbol{u}|\text{rest} &\sim N\left(\left(n\sigma_v^{-2} + \sigma_u^{-2}\right)^{-1}\sigma_v^{-2}\sum_{i=1}^{n}\boldsymbol{v}_i,\ \left(n\sigma_v^{-2} + \sigma_u^{-2}\right)^{-1}\boldsymbol{I}\right) \\
\sigma_\epsilon^2|\text{rest} &\sim \text{IG}\left(\frac{N+1}{2},\ \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Z}_i\boldsymbol{v}_i)^T(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Z}_i\boldsymbol{v}_i) + \frac{1}{a_\epsilon}\right) \\
\sigma_v^2|\text{rest} &\sim \text{IG}\left(\frac{nK+1}{2},\ \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{v}_i^T\boldsymbol{v}_i + \frac{1}{a_v}\right) \\
\sigma_u^2|\text{rest} &\sim \text{IG}\left(\frac{K+1}{2},\ \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} + \frac{1}{a_u}\right) \\
\Sigma|\text{rest} &\sim \text{IW}\left(n + \nu + 1,\ \sum_{i=1}^{n}\boldsymbol{C}_i\boldsymbol{C}_i^T + 2\nu\begin{bmatrix}1/a_1 & 0 \\ 0 & 1/a_2\end{bmatrix}\right) \\
a_\epsilon|\text{rest} &\sim \text{IG}\left(1, \frac{1}{\sigma_\epsilon^2} + \frac{1}{A_\epsilon^2}\right) \\
a_v|\text{rest} &\sim \text{IG}\left(1, \frac{1}{\sigma_v^2} + \frac{1}{A_v^2}\right) \\
a_u|\text{rest} &\sim \text{IG}\left(1, \frac{1}{\sigma_u^2} + \frac{1}{A_u^2}\right) \\
\text{and } a_j|\text{rest} &\sim \text{IG}\left(\frac{\nu+2}{2}, \Sigma_{jj}^{-1} + \frac{1}{A_j^2}\right),\ j = 1,\ 2.
\end{aligned}
$$

Here, "rest" means the set of other parameters in the model (3.13). Since all full conditional distributions are standard distributions, the sampling is then trivial.

3.4 Simulation Study

In this section, we evaluate and compare the proposed Bayesian mixed effects model with Berk et al. (2012)'s smoothing spline mixed effects model via simulation studies. Consider the following two mean functions

$$\mu_1(t) \quad = \quad t + \sin(t)$$

and

$$\mu_2(t) \quad = \quad 10 + 10\sin(0.7t),\ 0 \leq t \leq 10.$$

These two shapes are chosen to show that our proposed method can fit any shapes from smooth ($\mu_1(\cdot)$) to rough ($\mu_2(\cdot)$). We also consider samples sizes $n = 20, 50$ and 100 for each setting. For each setting, we generate individual functions as follows

$$f_i(t_j) = \mu_l(t_j) - \frac{c_1}{\sqrt{5}}\cos\left(\frac{\pi}{10}t_j\right) + \frac{c_2}{\sqrt{5}}\sin\left(\frac{\pi}{10}t_j\right),$$

where $c_1 \sim N(0,1)$, $c_2 \sim N(0,4)$, $l = 1,2$, $i = 1,\cdots,n$, and $j = 1,\cdots,m_i$. The measurement errors ϵ_{ij} are generated i.i.d. from the normal distribution with mean 0 and variance 0.001. This model will generate functions from the Gaussian process with mean function $\mu_l(t)$, and covariance function combined by two eigenfunctions $-\frac{1}{\sqrt{5}}\cos\left(\frac{\pi}{10}t\right)$ and $\frac{1}{\sqrt{5}}\sin\left(\frac{\pi}{10}t_j\right)$ with corresponding eigenvalues c_1 and c_2. To make the observations sparse, we choose the number of observations m_i for a given individual is distributed as Poisson(5) + 2, representing a mean number of 7 observations with a minimum of 2 for each individual. The positions of observation times, conditional on m_i, is uniformly distributed on the support of the function. For this simulation, we repeatedly draw 100 datasets from the above models.

We compare our method, denoted by BME, with the smoothing spline mixed effects model (Berk et al. 2012), denoted by SME, which can be implemented using the R package *sme* (Berk 2013). The MCMC is run $10,000$ iterations with a burn-in period of $5,000$. Figures (3.1) and (3.2) illustrate the posterior distributions of $\sigma^2_\epsilon, \sigma^2_v, \sigma^2_u$ and mean effect function of BME model for both settings with sample size $n = 100$.

Figures (3.3) and (3.4) illustrate the models fit for a single data set generated from the simulation setting one with sample size $n = 100$. Figure (3.3) shows that the SME model performs poorly on both individual and the mean functions. On the other hand, our proposed model, shown in Figure (3.4), preforms much better in terms of recovering the true function from the noisy data. Similar results can be found from Figures (3.5) and (3.6), which illustrate the model fit for the simulation setting two with $n = 100$.

The two methods are compared on a fine grid by two commonly used criteria in functional data analysis: the mean integrate squared error (MISE) and the point-wise mean squared error (MSE). The Monte Carlo version of these two measures are defined as follows:

$$\text{MSE}_{\text{MC}}(\hat{f}) = \frac{1}{100}\sum_{i=1}^{100}\left[\hat{f}_i(t) - f(t)\right]^2$$

and

$$\text{MISE}_{\text{MC}}(\hat{f}) = \frac{1}{100}\sum_{i=1}^{100}\int\left[\hat{f}_i(t) - f(t)\right]^2 dt,$$

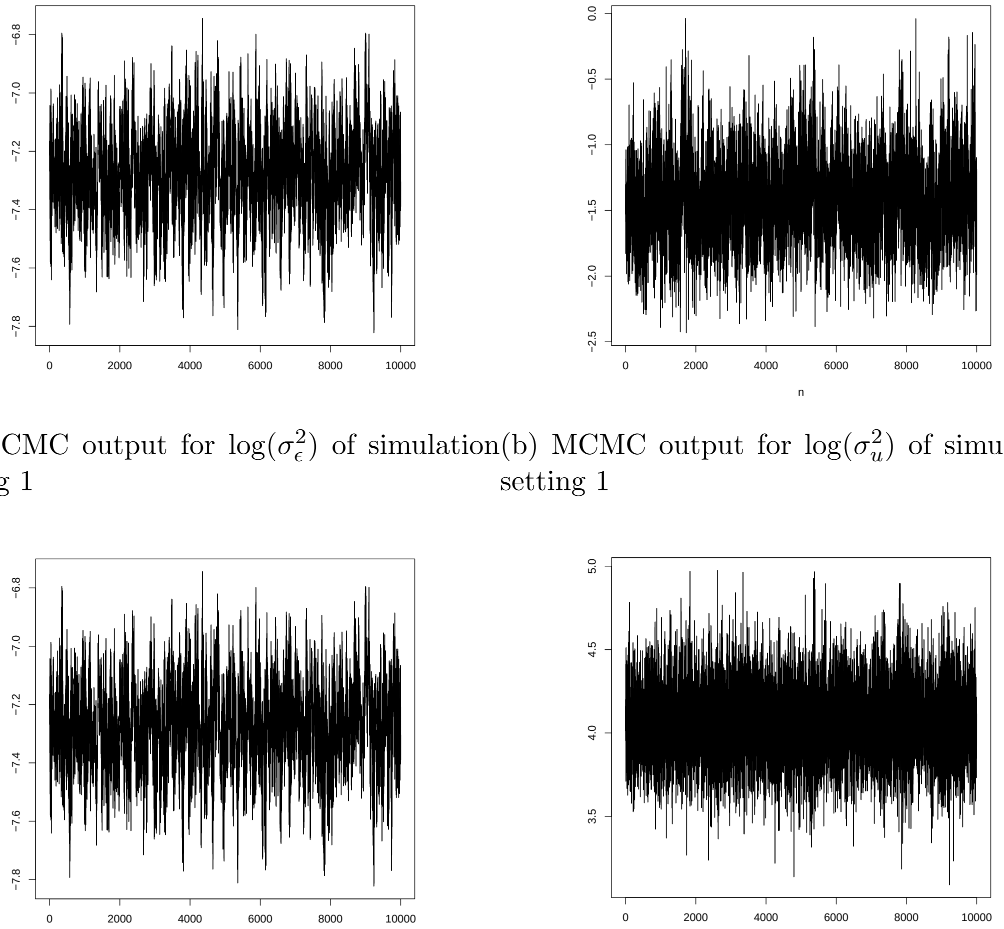
where $\int\left[\hat{f}_i(t) - f(t)\right]^2 dt$ is defined as ISE. The boxplots of the ISE for the mean function and the average ISE for individual functions are shown in Figures (3.7) and (3.8). Table (3.1) shows the MISE. The results for the point-wise MSE of the mean function are shown in Figures (3.9) and (3.10), while those averages for individual functions are shown in Figures (3.11) and (3.12).

From all the figures and tables listed above, we can see that our proposed method performs much better than the smoothing spline mixed effects model. The SME tends to work properly for regularly spaced sparse functional data, though the author claims that it should work for other scenarios as well. The reason that the SME fails to perform satisfactorily could be due to its model selection algorithm. The SME uses the leave-one-observation-out generalized cross validation (or other criteria, like AIC), which pools all the observations together to simultaneously optimize the smoothness for both the mean function and individual functions. For irregularly spaced sparse functional data, perhaps the leave-one-subject-out cross validation is more useful (Xiao et al. 2017). Comparing to the leave-one-observation-out cross validation, the leave-one-subject-out cross validation considers the dependence between observations for a given subject, and is supposed to be more robust against overfit (Reiss et al. 2010).

3.5 **Application**

We now illustrate our proposed method on a publicly available CD4 dataset from the Multicenter AIDs Cohort Study (MACS). The dataset is available in the R package *refund* (Crainiceanu et al. 2012). The CD4 cell is a type of white blood cell that sends signals to the human body to activate the immune response when viruses or bacteria are detected. Usually, the CD4 count is used as an important biomarker for assessing the health of HIV patients as HIV viruses attack or destroy the CD4 cells. The data contain CD4 cell counts for 366 HIV patents between months -18 to 42 since seroconversion (the time HIV becomes detectable) in a longitudinal study (Kaslow et al. 1987). Each individual has a number of observations between 1 to 11, and the total number of observations is $1,888$. Previous statistical analyses for this dataset can be found in Yao et al. (2005), Peng & Paul (2009), Goldsmith et al. (2011), Xiao et al. (2017).

In our analysis, we use the logarithm of the CD4 counts since the counts are skewed. We also remove all the individuals with only 1 count (17 individuals) since these individuals do not provide functional information in the model. Our analysis then includes a total of 349 subjects with 1871 data points on 61 (in months) possible observation times. With a burn-in period of $5,000$, $15,000$ samples are drawn from the posterior distribution using the sampling procedure described in Section 2.3. Figure (3.13) illustrates the fitting for both the mean curve and individual curves. The overall trend in Figure (3.13a) seems to be decreasing, which is consistent to previous findings. Finally, we show in Figures (3.13b-3.13d) the estimated trajectory of log (CD4 count) for three patients.



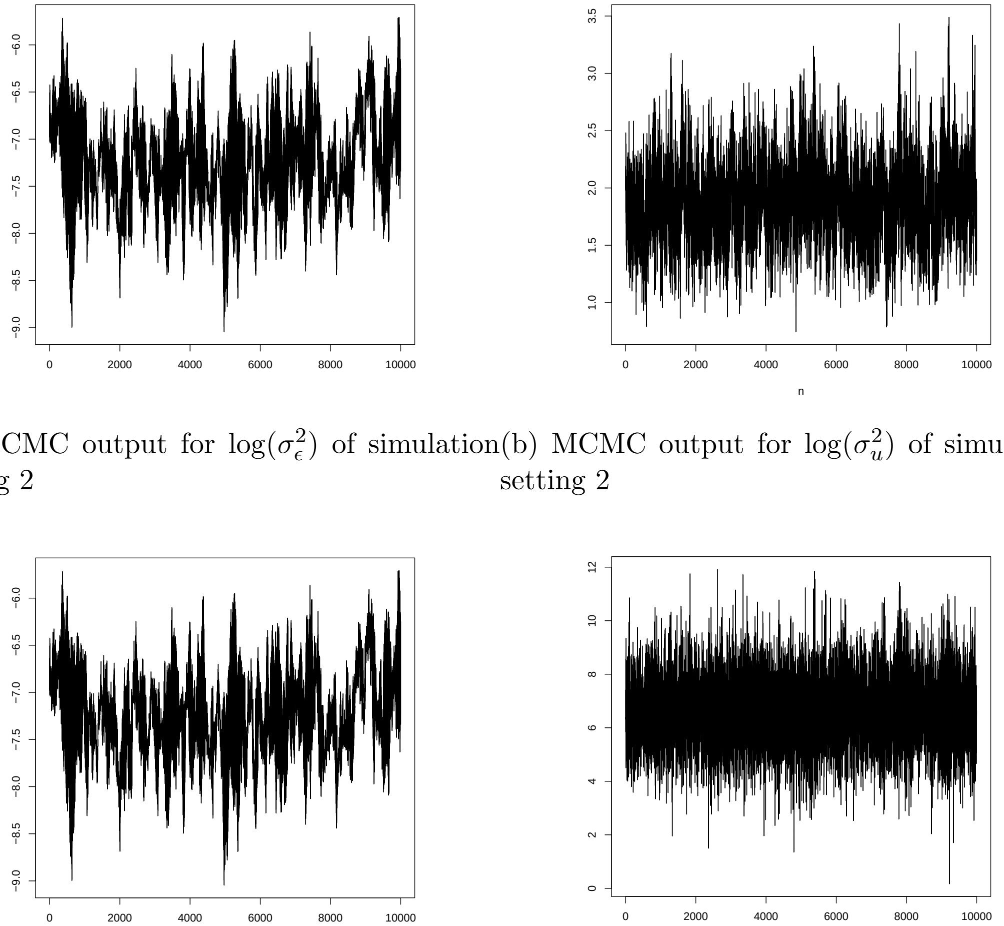
(a) MCMC output for $\log(\sigma_\epsilon^2)$ of simulation setting 1

(b) MCMC output for $\log(\sigma_u^2)$ of simulation setting 1

(c) MCMC output for $\log(\sigma_v^2)$ of simulation setting 1

(d) MCMC output for $\hat{f}$ at median of t_is of simulation setting 1

Figure 3.1: MCMC output for fitting Bayesian functional mixed effects model of simulation setting 1.



(a) MCMC output for $\log(\sigma_\epsilon^2)$ of simulation setting 2

(b) MCMC output for $\log(\sigma_u^2)$ of simulation setting 2

(c) MCMC output for $\log(\sigma_v^2)$ of simulation setting 2

(d) MCMC output for $\hat{f}$ at median of t_is of simulation setting 2

Figure 3.2: MCMC output for fitting Bayesian functional mixed effects model of simulation setting 2.

