Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Will Zhu	Date

Statistical Methods for Handling Missing Data in Functional Data Analysis

By

Will Zhu

Doctor of Philosophy

Biostatistics

Amita Manatunga, Ph.D.
Advisor

Qi Long, Ph.D.
Co-Advisor

Howard Chang, Ph.D.
Committee Member

Andrew Taylor, M.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Handling Missing Data in Functional Data Analysis

By

Will Zhu
B.S., Emory University, 2009

Advisors: Amita Manatunga, Ph.D. and Qi Long, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Abstract

Statistical analyses of functional data have drawn increased attention in recent years, yet handling missing data remains a notable obstacle in functional data analysis. This work is motivated by a renal study on detection of kidney obstruction, where up to two imaging scans, namely, baseline scan and the scan after furosemide treatment, are available for each kidney, resulting in two curves. In some cases, the kidney is judged to be non-obstructed and the patient does not receive furosemide, resulting in missing data for the second scan.

First, our objective is to develop a method that can impute the second curve based on the first curve, assuming that the first curve is informative about the missing second curve (Chapter 2). We model the curves for each individual using a set of potential basis functions and posit a sparse latent factor model for the basis coefficients, in which a shrinkage prior is assigned to the loadings to induce basis selection. We employ a Bayesian data augmentation algorithm to simultaneously estimate the model parameters and impute the missing curves. Our method is evaluated and compared to existing methods through a simulation study. We illustrate our method using a renal study, in which we impute the second curve for a kidney with a missing second curve, which can be useful in the interpretation of kidney obstruction.

In the same data situation with missing second curve, we consider an analysis of relationship between functional covariates and a binary outcome. We employ a Bayesian hierarchical model for jointly modeling the curves that are measured with error and the association between noise-free curves and the binary outcome in the presence of missing data. We consider two approaches of selecting basis functions for modeling the curves and for parameterizing functional coefficients in the functional generalized linear model used to model the association. In the first approach (Chapter 3), we use cubic B-spline basis functions and use deviance information criterion to select number of basis functions.

To overcome the difficulty in selecting basis functions, alternatively, we utilize functional principal component analysis (FPCA) to derive a more parsimonious model within the same framework, based on selecting functional principal components that explain large percent of variation in the curves (Chapter 4). We conduct simulation studies to assess the performance of the proposed methods in the presence of missing functional data. We illustrate our methods with the application to renal study.

Statistical Methods for Handling Missing Data in Functional Data Analysis

By

Will Zhu
Bachelors Degree, Emory University, 2009

Advisors: Amita Manatunga, Ph.D. and Qi Long, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Acknowledgement

I would like to express my deepest gratitude for my advisors, Dr. Amita Manatunga and Dr. Qi Long, for their guidance and patience over the past few years. I am especially very fortunate to have Dr. Manatunga as my advisor. She not only taught me how to conduct scientific research and how to write scientific paper, but also showed me how to be successful at work and helped me through many difficult times. I am also very grateful to have Dr. Long as my co-advisor. He encouraged me to think critically to solve problems I stumbled upon and provided helpful insights to issues I encountered in my research. Without their enormous and unconditional help, I would not be able to reach where I am.

I would also like to thank Dr. Howard Chang and Dr. Andrew Taylor for their time in serving on my dissertation committee. Their comments and suggestions led to substantial improvements in my dissertation and my presentations.

I would like to thank the faculty and staff here at the Department of Biostatistics and Bioinformatics for their great support in my graduate study. I learned a great deal from each and every one of you, whether through classes you taught, or attending a seminar you gave, or just by talking to you. I am indebted to Dr. Tianwei Yu, who introduced me to the field of biostatistics, to Dr. Michael Kutner and Dr. Lance Waller, who provided me exciting collaboration experiences in rehabilitation research and pediatric research, and to Paul Weiss, who showed me how to teach biostatistics to students with diverse backgrounds. I also want to thank all the administrative staff in the department, who have helped take care of many things to make sure everything goes smoothly for me.

Finally, my gratitude goes to my parents, whose faith and support made my life much easier, so I can focus on my dissertation. Without their help, this work would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent years, statistical analyses of functional data have drawn increased attention. Functional data are often encountered in biomedical studies and methods are available to include functional data both as a response and a predictor (Cai and Hall, 2006; Fan and Zhang, 2000; Ferraty and Vieu, 2003; Hyndman and Ullah, 2007; James and Hastie, 2001; James et al., 2000; Ramsay and Silverman, 2002; Yao et al., 2005a,b). A notable obstacle in the functional data analysis of biomedical data is how to handle missing data. The objective of this dissertation is to develop statistical methods for handling missing data for certain situations in the analysis of functional data. This work is motivated by a study, conducted in the Division of Nuclear Medicine at Emory University, aimed at improving renal image interpretations by radiologists and nuclear medicine physicians. Although our proposed methods are motivated by renal imaging study, they are general enough to have more broad applicability to many other settings. We describe the renal imaging study below.

1.2 Motivating Data

Radionuclide renal scans play an important role in the determination of kidney obstruction. When a kidney is obstructed, urine is unable to drain normally into the bladder; unless the obstruction is relieved and relieved in a timely manner, the kidney will lose its function and this loss of function become irreversible. Moreover, obstruction of a single kidney may be asymptomatic. Consequently, to preserve renal function, it is important to test for obstruction when there is a clinical suspicion. Radionuclide renal scans provide a non-invasive option and is usually the procedure of choice for evaluating suspected kidney obstruction. In academic institutions, nuclear scans are often interpreted by full time nuclear medicine physicians who had 36 months of training in their residencies. Unfortunately, a large percentage of the

radionuclide renal scans performed in the United States are interpreted at sites that perform fewer than 3 studies per week and are interpreted by private practice radiologists who have less than 4 months of training in all of nuclear medicine (IMV, 2003). Lack of training and limited experience coupled with the demands of interpreting a large variety of complex imaging studies at ever faster rates increases the error rate of the diagnosis (Taylor et al., 2008a).

In the Emory study, renal scans are acquired using a scintillation camera. Patients who are hydrated with 10 ounces of water are imaged in a supine position. Approximately 370 MBq (10 mCi) of Technetium-99m-mercaptoacetyltriglycine (MAG3), a gamma emitting tracer, is injected intravenously to the patient. Post injection, a scintillation camera, placed under the imaging table, is used to capture the photons emitted by the MAG3 as it is extracted from bloodstream by the kidneys, is transported into the kidney tubules and then travels down the ureters to the bladder. Multiple frames of data obtained during 24 minutes of image acquisition are recorded on a dedicated computer for subsequent processing. Regions of interest (ROI) are assigned over each kidney. The numbers of photons detected in each ROI in each frame are used to generate a time activity curve, called renogram or renal curve.

The process described above represents the baseline study for a patient. The baseline study for each patient resulting in a renogram is reviewed on site to determine if obstruction can be excluded. If obstruction can not be excluded in both kidneys, then a potent diuretic furosemide (Lasix) is injected intravenously and an additional set of images is acquired for 20 minutes. The average time between the injections of Tc-99m MAG3 and furosemide is greater than 30 minutes. The standard administered dose of furosemide is 40 mg but that can be increased if the patient is known to have elevated creatinine concentration or if the baseline study shows reduced MAG3 clearance (Bao et al., 2011; Taylor et al., 2008b). The post-furosemide component of the study resulting in another renogram is processed similarly to the baseline