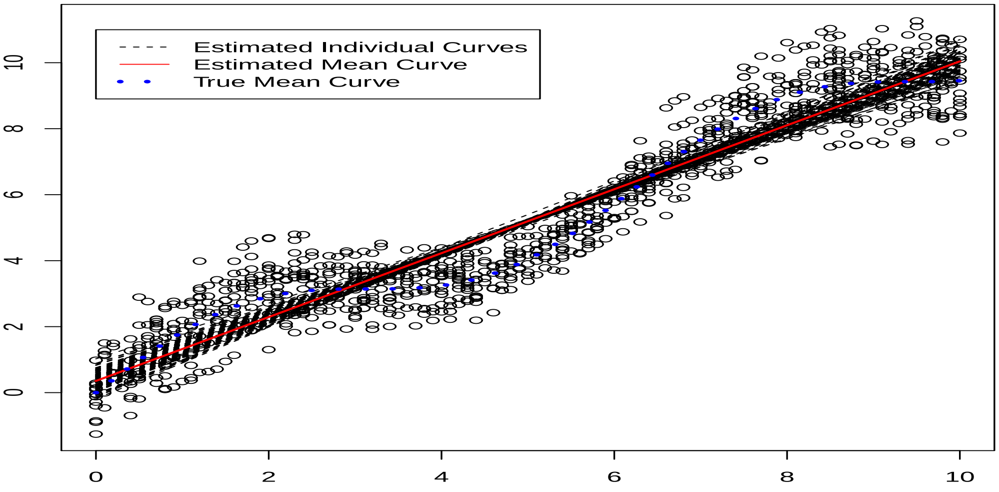


Figure 3.3: Smoothing spline mixed effects model for 100 sparse functional data generated from simulation setting 1.

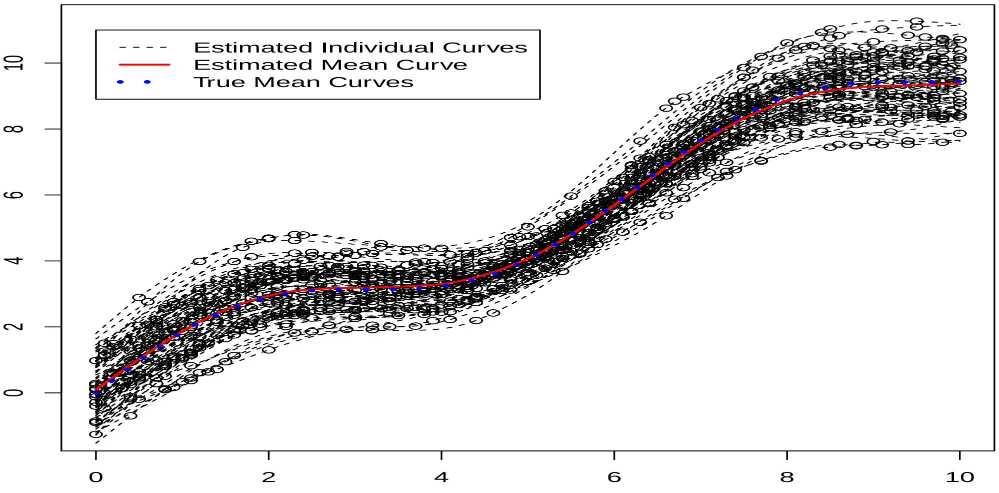


Figure 3.4: Bayesian functional mixed effects model for 100 sparse functional data generated from simulation setting 1.

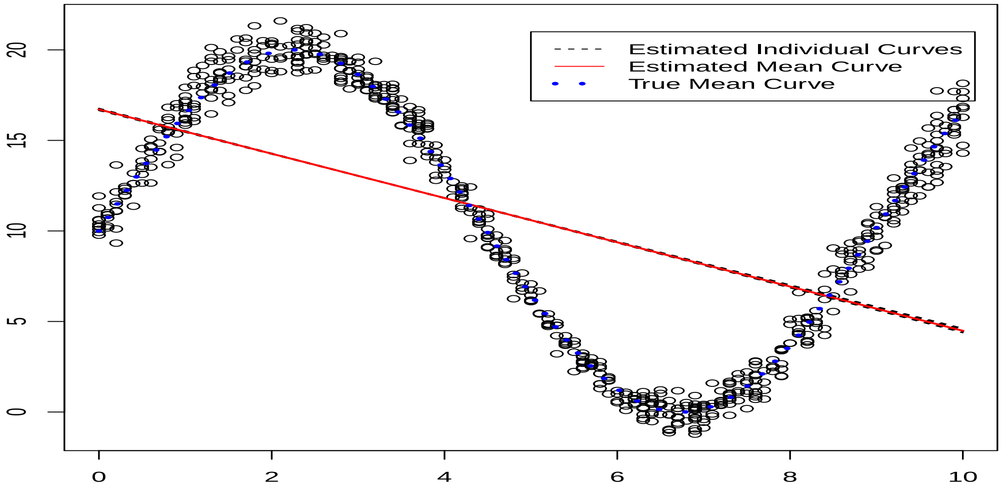


Figure 3.5: Smoothing spline mixed effects model for 100 sparse functional data generated from simulation setting 2.

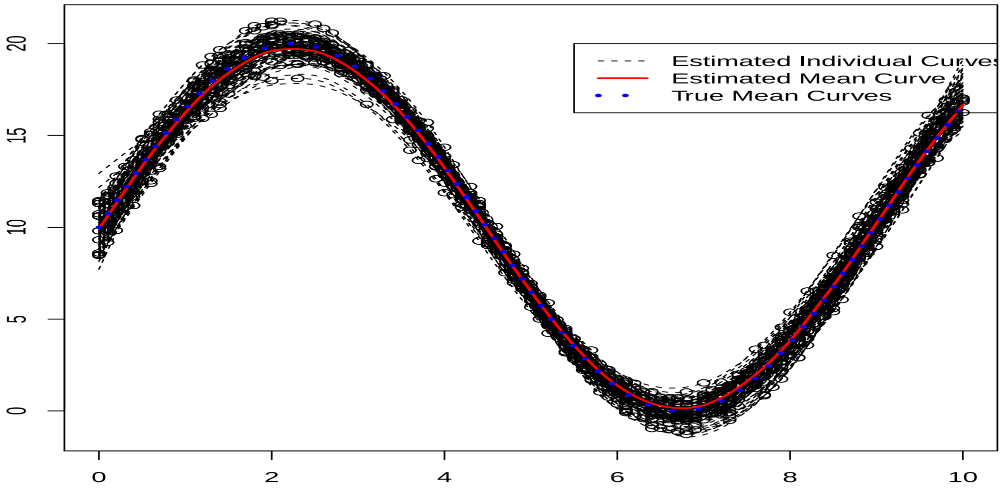
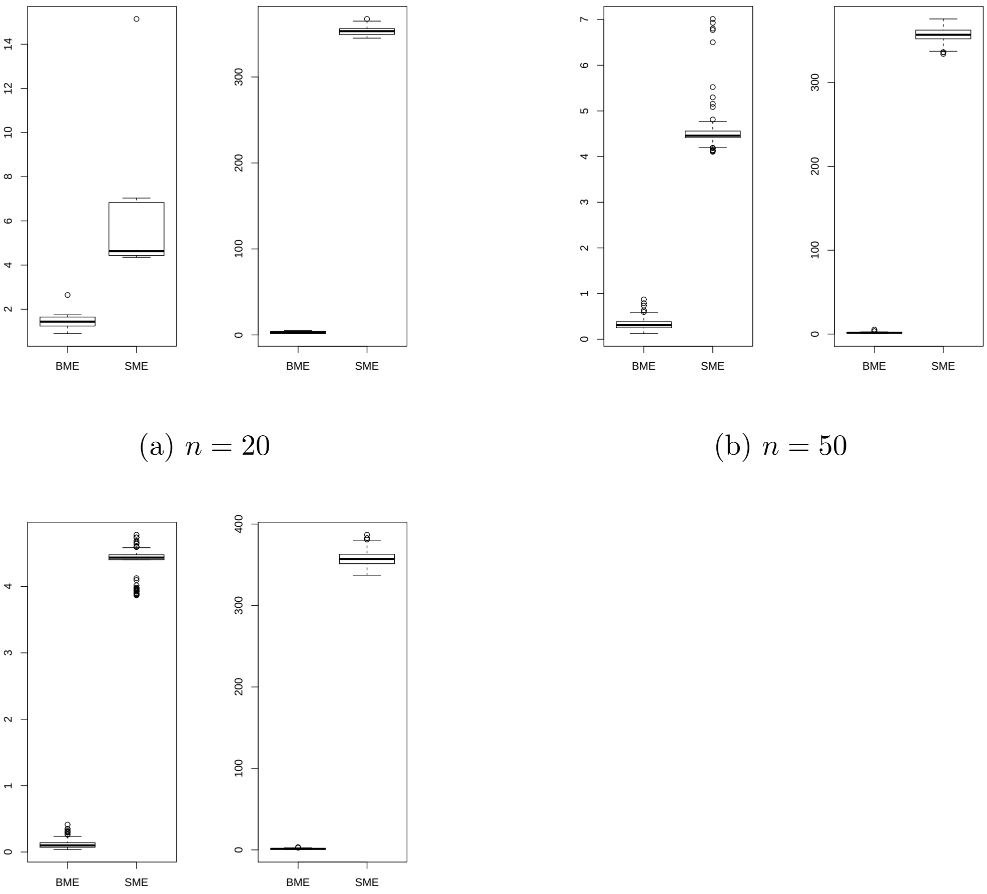


Figure 3.6: Bayesian functional mixed effects model for 100 sparse functional data generated from simulation setting 2.

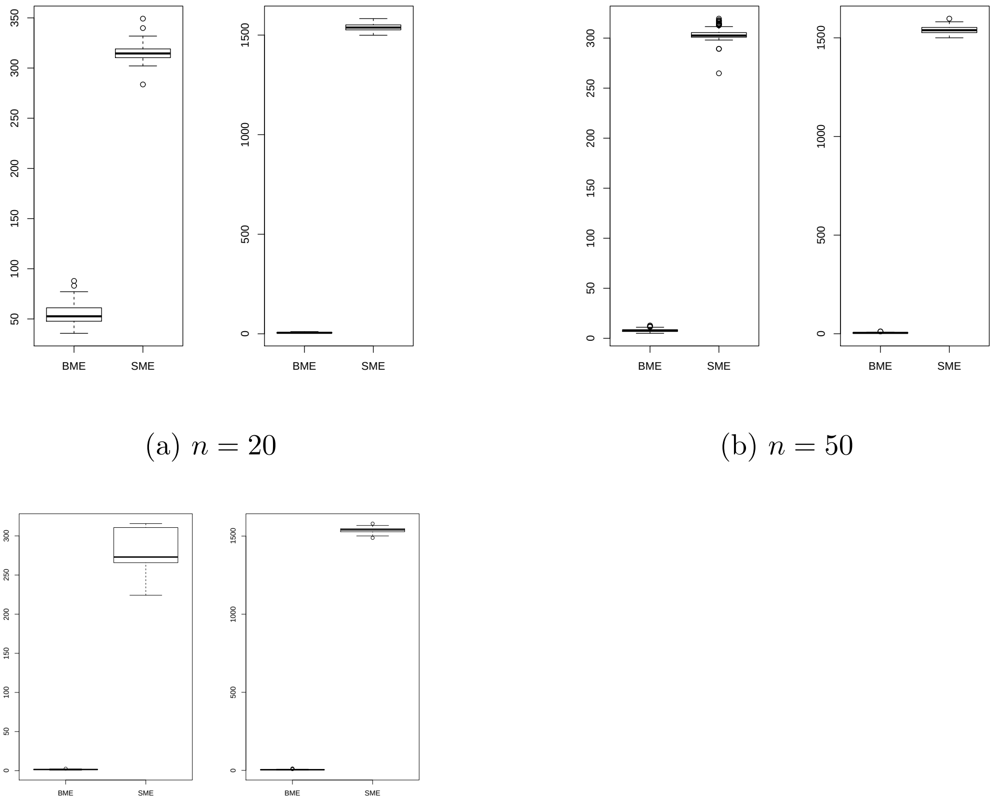


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.7: Boxplots for simulation setting 1 for $n = 20, 50, 100$. The left panel for each figure represents the ISE for the mean effect function; right panel for each figure represents the average ISE for the individual effects functions.

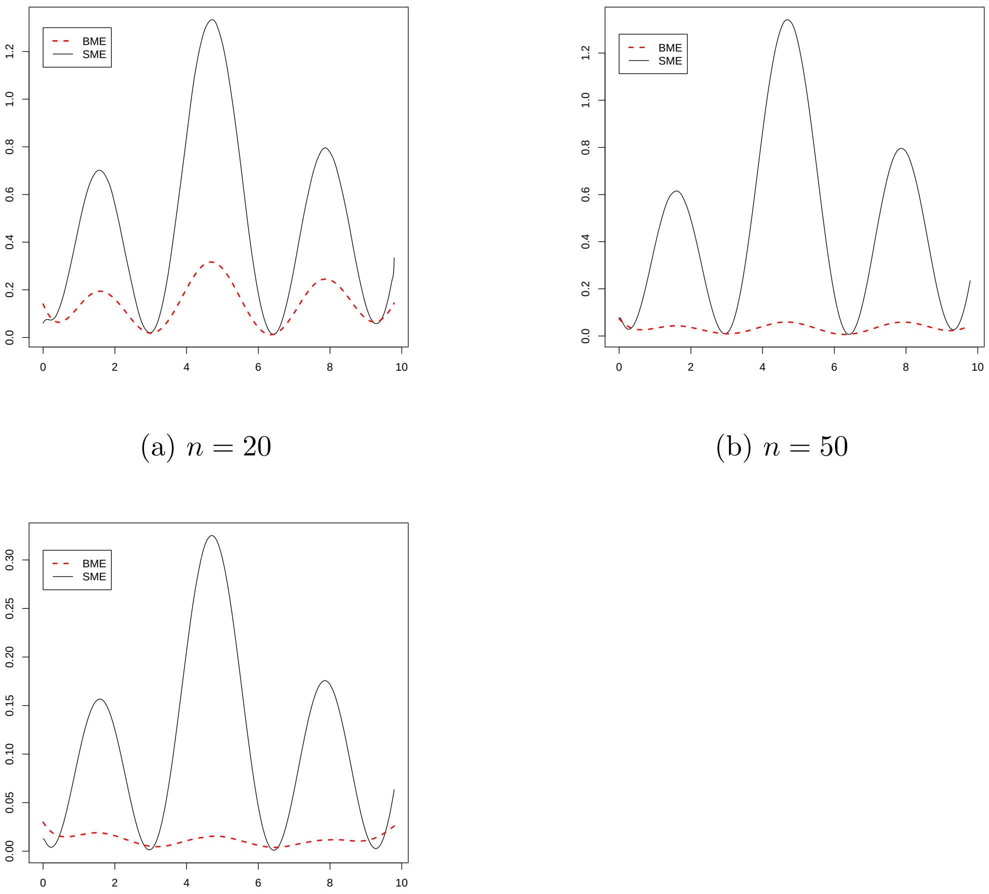


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.8: Boxplots for simulation setting 2 for $n = 20, 50, 100$. The left panel for each figure represents the ISE for the mean effect function; right panel for each figure represents the average ISE for the individual effects functions.

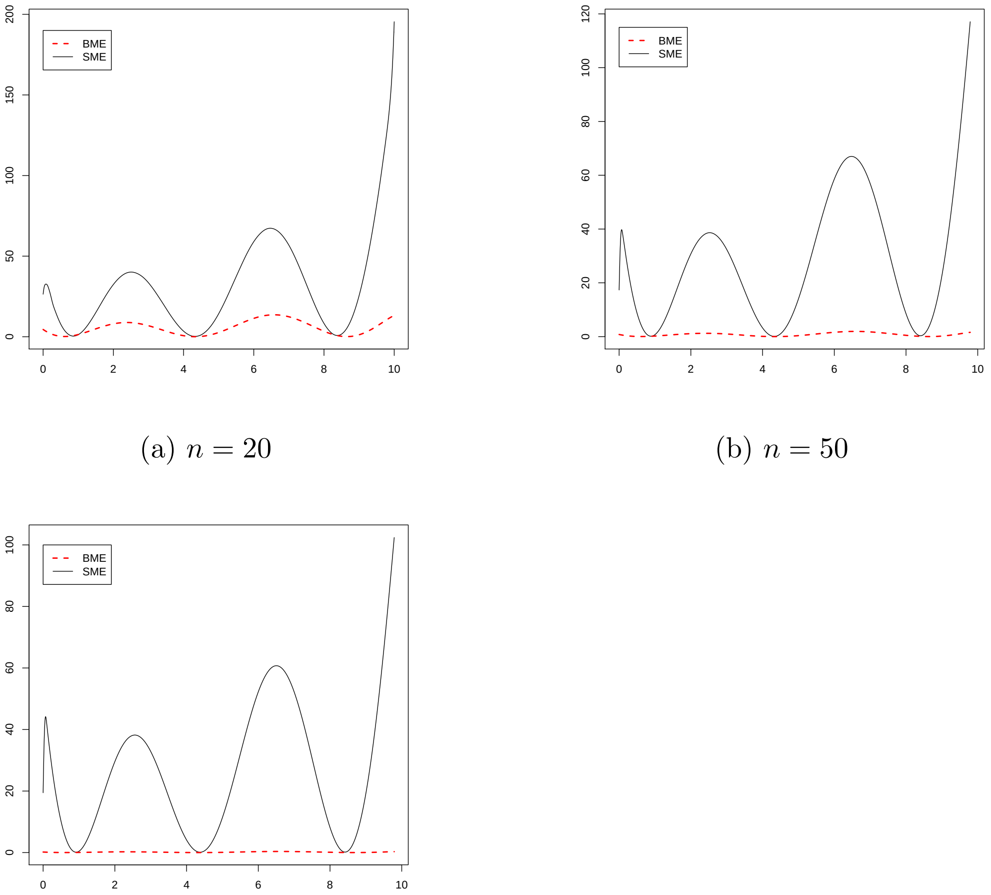


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.9: Pointwise empirical mean square error of the mean effect curves for simulation setting 1 for $n = 20, 50, 100$.

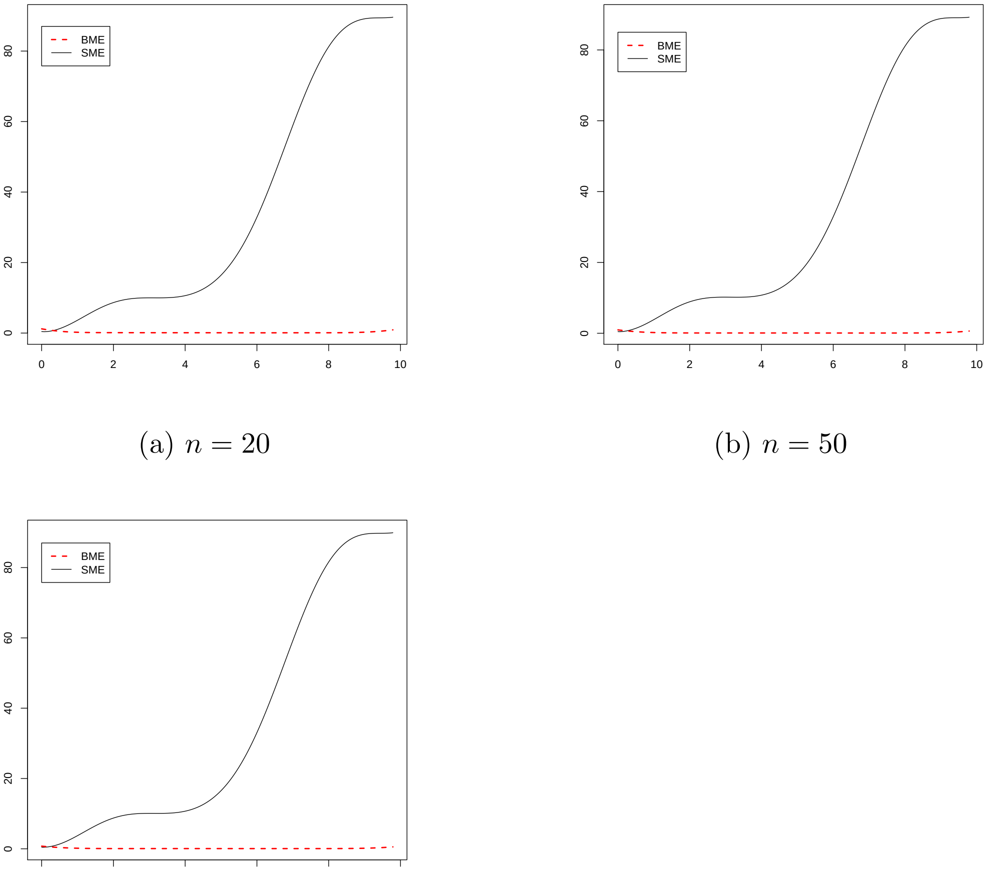


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.10: Pointwise empirical mean square error of the mean effect curves for simulation setting 2 for $n = 20, 50, 100$.

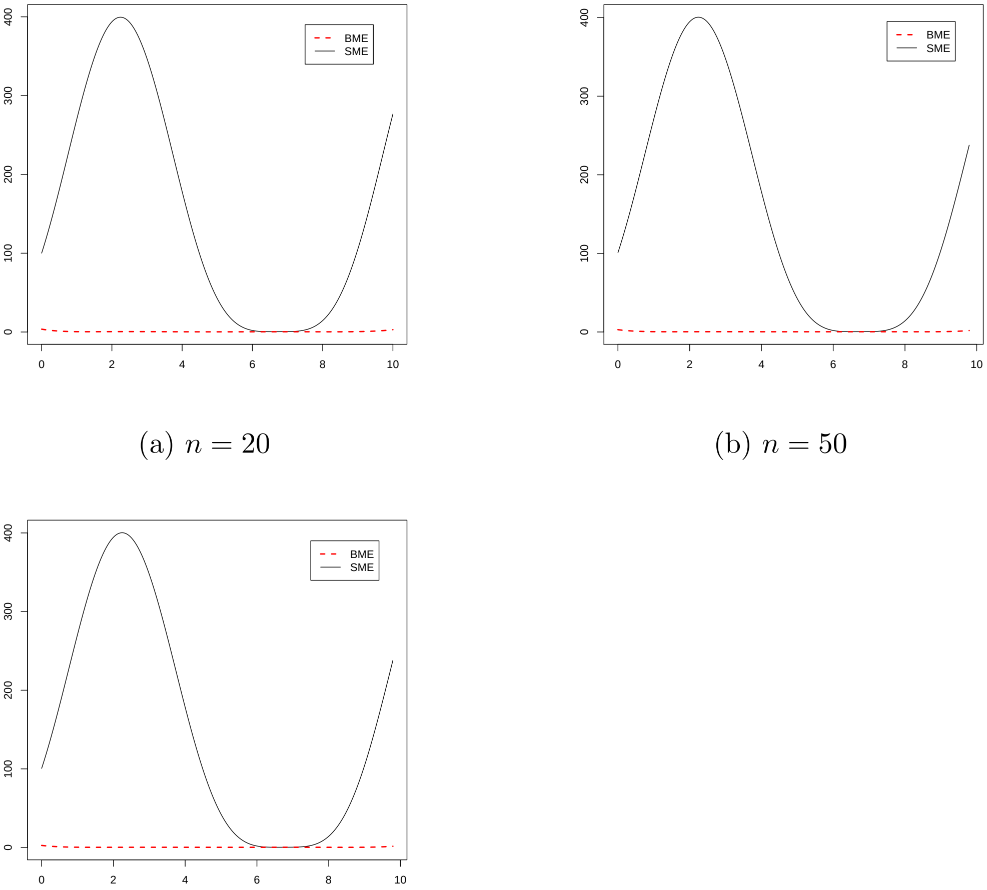


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.11: Pointwise empirical mean square error of the average individual effect curves for simulation setting 1 for $n = 20, 50, 100$.

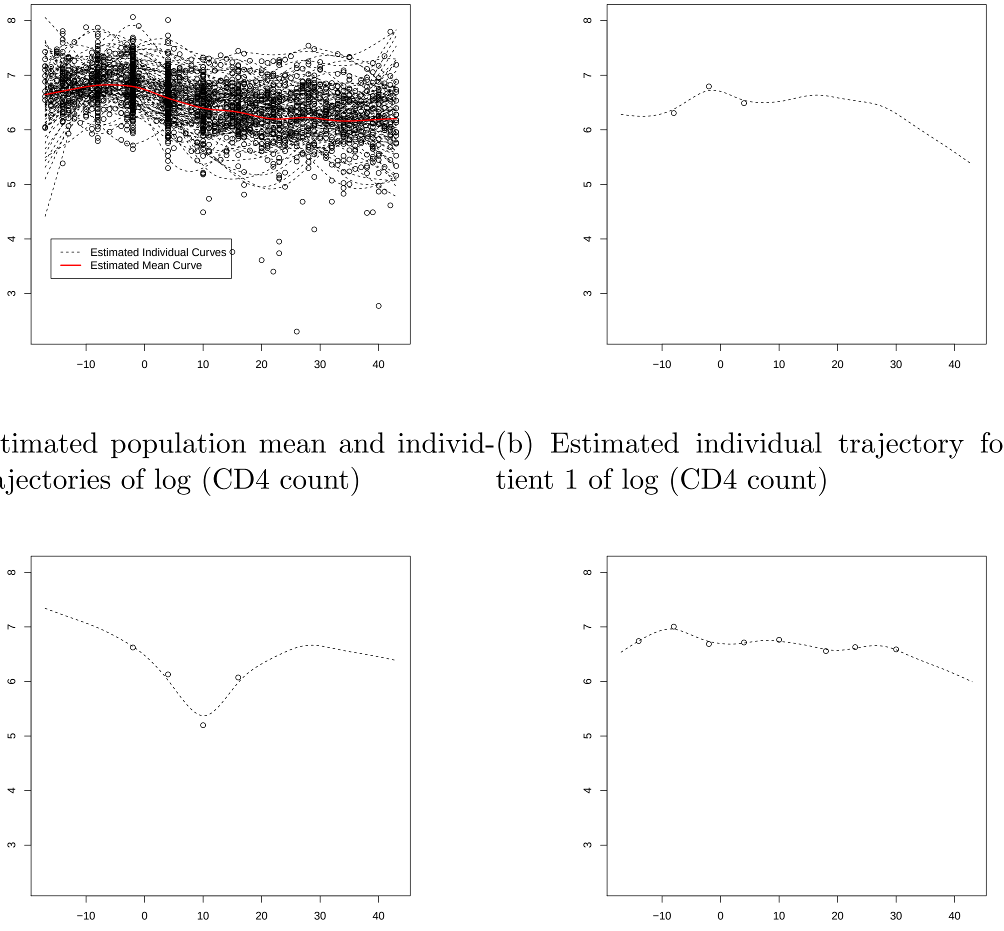


(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 3.12: Pointwise empirical mean square error of the average individual effect curves for simulation setting 2 for $n = 20, 50, 100$.



(a) Estimated population mean and individual trajectories of log (CD4 count)

(b) Estimated individual trajectory for patient 1 of log (CD4 count)

(c) Estimated individual trajectory for patient 2 of log (CD4 count)

(d) Estimated individual trajectory for patient 3 of log (CD4 count)

Figure 3.13: Observed and estimated log (CD4 count) trajectories of 349 HIV-infected patients.

Table 3.1: Monte Carlo approximation of MISE for two simulation settings. The replication size is 100. Two estimators, Bayesian mixed effects model (BME) and smoothing spline mixed effects model (SME), are compared by their mean effect function and average individual effect functions.

Mean effect function				
	Simulation 1		Simulation 2	
n	BME	SME	BME	SME
20	**1.4997**	6.0779	**2.7629**	355.52
50	**0.3414**	4.5999	**1.6401**	356.89
100	**0.1224**	4.3727	**1.3343**	357.91

Average individual effect functions				
	Simulation 1		Simulation 2	
n	BME	SME	BME	SME
20	**54.6357**	315.360	**7.8868**	1538.3
50	**7.8868**	303.826	**4.6889**	1539.1
100	**1.5427**	280.026	**4.389**	1538.2

CHAPTER IV

Bayesian Functional Mixed-effects Model with Grouped Smoothness

4.1 **Introduction**

The proposed Bayesian method in the previous chapter unifies the model selection and parameter estimation into a single framework, which overcomes the before mentioned disadvantages of existing nonparametric or semiparametric approaches for the sparse functional data. The simulation results have shown that our proposed method significantly outperforms the smoothing spline mixed effects model by Berk et al. (2012).

Our previously proposed model, among all other functional mixed effects models (James et al. 2000, Thompson & Rosen 2008, Wu & Zhang 2006), assumes a common underlying smoothness for all individual trajectories, which in general, may be a quite limited assumption. However, when functional data are sparse, it is not applicable to assume different smoothing parameters for individual curves separately, because for many individuals, only a small number of observations are available. Whether to assume a common smoothing parameter or different smoothing parameters is essentially a bias-and-variance trade-off. When the sample size (number of subjects) is small, pooling all observations together can almost always improve the estimation. However, as the sample size continues to increase to a relatively large number, the estimation performance will not necessarily be improved. This is because when the large number of functions are generated from some underlying Gaussian process, estimation of the assumed common smoothness can be affected substantially by outliers. The gain of a reduced variance due to a large sample size may be even smaller than the loss caused by an increased bias due to conditioning on a too general population. A potential strategy is to consider a model that can further stratify the underlying population.

In this chapter, we propose a new Bayesian functional mixed effects model which assumes two groups of functions with different smoothing parameters. We employ latent indicators to determine whether a given function belongs to a more jagged function group. We fit the proposed model through Markov Chain Monte Carlo (MCMC) methods, specifically the Metropolis within Gibbs type algorithm (Chib & Greenberg 1995). The general setting of our new model is similar to the semiparametric mixed effects model described in the previous chapter. Instead of specifying a single variance component σ_v^2 across all individual functions, we let this parameter equal to $\sigma_{v_1}^2 + \gamma_i \sigma_{v_2}^2$ for individual i, where γ_i is a latent indicator.

4.2 The Model and Prior Specification

4.2.1 The Model

Suppose we have n individual curves. As in the previous chapter, the response $y_i(t)$ for the i^{th} individual is written as

$$y_i(t) = \mu(t) + b_i(t) + \epsilon_i(t),\ 0 \leq t \leq T,\ i = 1, \ldots, n, \tag{4.1}$$

where $\mu(\cdot)$ is the mean function and $b_i(\cdot)$ is the individual effect function for subject i. Here, we assume $\epsilon_i(\cdot)$ a white-noise process with a constant variance σ_ϵ^2 and independent of $\mu(t)$ and $b_i(t)$.

Under the O'Sullivan penalized splines (Wand & Ormerod 2009), model (4.1) can be expressed as

$$y_i(t) = \beta_0 + \beta_1 t + \sum_{k=1}^{K} u_k Z_k(t) + a_{i0} + a_{i1} t + \sum_{k=1}^{K} b_{ik} Z_k(t) + \epsilon_i(t), \tag{4.2}$$

where $\{Z_1(\cdot), \cdots, Z_K(\cdot)\}$ is a given set of K-dimensional B-splines basis functions spanning over $[0, T]$, β_0, β_1 and u_k are coefficients for the mean functions, and α_{i0}, α_{i1} and v_{ik} are coefficients for individual effect functions.

With realizations on time points $\boldsymbol{t}_i = [t_{i1}, \cdots, t_{im_i}]^T$, each y_{ij} is modeled as

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} u_k Z_k(t_{ij}) + a_{i0} + a_{i1} t_{ij} + \sum_{k=1}^{K} b_{ik} Z_k(t_{ij}) + \epsilon_{ij}. \tag{4.3}$$

Let

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{bmatrix} \text{ and } \boldsymbol{Z}_i = \begin{bmatrix} Z_1(t_{i1}) & \cdots & Z_K(t_{i1}) \\ \vdots & \ddots & \vdots \\ Z_1(t_{im_i}) & \cdots & Z_K(t_{im_i}) \end{bmatrix},$$

then (4.3) can be rewritten in a matrix form as follows

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u} + \boldsymbol{X}_i \boldsymbol{a}_i + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i. \tag{4.4}$$

The shrinkage is controlled by assigning the following priors on the coefficients:

$$\boldsymbol{u} \sim N(\boldsymbol{0},\ \sigma_u^2 \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{b}_i \sim N(\boldsymbol{0},\ \sigma_{v_1}^2 + \gamma_i \sigma_{v_2}^2 \boldsymbol{I}). \tag{4.5}$$

where σ_u, $\sigma_{v_1}^2$ and $\sigma_{v_2}^2$ are variance components, and γ_i is a binary indicator. Apparently, instead of assuming a common smoothing parameter σ_v^2, we consider a mixture smoothing parameter for $\boldsymbol{b}_i$ conditioning on the latent indicator γ_i. The rationale is that we roughly stratify the individual curves into two groups, one of which is more jagged than the other. When $\gamma_i = 0$, the smoothing parameter equals to σ_{v_1}, for which the curve is said to have a “standard” smoothness. On the other hand, when $\gamma_i = 1$, the smoothing parameter equals to $\sigma_{v_1} + \sigma_{v_2}$, for which the curve is said to have a more jagged smoothness. To borrow information from each individual, we again model the linear parametric part of the individual effects as

$$\boldsymbol{a}_i \sim N(\mathbf{0},\ \Sigma), \tag{4.6}$$

where Σ is a 2 by 2 positive definite matrix. We rewrite the mixed effects model (4.4) in a hierarchical form:

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\alpha}_i + \boldsymbol{Z}_i \boldsymbol{v}_i + \epsilon_i \tag{4.7}$$

where

$$\boldsymbol{\alpha}_i \sim N(\boldsymbol{\beta},\ \Sigma),\ \boldsymbol{v}_i \sim N(\boldsymbol{u},\ (\sigma_{v_1}^2 + \gamma_i \sigma_{v_2}^2)\boldsymbol{I}),\ i = 1, \cdots, n. \tag{4.8}$$

For this model, we will need to estimate three variance components σ_u^2, $\sigma_{v_1}^2$, and $\sigma_{v_2}^2$, along with the latent indicators γ_is.

4.2.2 Prior Specification

Under the model (4.7), (4.5) and (4.8), we need to specify priors for the parameters $\{\boldsymbol{\beta}, \Sigma, \sigma_\epsilon, \sigma_u, \sigma_{v_1}, \sigma_{v_2}\}$, and latent indicators $\{\gamma_i\}$ for $i = 1, \cdots, n$. Most prior specifications are the same as in the previous chapter. The details are given as follows.

The prior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim N(\mathbf{0},\ c\boldsymbol{I}), \tag{4.9}$$

where the multiplier c is a prespecified constant. In our setting, we choose $c = 10^5$.

For the variance components $\{\sigma_\epsilon, \sigma_u, \sigma_{v_1}, \sigma_{v_2}\}$, we specify Half-Cauchy priors:

$$\begin{aligned}\sigma_\epsilon &\sim \text{Half-Cauchy}(A_\epsilon),\\ \sigma_u &\sim \text{Half-Cauchy}(A_u),\\ \sigma_{v_j} &\sim \text{Half-Cauchy}(A_v),\ j = 1, 2,\end{aligned} \tag{4.10}$$

where the scale parameters $\{A_\epsilon, A_u, A_v\}$ are specified as positive constants (10^5 in our setting).

We specify the hierarchical Inverse Wishart distribution (Huang et al. 2013) as a prior for the covariance matrix Σ:

$$\begin{aligned}\Sigma | a_1, a_2 &\sim \text{IW}(\nu + 1,\ 2\nu \,\text{diag}\left(\frac{1}{a_1},\ \frac{1}{a_2}\right))\\ \text{and } a_j &\sim \text{IG}\left(\frac{1}{2},\ \frac{1}{A_j^2}\right),\ j = 1 \text{ or } 2.\end{aligned} \tag{4.11}$$

We choose the hyperparameter $\nu = 2$, which implies a uniformly distributed prior for the correlation.

Finally, for the latent indicators $\{\gamma_i\}$, $i = 1, \cdots, n$, assume independent Bernoulli distributions with parameter π, that is

$$\gamma_i | \pi \sim \text{Bin}(1, \pi), \tag{4.12}$$

and assume the hyperprior for π a beta distribution

$$\pi \sim \text{Be}(A_\pi, B_\pi). \tag{4.13}$$

We choose hyperparameters A_π and B_π both equal to 1.

Combining models (4.5-4.8) and priors (4.9-4.13), we obtain the following hierarchical Bayesian semiparametric model:

$$\begin{aligned}
&\boldsymbol{y}_i|\boldsymbol{\alpha}_i, \boldsymbol{v}_i, \sigma_\epsilon^2 \sim N(\boldsymbol{X}_i\boldsymbol{\alpha}_i + \boldsymbol{Z}_i\boldsymbol{v}_i,\ \sigma_\epsilon^2\boldsymbol{I}_{m_i}),\quad 1 \le i \le n\\
&\boldsymbol{\alpha}_i|\boldsymbol{\beta}, \Sigma \sim N(\boldsymbol{\beta},\ \Sigma),\quad \boldsymbol{v}_i|\boldsymbol{u}, \sigma_{v_1}^2, \sigma_{v_2}^2, \gamma_i \sim N(\boldsymbol{u},\ (\sigma_{v_1}^2 + \gamma_i\sigma_{v_2}^2)\boldsymbol{I}_K),\quad \sigma_\epsilon^2|a_\epsilon \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_\epsilon}\right)\\
&\boldsymbol{\beta} \sim N(\boldsymbol{0},\ c\boldsymbol{I}_K),\quad \Sigma|a_1, a_2 \sim \mathrm{IW}(\nu+1,\ 2\nu\,\mathrm{diag}\left(\frac{1}{a_1},\ \frac{1}{a_2}\right))\\
&\boldsymbol{u}|\sigma_u \sim N(\boldsymbol{0},\ \ \sigma_u^2\boldsymbol{I}_K),\quad \gamma_i|\pi \sim \mathrm{Bin}(1,\pi),\quad 1 \le i \le n \qquad (4.14)\\
&\sigma_{v_1}^2|a_{v_1} \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_{v_1}}\right),\quad \sigma_{v_2}^2|a_{v_2} \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_{v_2}}\right),\quad \sigma_u^2|a_u \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{a_u}\right)\\
&\pi \sim \mathrm{Be}(A_\pi, B_\pi),\quad a_\epsilon \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_\epsilon^2}\right),\quad a_u \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_v^2}\right)\\
&a_{v_1} \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_{v_1}^2}\right),\quad a_{v_2} \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_{v_2}^2}\right),\quad a_j \sim \mathrm{IG}\left(\frac{1}{2},\ \frac{1}{A_j^2}\right),\ j = 1, 2
\end{aligned}$$

where $c = 10^5$, $A_\epsilon = A_u = A_{v_1} = A_{v_2} = A_1 = A_2 = 10^5$, $A_\pi = B_\pi = 1$, and $\nu = 2$.

4.3 Bayesian Inference

The hierarchical Bayesian model specified by (4.14) can be fitted using MCMC methods to sample from the posterior distribution. However, unlike the Gibbs sampler we specified in the previous chapter, the full conditional distribution of $\{\sigma_{v_1}^2, \sigma_{v_2}^2\}$ is not standard one. Here, we will sample the block $\{\sigma_{v_1}^2, \sigma_{v_2}^2\}$ with Metropolis-Hastings algorithm, leading to a Metropolis within Gibbs type algorithm. Details of the sampling scheme are given below.

We are interested in fitted functionals as posterior means

$$\hat{\boldsymbol{\mu}} = \mathrm{E}(\boldsymbol{X\beta} + \boldsymbol{Zu}|\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{E}(\boldsymbol{\beta}|\boldsymbol{y}) + \boldsymbol{Z}\mathrm{E}(\boldsymbol{u}|\boldsymbol{y})$$

and

$$\hat{\boldsymbol{f}}_i = \mathrm{E}(\boldsymbol{X}\boldsymbol{\alpha}_i + \boldsymbol{Z}\boldsymbol{v}_i|\boldsymbol{y}) = \boldsymbol{X}\mathrm{E}(\boldsymbol{\alpha}_i|\boldsymbol{y}) + \boldsymbol{Z}\mathrm{E}(\boldsymbol{v}_i|\boldsymbol{y}).$$

The posterior distributions of coefficients $\{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\alpha}_i, \boldsymbol{v}_i\}$, scale parameters $\{\sigma_\epsilon, \sigma_u, \sigma_{v_1}, \sigma_{v_2}, \Sigma\}$ and latent indicators $\{\gamma_i\}$, are not in closed form. Denote $\boldsymbol{C}_i = \begin{bmatrix} \boldsymbol{X}_i & \boldsymbol{Z}_i \end{bmatrix}$, $\sigma^2_{v_i} = \sigma^2_{v_1} + \gamma_i \sigma^2_{v_2}$, and $\gamma. = \sum_{i=1}^n \gamma_i$, the full conditionals of $\{\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\alpha}_i, \boldsymbol{v}_i, \sigma_\epsilon, \sigma_u, \Sigma, \gamma_i\}$, $i = 1, \cdots, n$, are shown to have the following distributions

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{v}_i \end{bmatrix} \Bigg| \text{rest} \quad &\sim \quad N\left(\left(\sigma_\epsilon^{-2} \boldsymbol{C}_i^T \boldsymbol{C}_i + \begin{bmatrix} \Sigma^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{v_i}^{-2}\boldsymbol{I} \end{bmatrix} \right)^{-1} \left(\sigma_\epsilon^{-2} \boldsymbol{C}_i^T \boldsymbol{y}_i + \begin{bmatrix} \Sigma^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{v_i}^{-2}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right), \right. \\
&\qquad \left. \left(\sigma_\epsilon^{-2} \boldsymbol{C}_i^T \boldsymbol{C}_i + \begin{bmatrix} \Sigma^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{v_i}^{-2}\boldsymbol{I} \end{bmatrix} \right)^{-1} \right) \\
\boldsymbol{\beta}|\text{rest} \quad &\sim \quad N\left(\left(n\Sigma^{-1} + c^{-1}\boldsymbol{I}\right)^{-1} \Sigma^{-1} \sum_{i=1}^n \boldsymbol{\alpha}_i, \ \left(n\Sigma^{-1} + c^{-1}\boldsymbol{I}\right)^{-1} \right) \\
\boldsymbol{u}|\text{rest} \quad &\sim \quad N\left(\left(\gamma.\sigma_{v_1}^{-2} + (n-\gamma.)(\sigma^2_{v_1} + \sigma^2_{v_2})^{-1} + \sigma_u^{-2}\right)^{-1} \sum_{i=1}^n \sigma_{v_i}^{-2} \boldsymbol{v}_i \ , \right. \\
&\qquad \left. \left(\gamma.\sigma_{v_1}^{-2} + (n-\gamma.)(\sigma^2_{v_1} + \sigma^2_{v_2})^{-1} + \sigma_u^{-2}\right)^{-1} \boldsymbol{I} \right) \\
\sigma^2_\epsilon|\text{rest} \quad &\sim \quad \mathrm{IG}\left(\frac{N+1}{2}, \ \frac{1}{2} \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Z}_i\boldsymbol{v}_i)^T (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Z}_i\boldsymbol{v}_i) + \frac{1}{a_\epsilon} \right) \\
\sigma^2_u|\text{rest} \quad &\sim \quad \mathrm{IG}\left(\frac{K+1}{2}, \ \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} + \frac{1}{a_u} \right) \\
\Sigma|\text{rest} \quad &\sim \quad \mathrm{IW}\left(n + \nu + 1, \ \sum_{i=1}^n \boldsymbol{C}_i \boldsymbol{C}_i^T + 2\nu \begin{bmatrix} 1/a_1 & 0 \\ 0 & 1/a_2 \end{bmatrix} \right) \\
a_\epsilon|\text{rest} \quad &\sim \quad \mathrm{IG}\left(1, \frac{1}{\sigma^2_\epsilon} + \frac{1}{A^2_\epsilon} \right) \\
a_{v_1}|\text{rest} \quad &\sim \quad \mathrm{IG}\left(1, \frac{1}{\sigma^2_{v_1}} + \frac{1}{A^2_{v_1}} \right) \\
a_{v_2}|\text{rest} \quad &\sim \quad \mathrm{IG}\left(1, \frac{1}{\sigma^2_{v_2}} + \frac{1}{A^2_{v_2}} \right)
\end{aligned}$$

$$
\begin{aligned}
a_u|\text{rest} &\sim \text{IG}\left(1, \frac{1}{\sigma_u^2} + \frac{1}{A_u^2}\right) \\
a_j|\text{rest} &\sim \text{IG}\left(\frac{\nu+2}{2}, \Sigma_{jj}^{-1} + \frac{1}{A_j^2}\right),\ j = 1,\ 2 \\
\gamma_i|\text{rest} &\sim \text{Bin}\left(1, \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\
\text{and } \pi|\text{rest} &\sim \text{Be}(\gamma. + A_\pi, n + B_\pi - \gamma.),
\end{aligned}
$$

where $\eta_i = -\frac{k}{2}(\log(\sigma_{v_1}^2 + \sigma_{v_2}^2) - \log(\sigma_{v_1}^2)) + \frac{1}{2}(\frac{1}{\sigma_{v_1}^2} - \frac{1}{\sigma_{v_1}^2+\sigma_{v_2}^2})v_i^T v_i + \text{logit}(\pi)$.

However, direct sampling from the full conditional distribution of $\{\sigma_{v_1}^2, \sigma_{v_2}^2\}$ is not applicable. We employ here the random direction Adaptive Rejection Metropolis Sampling algorithm introduced by Petris & Tardella (2003) to draw samples from their full conditional distributions separately.

4.4 Simulation Study

In this section, we evaluate the proposed Bayesian mixed effects model with grouping by a simulation study. Consider the mean function specified as

$$\mu(t) \quad = \quad t + \sin(t)$$

with sample sizes $n = 50, 100$. For each individual i, the function is randomly generated from either of the two types of random functions with probability 0.5. The first type of random functions are

$$f_i(t_j) = \mu_l(t_j) - \frac{c_1}{\sqrt{5}}\cos\left(\frac{\pi}{10}t_j\right) + \frac{c_2}{\sqrt{5}}\sin\left(\frac{\pi}{10}t_j\right),$$

The second type of random functions are generated from

$$f_i(t_j) = \mu_l(t_j) - \frac{c_1}{\sqrt{5}}\cos\left(10\pi t_j\right) + \frac{c_2}{\sqrt{5}}\sin\left(10\pi t_j\right),$$

which are more jagged than the first type. The coefficients are distributed as $c_1 \sim N(0,1)$, $c_2 \sim N(0,4)$, $l = 1,2$, $i = 1,\cdots,n$, and $j = 1,\cdots,m_i$. The measurement errors ϵ_{ij} are generated i.i.d. from a normal distribution with mean 0 and variance 0.001. The number of observations m_i for a given individual is distributed as Poisson$(5)+2$. The locations of observation times, conditional on m_i, is uniformly distributed on the support of the function. The replication size is 20.

Our newly proposed Bayesian functional mixed effects model (denoted by BME2) is then compared with the model we proposed in the previous chapter (BME). We evaluate the methods using the following two criteria: the mean integrate squared error (MISE) and the point-wise mean squared error (MSE). The Monte Carlo version of these two measures are defined as follows:

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{f}) = \frac{1}{100}\sum_{i=1}^{100}\left[\hat{f}_i(t) - f(t)\right]^2$$

and

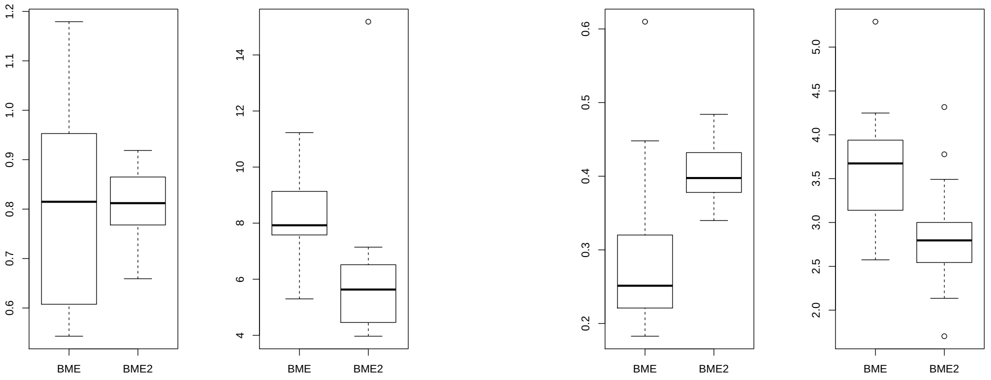
$$\mathrm{MISE}_{\mathrm{MC}}(\hat{f}) = \frac{1}{100}\sum_{i=1}^{100}\int\left[\hat{f}_i(t) - f(t)\right]^2 dx,$$

where $\int\left[\hat{f}_i(t) - f(t)\right]^2 dx$ is defined as ISE. For BME, the sampling procedure in Section 2.3 is used on each dataset to produce $10,000$ iterations with a burn-in period of $5,000$, while for BME2, the sampling procedure in Section 3.3 is used to produce $15,000$ iterations with a burn-in period of $5,000$. The boxplots of both the ISE for the mean curve and the average ISE for individual curves are shown in Figure (4.1). The MISE results are given in Table (4.1). The results for the point-wise MSE for the mean curves are shown in Figure (4.2), while the average MSE for individual curves are shown in Figure (4.3).

The results show that our newly proposed method in general preforms better on individual curve estimations, which is as expected since the individual curves are further stratified. However, it does not seem to improve the performance on the mean curve estimation. This potential trade-off is worth further investigations.

4.5 Application

We use the same CD4 dataset as in the previous chapter for a real data analysis. Recall, the dataset includes a total number of 349 subjects with 1871 data points on 61 (in months) possible observation times. We run MCMC for $25,000$ iterations with a burn-in size of $5,000$ using the sampling procedure described in Section 3.3. Figure (4.4) illustrates the fitted mean curve and individual curves. Finally, we show in Figures (4.4b-4.4d) the estimated trajectory of log (CD4 count) for three patients. The fitted curves seem to be identical to those in the previous chapter.



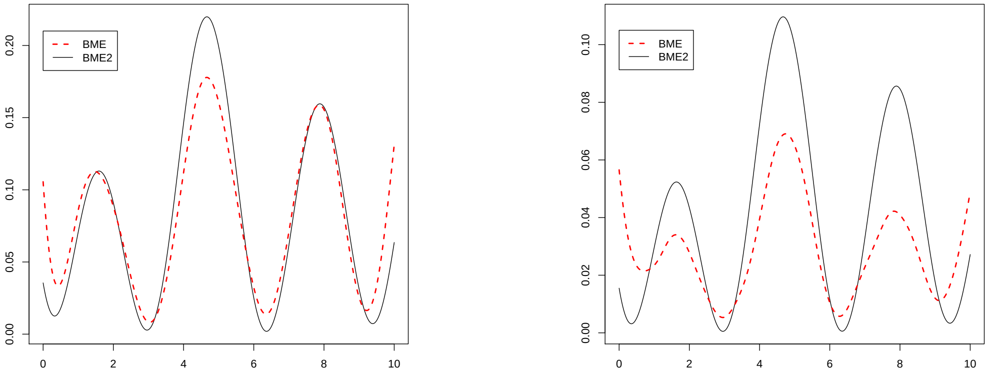
(a) $n = 50$

(b) $n = 100$

Figure 4.1: Boxplots for ISE for $n = 20, 50, 100$. The left panel for each figure represents the ISE for the mean effect function; right panel for each figure represents the average ISE for the individual effects functions.

Table 4.1: Monte Carlo approximation of MISE. The replication size is $m = 20$. Two estimators, Bayesian functional mixed effects model (BME) and Bayesian functional mixed effects model with grouped smoothness (BME2), are compared by their mean effect function and average individual effect functions.

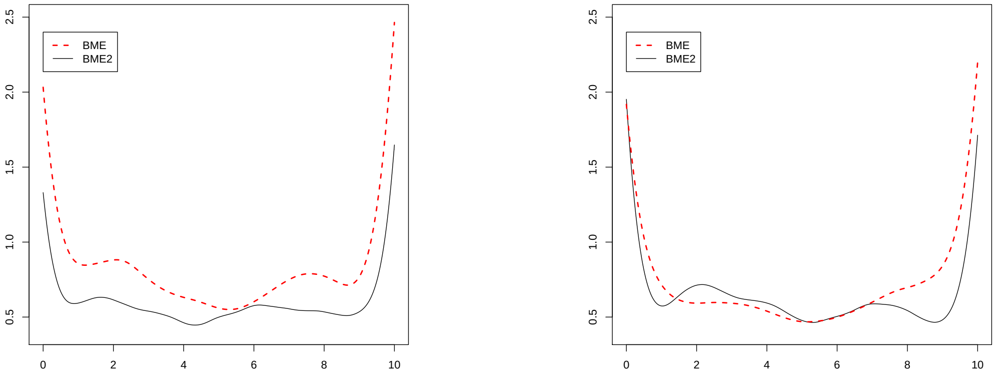
	Mean effect function		Average individual functions	
n	BME	BME2	BME	BME2
50	**0.8054**	0.8090	8.3110	**5.9551**
100	**0.2867**	0.4012	3.5775	**2.8325**



(a) $n = 50$

(b) $n = 100$

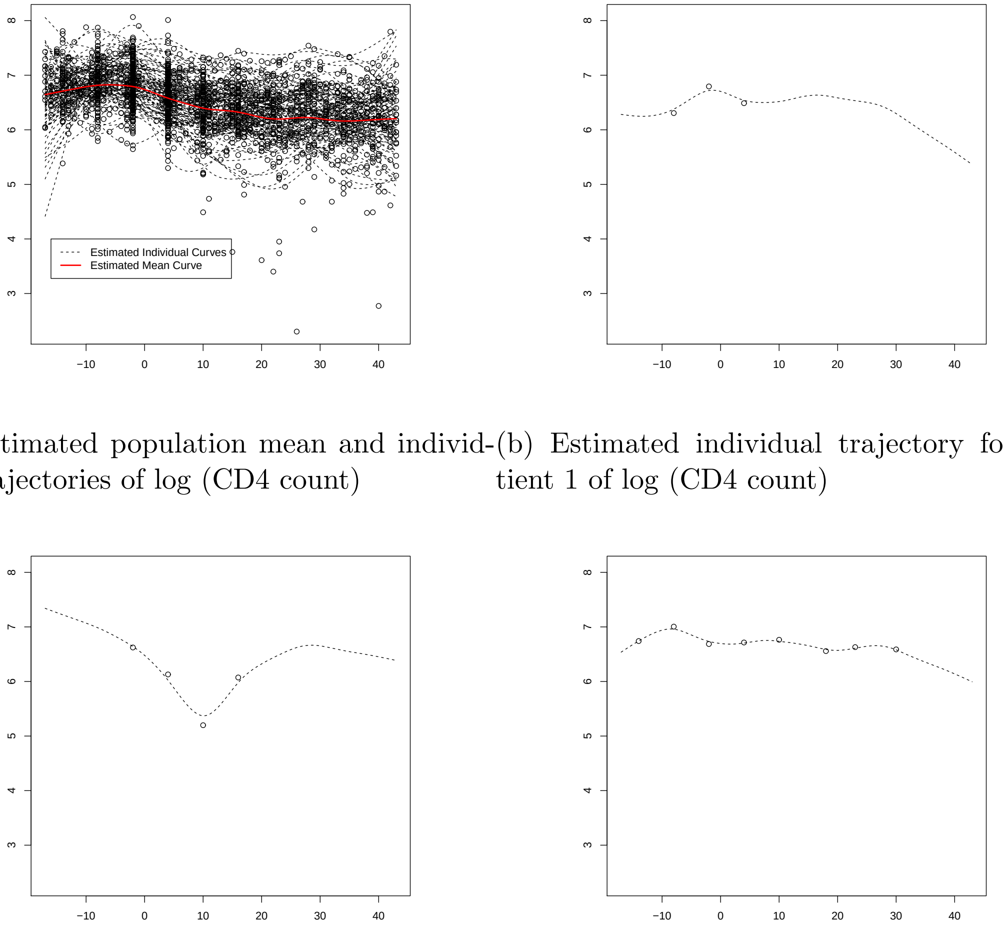
Figure 4.2: Pointwise empirical mean square error of the mean effect curves for $n = 20, 50, 100$.



(a) $n = 50$

(b) $n = 100$

Figure 4.3: Pointwise empirical mean square error of the average individual effect curves for $n = 20, 50, 100$.



(a) Estimated population mean and individual trajectories of log (CD4 count)

(b) Estimated individual trajectory for patient 1 of log (CD4 count)

(c) Estimated individual trajectory for patient 2 of log (CD4 count)

(d) Estimated individual trajectory for patient 3 of log (CD4 count)

Figure 4.4: Observed and estimated log (CD4 count) trajectories of 349 HIV-infected patients.

CHAPTER V

Orthogonal Series Density Estimation for Complex Surveys

5.1 **Introduction**

In this chapter, on a somewhat independent track, we consider orthogonal series density estimator for complex surveys, where sample design informations are considered in the model, which can cause samples neither independent nor identically distributed.

Nonparametric methods are popular for density estimations. Most work in the area of nonparametric density estimation was for independent and identically distributed samples. However, both assumptions are violated if the samples are from a finite population using a complex sampling design. Bellhouse & Stafford (1999) and Buskirk (1999) proposed kernel density estimators (KDE) by incorporating sampling weights, and their asymptotic properties were studied by Buskirk & Lohr (2005). Kernel methods for clustered samples and stratified samples were studied in Breunig (2001) and Breunig (2008), respectively.

One disadvantage of the KDE is that all samples are needed to evaluate the estimator. However, in some circumstances, there is a practical need to evaluate the estimator without using all samples for confidentiality or storage reasons. For example, many surveys are routinely conducted and sampling data are constantly collected. Data managers want to publish exact estimators without releasing all original data. In Section 6, we provide a real data example from Oklahoma M-SISNet, which is a routinely conducted survey on climate policies and public views. The orthogonal series estimators are useful alternatives to KDEs, without needing to release or store all samples.

The basic idea of the orthogonal series method is that any square integrable function f, in our case a density function, can be projected onto an orthogonal basis $\{\varphi_j\}$: $f(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x)$, where

$$\theta_j = \int \varphi_j(x) f(x) dx = \mathrm{E}(\varphi_j(X)) \tag{5.1}$$

is called the jth Fourier coefficient. Some of the work in orthogonal series density estimation (OSDE) was covered in monographs by Efromovich (1999) and Tarter & Lock (1993), among others. Efromovich (2010) gave a brief introduction of this method. Walter (1994) discussed properties of different bases. Donoho et al. (1996) and Efromovich (1996) studied data driven estimators. Asymptotic properties were studied by Pinsker (1980) and Efromovich & Pinsker (1982).

In this paper, we study the OSDE for samples from complex surveys. To the best of our knowledge, no previous work has been done on developing OSDE for finite populations. We propose a Horvitz-Thompson type of OSDE, incorporating sampling weights from the complex survey. We show that the proposed OSDE is design-unbiased and asymptotically design-consistent. We further prove the asymptotic normality of the proposed estimator. We compare the lower bound of minimax mean integrated squared error (MISE) with the I.I.D. case in Efromovich & Pinsker (1982). We propose two data driven estimators and show their efficiency in a simulation study. Finally, we analyze the M-SISNet survey data using the proposed estimation. All proofs to theorems and corollaries are given in the appendix.

5.2 Notations

Consider a finite population labeled as $U = \{1, 2, ..., N\}$. A survey variable x is associated with each unit in the finite population. A subset s of size n is selected from U according to some fixed-size sampling design $\mathcal{P}(\cdot)$. The first and second order inclusion probabilities from the sampling design $\mathcal{P}(\cdot)$ are $\pi_i = \Pr(i \in s)$ and $\pi_{ij} = \Pr(i, j \in s)$, respectively. The inverse of the first order inclusion probability defines the sampling weight $d_i = \pi_i^{-1}$, $\forall i \in s$.

The inference approach used in this paper for complex surveys is the combined design-model-based approach originated in Hartley & Sielken (1975). This approach accounts for two sources of variability. The first one is from the fact that the finite population is a realization from a superpopulation, that is, the units $\boldsymbol{x}_U = \{x_1, x_2, ..., x_N\}$ are considered independent random variables with a common distribution function F, whose density function is f. The second one is from the complex sampling procedure which leads to a sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Denote $\boldsymbol{w} = \{w_1, w_2, \ldots, w_n\}$ design variables that determine the sampling weights. The sampling design $\mathcal{P}(\cdot)$ is embedded within a probability space $(S, \mathcal{J}, P_{\mathcal{P}})$. The expectation and variance operator with respect to the sampling design are denoted by $\mathrm{E}_{\mathcal{P}}(\cdot) = \mathrm{E}_{\mathcal{P}}(\cdot \mid \boldsymbol{x}_U)$ and $\mathrm{Var}_{\mathcal{P}}(\cdot) = \mathrm{Var}_{\mathcal{P}}(\cdot \mid \boldsymbol{x}_U)$, respectively. The superpopulation ξ, from which the finite population is realized, is embedded within a probability space $(\Omega, \mathcal{F}, P_\xi)$. The sample $\boldsymbol{x}$ and the design variables $\boldsymbol{w}$ are ξ-measurable. The expectation and variance operator with respect to the model are denoted by $\mathrm{E}_\xi(\cdot)$ and $\mathrm{Var}_\xi(\cdot)$, respectively. Assume that, given the design variables $\boldsymbol{w}$, the product space, which couples the model and the design spaces, is $(\Omega \times S, \mathcal{F} \times \mathcal{J}, P_\xi \times P_{\mathcal{P}})$. The combined expectation and variance operators are denoted by $\mathrm{E}_C(\cdot)$ and $\mathrm{Var}_C(\cdot)$, where $\mathrm{E}_C(\cdot) = \mathrm{E}_\xi[\mathrm{E}_{\mathcal{P}}(\cdot \mid \boldsymbol{x}_U)]$ and $\mathrm{Var}_C(\cdot) = \mathrm{E}_\xi[\mathrm{Var}_{\mathcal{P}}(\cdot \mid \boldsymbol{x}_U)] + \mathrm{Var}_\xi[\mathrm{E}_{\mathcal{P}}(\cdot \mid \boldsymbol{x}_U)]$.

5.3 Main Results

Consider a sample $s = \{x_1, x_2, ..., x_n\}$ drawn from a finite population $\boldsymbol{x}_U$ using some fixed-size sampling design $\mathcal{P}(\cdot)$. Our goal is to estimate the hypothetical density function f of the superpopulation. Equation (5.1) implies that θ_j can be estimated using the Horvitz-Thompson (HT) estimator for the finite population mean

$$\hat{\theta}_j = N^{-1} \sum_{i=1}^{n} d_i \varphi_j(x_i), \tag{5.2}$$

where N is the finite population size and $d_i = \pi_i^{-1}$ is the sampling weight for unit i. The HT estimator is a well known design unbiased estimator (Fuller 2009). The basis $\{\varphi_j\}$ can be Fourier, polynomial, spline, wavelet, or others. Properties of different bases are discussed in Efromovich (2010). We consider the cosine basis throughout the paper, which is defined as $\{\varphi_0 = 1, \varphi_j = \sqrt{2}\cos(\pi j x)\}, j = 1, 2, \cdots, x \in [0, 1]$. Regarding the compact support $[0, 1]$ for the density, we adopt the argument in Wahba (1981):"it might be preferable to assume the true density has compact support and to scale the data to interior of $[0, 1]$." Analogous to Efromovich (1999), we propose an orthogonal series estimator in the form

$$\hat{f}(x) = \hat{f}(x, \{w_j\}) = 1 + \sum_{j=1}^{\infty} w_j \hat{\theta}_j \varphi_j(x), \tag{5.3}$$

where $\hat{\theta}_j$ is the HT estimator for the Fourier coefficient as in (5.2) and $w_j \in [0, 1]$ is a shrinking coefficient. Note that $\theta_0 = \int_0^1 f(x)dx = 1$. If $\boldsymbol{x}_U$ is known for all units in the finite population, we can write the population estimator for $f(x)$ as

$$f_U(x) = f_U(x, \{w_j\}) = 1 + \sum_{j=1}^{\infty} w_j \theta_{U,j} \varphi_j(x), \tag{5.4}$$

where $\theta_{U,j} = N^{-1} \sum_{i=1}^{N} \varphi_j(x_i)$.

The following theorems and a corollary show properties of our proposed estimator under both design and combined spaces. Theorem 5.3.1 considers unbiasedness and consistency under the design space.

Theorem 5.3.1 *Suppose* $f \in L_2(\mathbb{R})$, $\delta = N^{-1}\sum\sum_{i\neq k}\frac{\pi_{ik}}{\pi_i\pi_k} - N < \infty$ *and* $\sum_{i=1}^{\infty} w_i^2 < \infty$. *Then, the estimator* $\hat{f}(x,\{w_j\})$ *is design-unbiased and asymptotically design-consistent for* $f_U(x,\{w_j\})$, *i.e.,*

$$E_{\mathcal{P}}\left[\hat{f}(x,\{w_j\})\right] = f_U(x,\{w_j\}) \textit{ and } \Gamma_{\mathcal{P}} = Var_{\mathcal{P}}\left[\hat{f}(x,\{w_j\})\right] \to 0 \textit{ as } N \to \infty.$$

Theorem 5.3.2 shows the asymptotic normality of the proposed estimator $\hat{f}(x,\{w_j\})$ under the design space.

Theorem 5.3.2 *Suppose that all assumptions in Theorem* 5.3.1 hold. As $N \to \infty$,

$$\frac{\hat{f}(x,\{w_j\}) - f_U(x,\{w_j\})}{\hat{\Gamma}_{\mathcal{P}}} \xrightarrow{L_{\mathcal{P}}} N(0,1), \tag{5.5}$$

where

$$\hat{\Gamma}_{\mathcal{P}} = N^{-1}\sum_{j=1}^{J} w_j^2(1 + 2^{-1/2}\hat{\theta}_{2j} + \delta\hat{\theta}_j^2)(1 + 2^{-1/2}\varphi_{2j}(x)).$$

We then show the asymptotic normality of the proposed estimator $\hat{f}(x,\{w_j\})$ under the combined inference. Define a *Sobolev Class* of k-fold differentiable densities as $\mathcal{F}(k,Q) = \{f : f(x) = 1 + \sum_{j=1}^{\infty}\theta_j\varphi_j(x),\ \sum_{j=1}^{\infty}(\pi j)^{2k}\theta_j^2 \leq Q < \infty\}$, $k \geq 1$. Note that for any $f \in \mathcal{F}(k,Q)$, f is 1-periodic, $f^{(k-1)}$ is absolute differentiable and $f^{(k)} \in L_2(\mathbb{R})$.

Theorem 5.3.3 *Suppose that* $f \in \mathcal{F}(k,Q)$ *and all assumptions in Theorem 2 hold.*

Then,

$$\frac{\hat{f}(x,\{w_j\}) - f(x)}{Var_C\left[\hat{f}(x,\{w_j\})\right]} \xrightarrow{L_C} N(0,1) \text{ as } N \to \infty, \tag{5.6}$$

where $Var_C\left[\hat{f}(x,\{w_j\})\right] = N^{-1}\sum_{j=1}^{J} w_j^2 b_j(1 + 2^{-1/2}\varphi_{2j}(x))$ *and* $b_j = 2 + 2^{1/2}\theta_{2j} + (\delta - 1)\theta_j^2 + o_N(1)$.

The following corollary is a direct result of using Theorem 5.3.3 and Efromovich & Pinsker (1982). It shows the lower bound of the minimax MISE for the proposed estimator $\hat{f}(x,\{w_j\})$ under the Sobolev class.

Corollary 5.3.1 *Let* $f \in \mathcal{F}(k,Q)$ *and* $\hat{f}(x,\{w_j\})$ *be the estimator in Theorem* 5.3.3. The lower bound of the minimax MISE, under the combined inference approach, is given by:

$$R(\mathcal{F}) = \inf_{\{w_j\}} \sup_{f\in\mathcal{F}(k,Q)} \text{MISE}_C\left[\hat{f}(x,\{w_j\})\right] \geq N^{-2k/(2k+1)}P(k,Q,b)(1+o_N(1)), \tag{5.7}$$

where $P(k,Q,b) = Q^{1/(2k+1)}\left\{\frac{k}{\pi(k+1)b}\right\}^{2k/(2k+1)}$ and $b = 2$.

Remark that this lower bound is of the same form as the I.I.D. case in Efromovich & Pinsker (1982), but with $b = 2$ instead of $b = 1$.

5.4 Data Driven Estimators

The choice of shrinking coefficients $\hat{w}_j$ is not unique. To get a proper data driven estimator, we start with the oracle estimator (5.3), and then obtain $\hat{w}_j$ by minimizing the MISE for the oracle estimator. Here, we propose two estimators: a truncated estimator and a smoothed truncated estimator, mimicking those in the I.I.D. case.

The truncated estimator, denoted by $\hat{f}_T$, is an estimator with $\hat{w}_j = 1$ for $j \leq J$, and $\hat{w}_j = 0$ for $j > J$. Alternatively we can write $\hat{w}_j = I_{j \leq J}$. Then, only the truncation parameter J needs to be estimated. Notice that the MISE of this estimator is

$$\text{MISE}_C \left[\hat{f}(x, \{w_j\}) \right] = \sum_{j=1}^{J} \left[\text{Var}_C(\hat{\theta}_j) - \theta_j^2 \right] - \int f^2(x) dx.$$

Since $\int f^2(x)dx$ is fixed and an unbiased estimator for θ_j^2 is $\hat{\theta}_j^2 - N^{-1} b_j$, a data-driven estimate for J can be obtained from

$$\hat{J} = \arg\min \sum_{j=1}^{J} (2N^{-1} \hat{b}_j - \hat{\theta}_j^2),$$

where $\hat{b}_j$ is the plug-in estimator of b_j. In practice, the solution is obtained through a numerical search. Efromovich (1999) suggests to set the upper bound for $\hat{J}$ to be $\lfloor 4 + 0.5 \ln n \rfloor$ for the search. Theoretically, the minimum of the MISE can be approximated in the following corollary.

Corollary 5.4.1 *Let $f \in \mathcal{F}(k, Q)$, $k > 1/2$. The MISE of $\hat{f}_T$ is minimized when*

$$J \approx N^{1/(2k+1)} H_1(k, b, c), \tag{5.8}$$

and the minimum is approximately

$$R(\hat{f}_T) = MISE_C(\hat{f}_T(x, \{\hat{w}_j\})) \approx N^{-2k/(2k+1)} H_2(k, b, c), \tag{5.9}$$

where $H_1(k, b, c) = b^{-1/(2k+1)} \left(\frac{2k+1}{(2k+2)c} \right)^{-1/(2k+1)}$, $H_2(k, b, c) = b^{2k/(2k+1)} \left(\frac{2k+1}{(2k+2)c} \right)^{-1/(2k+1)}$, and c is a constant.

One possible modification for $\hat{f}_T$ is to shrink each Fourier coefficient toward zero.

We call this estimator the smoothed truncated estimator, denoted by $\hat{f}_S$. It is constructed similarly as the truncated estimator, with the first J Fourier coefficients shrunk by multiplying the optimal smoothing coefficients w_j^*, obtained from the proof of Corollary 5.3.1. Mathematically, $\hat{w}_j = \hat{w}_j^* I_{j \leq J}$, where $\hat{w}_j^* = (\hat{\theta}_j^2 - N^{-1}\hat{b}_j)/\hat{\theta}_j^2$ is a direct plug-in estimator for w_j^*.

A potential problem of the nonparametric density estimation is that the estimator may not be a valid density function. A simple modification is to define the L^2-projection of $\hat{f}_T$ (or $\hat{f}_S$) onto a class of non-negative densities, $\tilde{f}_T(x) = \max\{0, \hat{f}_T(x) - \text{const.}\}$, where the normalizing constant is to make $\tilde{f}_T$ integrate to 1. It has been proved that the constant always exists and is unique (Glad et al. 2003).

5.5 Simulation

We compared our proposed estimators with the series estimator that ignores the finite population and sampling designs, through a Monte Carlo simulation study. We considered estimating density functions for three sampling designs: (1) the simple random sample without replacement (SRSWOR), (2) the stratified sampling and (3) the Poisson sampling. Note that the Poisson sampling has a random size with units independently sampled and hence violates our assumption of fixed size sampling.

1. For the SRSWOR, we considered two superpopulations: the standard normal distribution $\mathrm{N}(0,1)$ and a mixture normal distribution $0.4\mathrm{N}(-1, 0.5) + 0.6\mathrm{N}(1,1)$.

2. For the stratified sampling, we considered two superpopulations: a two-component mixture normal $0.4\mathrm{N}(-1, 0.5) + 0.6\mathrm{N}(1,1)$ and a three-component mixture normal $0.3\mathrm{N}(-1, 0.15) + 0.4\mathrm{N}(0, 0.15) + 0.3\mathrm{N}(1, 0.15)$. We designed two strata for the two-component mixture and three strata for the three-component mixture. A proportional stratified sampling is used.

3. For the Poisson sampling, we considered the same two superpopulations as in (1). We specified the expected sample size for the Poisson sampling to be n, and generated the first order inclusion probabilities for the Poisson sampling using the function "inclusionprobabilities" in the R package "sampling" (Till & Matei 2016).

For all cases, we considered a finite population of size $N = 1,000$ drawn from each of the superpopulations. We repeated drawing the finite population for $m_1 = 100$ times. For each of the finite population, we drew samples according to the sampling design, with increasing sample sizes: $n = 20, 40, 60$ and 80. The replication number for each finite population is $m_2 = 10,000$. The performance of estimators is measured by a Monte Carlo approximation of the MISE:

$$\text{MISE}_{\text{MC}}(\tilde{f}) = \int \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left[\tilde{f}_{ij}(x) - f(x) \right]^2 dx.$$

The results of the simulation study are shown in Table 1. In general, the I.I.D. series estimator, which ignores the sampling design, performs the worst in all cases. However, it is not surprising to see that the improvement for the proposed estimators is much more in stratified sampling than in SRSWOR or Poisson sampling. It confirms the necessity of incorporating stratification sampling weights into the series estimator for a complex survey. Lastly, the smoothed truncated estimator performs better than the truncated estimator in most scenarios.

5.6 Oklahoma M-SISNet Survey

The Oklahoma Weather, Society and Government Survey conducted by Meso-Scale Integrated Sociogeographic Network (M-SISNet) measures Oklahomans' perceptions of weather in the state, their views on government policies and societal issues and their use of water and energy. The survey is routinely conduced at the end of each season. Until the end of 2016, 12 waves of survey data have been collected. It is desired that estimates can be obtained without constantly pulling out the original data. The sampling design has two separated phases. In Phase I, a simple random sample of size $n = 1,500$ is selected from statewide households. In Phase II, a stratified oversample is selected from five special study areas: Payne County, Oklahoma City County, Kiamichi County, Washita County and Canadian County. In each stratum, the sample size is fixed to be 200. The second phase can be viewed as a stratified sampling over the entire state with six strata: $n_1 = \cdots = n_5 = 200$ and $n_6 = 0$, where the sixth stratum contains households not in the five special study areas. This design with oversampling is not a typical fixed-size complex survey. The first-order inclusion probabilities are approximately $\pi_{hi} = n_h/N_h + n/N$, for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 6$. Note that for units not in the five areas, this inclusion probability is simply n/N. We presents OSDEs for two continuous variables for illustration: the monthly electricity bill and the monthly water bill. Figure 5.1 shows OSDEs of the two variables for all seasons in 2015.

Appendix

Proof of Theorem 3.3.1

Proof. We first show that $\hat{f}(x, \{w_j\})$ is design-unbiased:

$$\begin{aligned}
\mathrm{E}_{\mathcal{P}}\left[\hat{f}(x, \{w_j\})\right] &= \mathrm{E}_{\mathcal{P}}\left[1 + \sum_{j=1}^{\infty} w_j \hat{\theta}_j \varphi_j(x)\right] \\
&= 1 + \sum_{j=1}^{\infty} w_j \mathrm{E}_{\mathcal{P}}(\hat{\theta}_j) \varphi_j(x) \\
&= 1 + \sum_{j=1}^{\infty} w_j \theta_{U,j} \varphi_j(x) \\
&= f_U(x, \{w_j\}).
\end{aligned}$$

It remains to show that $\hat{f}(x,\{w_j\})$ is asymptotically design-consistent, that is, the design-variance of $\hat{f}(x,\{w_j\})$ approaches zero in the limit. We need the simple fact that

$$\varphi_j^2(x) = [\sqrt{2}\cos(\pi jx)]^2 = 1+\cos(\pi 2jx) = 1+2^{-1/2}\varphi_{2j}(x).$$

Then, we have

$$\begin{aligned}
\Gamma_{\mathcal{P}} &= \mathrm{Var}_{\mathcal{P}}\left[1+\sum_{j=1}^{\infty} w_j\hat{\theta}_j\varphi_j(x)\right] \\
&= \sum_{j=1}^{\infty} w_j^2\varphi_j^2(x)\mathrm{Var}_{\mathcal{P}}(\hat{\theta}_j) \\
&= \sum_{j=1}^{\infty} w_j^2\left[1+2^{-1/2}\varphi_{2j}(x)\right]N^{-2}\mathrm{Var}_{\mathcal{P}}\left[\sum_{i=1}^{n} d_i\varphi_j(x_i)\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}_{\mathcal{P}}\left[\sum_{i=1}^{n} d_i\varphi_j(x_i)\right] &= \mathrm{Var}_{\mathcal{P}}\left[\sum_{i=1}^{N} I_i d_i\varphi_j(x_i)\right] \\
&= \mathrm{E}_{\mathcal{P}}\left[\sum_{i=1}^{N} I_i d_i\varphi_j(x_i)\right]^2 - \left\{\mathrm{E}_{\mathcal{P}}\left[\sum_{i=1}^{N} I_i d_i\varphi_j(x_i)\right]\right\}^2 \\
&= \sum_{i=1}^{N}\mathrm{E}_{\mathcal{P}}(I_i^2)d_i^2\mathrm{E}_{\mathcal{P}}\left[\varphi_j^2(x_i)\right] + \sum\sum_{i\neq k}\pi_{ik}d_id_k\mathrm{E}_{\mathcal{P}}\left[\varphi_j(x_i)\right]\mathrm{E}_{\mathcal{P}}\left[\varphi_k(x_k)\right] \\
&\quad -\left\{\sum_{i=1}^{N}\mathrm{E}_{\mathcal{P}}(I_i)d_i\mathrm{E}_{\mathcal{P}}\left[\varphi_j(x_i)\right]\right\}^2 \\
&= \sum_{i=1}^{N}\mathrm{E}_{\mathcal{P}}\left[1+2^{-1/2}\varphi_{2j}(x_i)\right] + \sum\sum_{i\neq k}\frac{\pi_{ik}}{\pi_i\pi_k}\theta_{U,j}^2 - N^2\theta_{U,j}^2 \\
&= N(1+2^{-1/2}\theta_{U,2j}+\delta\theta_{U,j}^2) \\
&\leq NM,
\end{aligned}$$

where $1+2^{-1/2}\theta_{U,2j}+\delta\theta_{U,j}^2 \leq M < \infty$ for every j.

Hence, $\Gamma_{\mathcal{P}} \leq N^{-1}M\sum_{j=1}^{\infty} w_j^2 \to 0$ as $N\to\infty$. □

Proof of Theorem 3.3.1

Proof. By the definition of $\hat{\theta}_j$ and $\theta_{U,j}$, we have

$$\begin{aligned}\hat{f}(x, \{w_j\}) &= 1 + \sum_{j=1}^{\infty} w_j \hat{\theta}_j \varphi_j(x) \\ &= 1 + \sum_{i=1}^{N} I_i d_i \sum_{j=1}^{\infty} w_j \varphi_j(x) \varphi_j(x_i),\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left[I_i d_i \sum_{j=1}^{\infty} w_j \varphi_j(x) \varphi_j(x_i) \right] &= \sum_{j=1}^{\infty} w_j \varphi_j(x) \mathrm{E}\left[\varphi_j(x_i)\right] \\ &= \sum_{j=1}^{\infty} w_j \theta_{U,j} \varphi_j(x).\end{aligned}$$

Also, from the proof of Theorem 1, we have

$$\begin{aligned}\mathrm{Var}\left[I_i d_i \sum_{j=1}^{\infty} w_j \varphi_j(x) \varphi_j(x_i) \right] &= \sum_{j=1}^{\infty} w_j^2 (1 + 2^{-1/2} \theta_{U,2j} + \delta \theta_{U,j}^2) \\ &\leq B \sum_{j=1}^{\infty} w_j^2 < \infty \text{ by assumption.}\end{aligned}$$

Therefore, by the Lindeberg-L*é*vy central limit theorem, we have

$$\frac{\hat{f}(x, \{w_j\}) - f_U(x, \{w_j\})}{\Gamma_{\mathcal{P}}} \xrightarrow{L_{\mathcal{P}}} N(0, 1). \tag{5.10}$$

It remains to show that $\hat{\Gamma}_{\mathcal{P}}$ is consistent for $\Gamma_{\mathcal{P}}$ under design, or equivalently,

$$|\hat{\Gamma}_{\mathcal{P}} - \Gamma_{\mathcal{P}}| \xrightarrow{P_{\mathcal{P}}} 0, \text{ as } n \to N. \tag{5.11}$$

Condition (5.11) can be proved by using the facts that $\hat{\theta}_j$ is design unbiased and $\mathrm{E}(\hat{\theta}_j^2) = \theta_j^2 + \mathrm{Var}(\hat{\theta}_j) \to \theta_j^2$ as $n \to N$.

Then, Theorem 2 is proved by using the equations (5.10) and (5.11) in conjunction with

Slutsky's theorem. □

Proof of Theorem 3.3.3

Proof. Since $f_U(x, \{w_j\})$ is the standard OSDE from an I.I.D. sample which is the finite population, then

$$\frac{f_U(x, \{w_j\}) - f(x)}{\mathrm{Var}_\xi\left[f_U(x, \{w_j\})\right]} \xrightarrow{L_\xi} N(0, 1). \tag{5.12}$$

The asymptotic distribution of the I.I.D. OSDE under Sobolev class is obtained from Efromovich (1999), Chapter 7. Also,

$$\begin{aligned} \mathrm{Var}_C\left[\hat{f}(x, \{w_j\})\right] &= \sum_{j=1}^{J} \mathrm{Var}_C\left[w_j \hat{\theta}_j \varphi_j(x)\right] \\ &= \sum_{j=1}^{J} w_j^2 (1 + 2^{-1/2}\varphi_{2j}(x)) \mathrm{Var}_C(\hat{\theta}_j) \end{aligned} \tag{5.13}$$

Next, we calculate the variance of $\hat{\theta}_j$ by using Theorem 1:

$$\begin{aligned} \mathrm{Var}_C(\hat{\theta}_j) &= \mathrm{E}_\xi\left[\mathrm{Var}_{\mathcal{P}}(\hat{\theta}_j)\right] + \mathrm{Var}_\xi\left[\mathrm{E}_{\mathcal{P}}(\hat{\theta}_j)\right] \\ &= \mathrm{E}_\xi\left[N^{-1}(1 + 2^{-1/2}\theta_{U,2j} + \delta\theta_{U,j}^2)\right] + \mathrm{Var}_\xi(\theta_{U,j}) \\ &= N^{-1}\left[1 + 2^{-1/2}\theta_{2j} + \delta \mathrm{E}_\xi(\theta_{U,j}^2)\right] + \mathrm{Var}_\xi(\theta_{U,j}) \end{aligned} \tag{5.14}$$

Then, we evaluate $\mathrm{E}_\xi(\theta_{U,j}^2)$ and $\mathrm{Var}_\xi(\theta_{U,j})$ separately. Based on a standard result in the I.I.D. case, we have

$$\mathrm{Var}_\xi(\theta_{U,j}) = N^{-1}(1 + 2^{-1/2}\theta_{2j} - \theta_j^2) \tag{5.15}$$

and

$$\mathrm{E}_\xi(\theta_{U,j}^2) = \mathrm{E}_\xi^2(\theta_{U,j}) + \mathrm{Var}_\xi(\theta_{U,j})$$

$$= \quad N^{-1}(1 + 2^{-1/2}\theta_{2j} - \theta_j^2) + \theta_j^2. \tag{5.16}$$

Then, plug equations (5.15) and (5.16) into (5.14), we have

$$\mathrm{Var}_C(\hat{\theta}_j) = N^{-1}\left[2 + 2^{1/2}\theta_{2j} + (\delta - 1)\theta_j^2 + o_N(1)\right] = N^{-1}b_j. \tag{5.17}$$

Hence, plug (5.17) into (5.13) we can get the variance of $\hat{f}$ under the combined inference approach.

Finally, apply Theorem 5.1 in Bleuer & Kratina (1999), Theorem 3 is proved. □

Proof of Corollary 3.3.1

Proof. The proof is similar to Efromovich & Pinsker (1982). We sketch the steps as follows. We first evaluate the linear minimax MISE for the functions in the Sobolev class defined above. That is, we optimize w_j^*'s that minimize $\mathrm{MISE}_C(\hat{f})$. Notice that $\mathrm{E}_C(\hat{\theta}_j) = \mathrm{E}_\xi[\mathrm{E}_{\mathcal{P}}(\hat{\theta}_j)] = \mathrm{E}_\xi(\theta_{U,j}) = \theta_j$ implying that $\hat{\theta}_j$ is an unbiased estimator of θ_j. Therefore,

$$\begin{aligned}\mathrm{MISE}_C\left[\hat{f}(x, \{w_j\})\right] &= \mathrm{E}_C\left[\int (f - \hat{f})^2\right] \\ &= \sum_{j=1}^{\infty}\left\{w_j^2\left[\mathrm{Var}_C(\hat{\theta}_j) + \theta_j^2\right] - 2w_j\theta_j^2 + \theta_j^2\right\}. \end{aligned} \tag{5.18}$$

A straightforward calculation yields that

$$w_j^* = \frac{\theta_j^2}{\theta_j^2 + \mathrm{Var}_C(\hat{\theta}_j)}. \tag{5.19}$$

Plug equation (5.19) into (5.18),

$$\begin{aligned}R_L(\mathcal{F}) &= \inf_{\{w_j\}} \sup_{f \in \mathcal{F}(k,Q)} \mathrm{MISE}_C\left[\hat{f}(x, \{w_j\})\right] \\ &\geq \sup_{f \in \mathcal{F}(k,Q)} \sum_{j=1}^{\infty} \frac{\theta_j^2 \mathrm{Var}_C(\hat{\theta}_j)}{\theta_j^2 + \mathrm{Var}_C(\hat{\theta}_j)}, \end{aligned} \tag{5.20}$$

where $\mathrm{Var}_C(\hat{\theta}_j)$ is of the form (5.17). Plug (5.17) into (5.20), and use the Lagrange multiplier to show that the maximum of (6) is attained at

$$\theta_j^2 = N^{-1}(\mu/(\pi j)^k - b_j)_+, \tag{5.21}$$

where μ is determined by the constraint $\sum_{j=1}^{\infty}(\pi j)^{2k}\theta_j^2 \leq Q$. Plug equation (5.21) back to (5.20), we obtain

$$R_L(\mathcal{F}) \geq N^{-2k/(2k+1)}P(k,Q,b).$$

Pinsker (1980) shows that for Sobolev ball $\mathcal{F}$, the linear minimax risk is asymptotically equal to the minimax risk, that is, $R(\mathcal{F}) = R_L(\mathcal{F})(1 + o_N(1))$. Therefore Corollary 1 is proved. □

Proof of Corollary 3.4.1

Proof. Let $\hat{w}_j = I_{j\leq J}$. Plug equation (5.17) into (5.18), we have

$$R(\hat{f}_T) = N^{-1}\sum_{j=1}^{J} b_j + \sum_{j=J+1}^{\infty}\theta_j^2 \approx N^{-1}bJ + \sum_{j=J+1}^{\infty}\theta_j^2. \tag{5.22}$$

Notice that for $f \in \mathcal{F}(k,Q)$. By a straightforward calculation, we have $\theta_j^2 = cj^{-2(k+1)}$ (Efromovich 1999). Therefore,

$$\sum_{j=J+1}^{\infty}\theta_j^2 \approx c\int_J^{\infty} j^{-2(k+1)}dj = \frac{c}{2k+1}J^{-2k-1}. \tag{5.23}$$

Plug (5.23) into (5.22) and optimize J, Corollary 2 is proved. □

Table 5.1: Monte Carlo approximation of MISE for three sampling designs and two superpopulations. The finite population size is $N = 1,000$. The replication size of the finite population is $m_1 = 100$, and the replication size of the sample is $m_2 = 10,000$. Three estimators are compared: the truncated estimator, the smoothed estimator and the series estimator ignoring finite population and sampling design (I.I.D.).

SRSWOR						
	Standard Normal			Mixture Normal		
n	Truncated	Smoothed	I.I.D.	Truncated	Smoothed	I.I.D.
20	0.0232	**0.0220**	0.0290	0.0498	**0.0480**	0.0535
40	0.0150	**0.0140**	0.0157	**0.0311**	0.0318	0.0388
60	0.0116	**0.0109**	0.0121	**0.0226**	0.0234	0.0335
80	0.0094	**0.0089**	0.0100	**0.0173**	0.0180	0.0219
Poisson Sampling						
	Standard Normal			Mixture Normal		
n	Truncated	Smoothed	I.I.D.	Truncated	Smoothed	I.I.D.
20	0.0497	**0.0481**	0.0527	0.0580	**0.0442**	0.0705
40	0.0281	**0.0270**	0.0392	0.0344	**0.0294**	0.0399
60	0.0241	**0.0229**	0.0237	0.0283	**0.0280**	0.0322
80	0.0201	**0.0190**	0.0211	0.0235	**0.0234**	0.0285
Stratified Sampling						
	Two Strata			Three Strata		
n	Truncated	Smoothed	I.I.D.	Truncated	Smoothed	I.I.D.
20	0.0415	**0.0409**	0.0739	0.2847	**0.2826**	0.3106
40	0.0231	**0.0230**	0.0688	0.2731	**0.2718**	0.3309
60	0.0181	**0.0180**	0.0672	0.0426	**0.0419**	0.1132
80	**0.0142**	**0.0142**	0.0675	0.0412	**0.0406**	0.1175

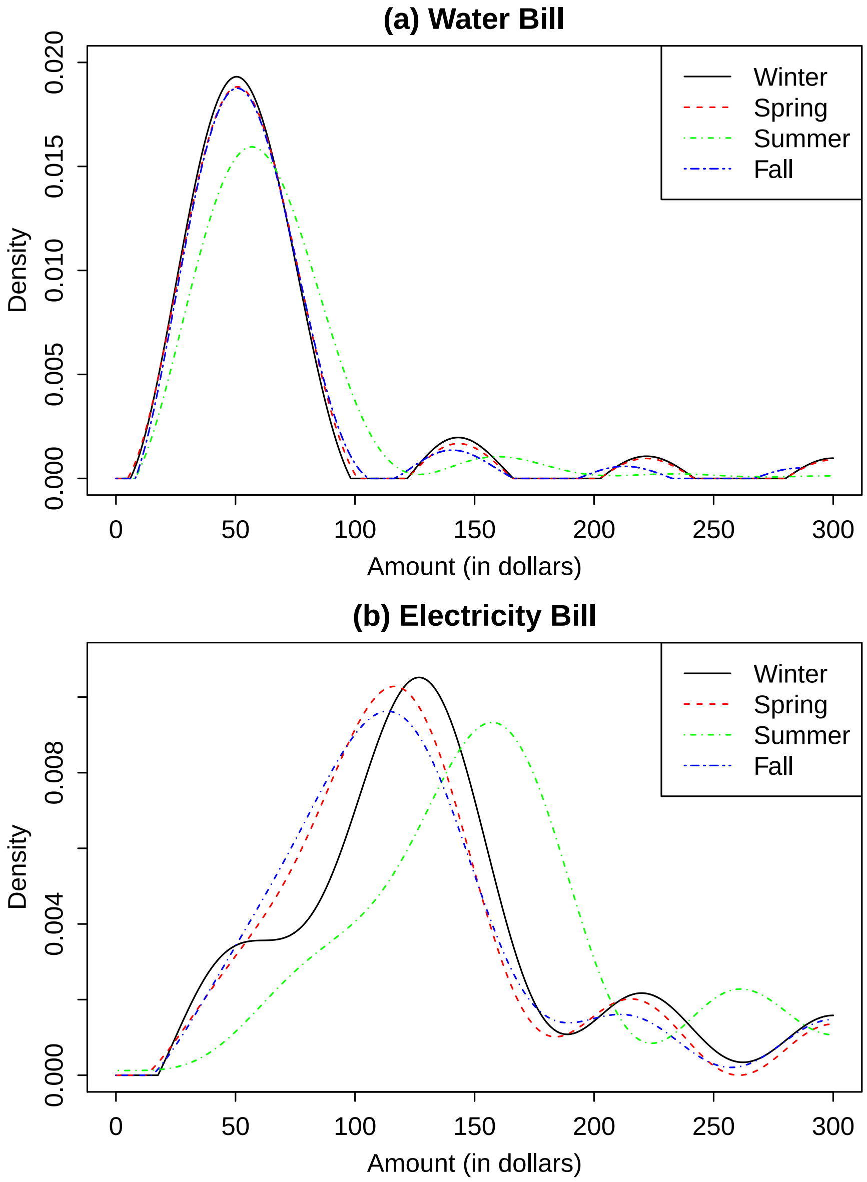


Figure 5.1: OSDEs of the electricity bill and the water bill for seasonal waves in 2015.

Bibliography

Antoniadis, A., Brossat, X., Cugliari, J. & Poggi, J.-M. (2013), 'Clustering functional data using wavelets', *International Journal of Wavelets, Multiresolution and Information Processing* **11**, 1350003.

Bellhouse, D. R. & Stafford, J. E. (1999), 'Density estimation from complex surveys', *Statistica Sinica* **9**, 407–424.

Benhenni, K. & Degras, D. (2014), 'Local polynomial estimation of the mean function and its derivatives based on functional data and regular designs', **18**, 881–899.

Berk, M. (2012), Statistical Methods for Replicated, High-Dimensional Biological Time Series, PhD thesis, Imperial College London.

Berk, M. (2013), 'sme: Smoothing-splines mixed-effects models', *R package version 0.8. h. See https://CRAN. R-project. org/package= sme* .

Berk, M., Hemingway, C., Levin, M. & Montana, G. (2012), Longitudinal analysis of gene expression profiles using functional mixed-effects models, *in* 'Advanced Statistical Methods for the Analysis of Large Data-Sets', Springer, pp. 57–67.

Besse, P. & Ramsay, J. O. (1986), 'Principal components analysis of sampled functions', *Psychometrika* **51**(2), 285–311.

Bleuer, S. R. & Kratina, I. S. (1999), 'On the two-phase framework for joint model and design-based inference', *The Annals of Statistics* **33**, 2789–2810.

Breunig, R. V. (2001), 'Density estimation for clustered data', *Econometric Reviews* **20**, 353–367.

Breunig, R. V. (2008), 'Nonparametric density estimation for stratified samples', *Statistics and Probability Letters* **78**, 2194–2200.

Brumback, B. A. & Rice, J. A. (1998), 'Smoothing spline models for the analysis of nested and crossed samples of curves', *Journal of the American Statistical Association* **93**(443), 961–976.

Buskirk, T. D. (1999), Using nonparametric methods for density estimation with complex survey data, Technical report, PhD thesis, Department of Mathematics, Arizona State University.

Buskirk, T. D. & Lohr, S. L. (2005), 'Asymptotic properties of kernel density estimation with complex survey data', *Journal of Statistical Planning and Inference* **128**, 165–190.

Cai, T. & Yuan, M. (2010), 'Nonparametric covariance function estimation for functional and longitudinal data', *University of Pennsylvania and Georgia inistitute of technology* .

Chib, S. & Greenberg, E. (1995), 'Understanding the metropolis-hastings algorithm', *The american statistician* **49**(4), 327–335.

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L., Scheipl, F., Greven, S., Harezlak, J., Kundu, M. & Zhao, Y. (2012), 'refund: Regression with functional data', *R package version 0.1-6* .

de Boor, C. (1972), 'On calculating with b-splines', *Journal of Approximation Theory* **6(1)**, 50–62.

de Boor, C. (1978), *A Pratical Guide to Splines*, Springer, New York.

Donoho, D., Johnstone, I., Kerkyacharian, G. & Picard, D. (1996), 'Density estimation by wavelet thresholding', *Annals of Statistics* **24**, 508–539.

Dung, V. T. & Tjahjowidodo, T. (2017), 'A direct method to solce optimal knots of b-spline curves fitting', *PLoS One* **12(3)**, e0173857.

Durban, M., Harezlak, J., P., W. M. & Carroll, R. J. (2005), 'Simple fitting of subject-specific curves for longitudinal data', *Statistics in Medicine* **24**, 1153–1167.

Efromovich, S. (1996), 'Adaptive orthogonal series density estimation for small samples', *Computational Statistics and Data Analysis* **22**, 599–617.

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theorey and Applications*, New York: Springer.

Efromovich, S. (2010), 'Orthogonal series density estimation', *WIREs Comp Stat* **2**, 467–476.

Efromovich, S. & Pinsker, M. S. (1982), 'Estimation of square-integrable probability density of a random variable', *Problems of Information Transmission* **18**, 19–38.

Fan, J. & Gijbels, I. (1996), 'Local polynomial modelling and its applications', *Monographs on Statistics and Applied Probability. Chapman & Hall/CRC* .

Fuller, W. A. (2009), *Sampling Statistics*, Wiley, New York.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, CRC press.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian analysis* **1**(3), 515–534.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. J. (1992), 'Practical markov chain monte carlo', *Statistical Science* pp. 473–483.

Giacofci, M., Lambert-Lacroix, S., Marot, G. & Picard, F. (2013), 'Wavelet-based clustering for mixed-effects functional models in high dimension', *Biometrics* **69**, 31–40.

Givens, G. H. & Hoeting, J. A. (2005), *Computational Statistics*, John Wiley and Sons.

Glad, I. K., Hjort, N. L. & Ushakov, N. G. (2003), 'Correction of density estimators that are not densities', *Scandinavian Journal of Statistics* **30**, 415–427.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. & Reich, D. (2011), 'Penalized functional regression', *Journal of Computational and Graphical Statistics* **20**(4), 830–851.

Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall.

Gu, C. (2013), *Smoothing Spline ANOVA Models*, Springer.

Hartley, H. O. & Sielken, R. L. (1975), 'A super-population viewpoint for finite population sampling', *Biometrics* **31**, 411–422.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009), *The elements of statistical learning*, Springer, New York.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**, 97–109.

Huang, A., Wand, M. P. et al. (2013), 'Simple marginally noninformative prior distributions for covariance matrices', *Bayesian Analysis* **8**(2), 439–452.

James, G. (2010), Sparseness and funtional data analysis, *in* F. Ferraty & Y. Romain, eds, 'The Oxford Handbook of Functional Data Analysis', Oxford: Oxford University Press.

James, G. M., Hastie, T. J. & Sugar, C. A. (2000), 'Principal component models for sparse functional data', *Biometrika* **87**(3), 587–602.

James, M., Hastie, T. & Friedman, J. (2005), 'Principle component models for sparse functional data', *Biometrika* **87**, 587–602.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., RINALDO Jr, C. R. & STUDY, M. A. C. (1987), 'The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants', *American journal of epidemiology* **126**(2), 310–318.

Kaufman, C. G. & Sain, S. R. (2010), 'Bayesian functional anova modeling using gaussian process prior distributions', *Bayesian Analysis* **5**, 123–150.

Peng, J. & Paul, D. (2009), 'A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data', *Journal of Computational and Graphical Statistics* **18**(4), 995–1015.

Petris, G. & Tardella, L. (2003), 'A geometric approach to transdimensional markov chain monte carlo', *Canadian journal of Statistics* **31**(4), 469–482.

Pinsker, M. S. (1980), 'Optimal filtration of square-integrable signals in Gaussian noise', *Problems Inform. Transmission* **16**, 53–68.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York.

Reiss, P. T., Huang, L. & Mennes, M. (2010), 'Fast function-on-scalar regression with penalized basis expansions', *The international journal of biostatistics* **6**(1).

Rice, J. & Silverman, B. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society, Series B* **53**, 233–243.

Robert, C. P. & Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer.

Ruppert, D. (2002), 'Selecting the number of knots for penalized splines', *Journal of Computational and Graphical Statistics* **11:23**, 735–757.

Ruppert, D., Wand, W. P. & Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Tarter, M. E. & Lock, M. D. (1993), *Model-Free Curve Estimation*, New York: Chapman and Hall.

Thompson, W. K. & Rosen, O. (2008), 'A bayesian model for sparse functional data', *Biometrics* **64:1**, 54–63.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *The Annals of Statistics* **22(4)**, 1701–1728.

Till, Y. & Matei, A. (2016), *sampling: Survey Sampling*. R package version 2.8.
URL: *https://CRAN.R-project.org/package=sampling*

Wahba, G. (1981), 'Data-based optimal smoothing of orthogonal series density estimates', *The Annals of Statistics* **9**, 146–156.

Walter, G. G. (1994), *Wavelets and other Orthogonal Systems with Applications*, London: CRC Press.

Wand, M. P. (2003), 'Smoothing and mixed models', *Computational Statistics* **18**, 223–249.

Wand, M. P. & Ormerod, J. T. (2009), 'On semiparametric regression with o'sullivan penalized splines', *Australian and New Zealand Journal of Statistics* **50**, 179–198.

Wand, M. P. & Ormerod, J. T. (2011), 'Penalized wavelets: Embedding wavelets into semiparametric regression', *Electronic Journal of Statistics* **5**, 1654–1717.

Wand, M. P., Ormerod, J. T., Padoan, S. A., Frühwirth, R. et al. (2011), 'Mean field variational bayes for elaborate distributions', *Bayesian Analysis* **6**(4), 847–900.

Wu, H. & Zhang, J. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*, Wiley, New Jersey.

Xiao, L., Li, C., Checkley, W. & Crainiceanu, C. (2017), 'Fast covariance estimation for sparse functional data', *Statistics and Computing* pp. 1–12.

Yao, F., Muller, H. & Wang, J. (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association* **100:470**, 577–590.

Zhang, X. & Wang, J. (2016), 'From sparse to dense functional data and beyond', *The Annuals of Statistics* **44:5**, 2281–2321.

Zhao, Y., Ogden, R. T. & Reiss, P. T. (2012), 'Wavelet-based lasso in functional linear regression', *Journal of Computational and Graphical Statistics* **21:3**, 600–617.

VITA

Shangyuan Ye

Candidate for the Degree of

Doctor of Philosophy

Dissertation: BAYESIAN ANALYSIS FOR SPARSE FUNCTIONAL DATA

Major Filed: Statistics

Biographical:

- Education:

  Complete the requirements for the Doctor of Philosophy in Statistics at Oklahoma State University, Stillwater, Oklahoma in May, 2018.

  Complete the requirements for the Bachelor of Art in Mathematics at The University of Texas at Dallas, Richardson, Texas in May, 2014.

- Experience:

  Employed as graduate teaching assistant and instructor by the Department of Statistics, Oklahoma State University between August 2014 to July 2018.

- Professional Memberships:

  Member of the American Statistical Association since February 2017.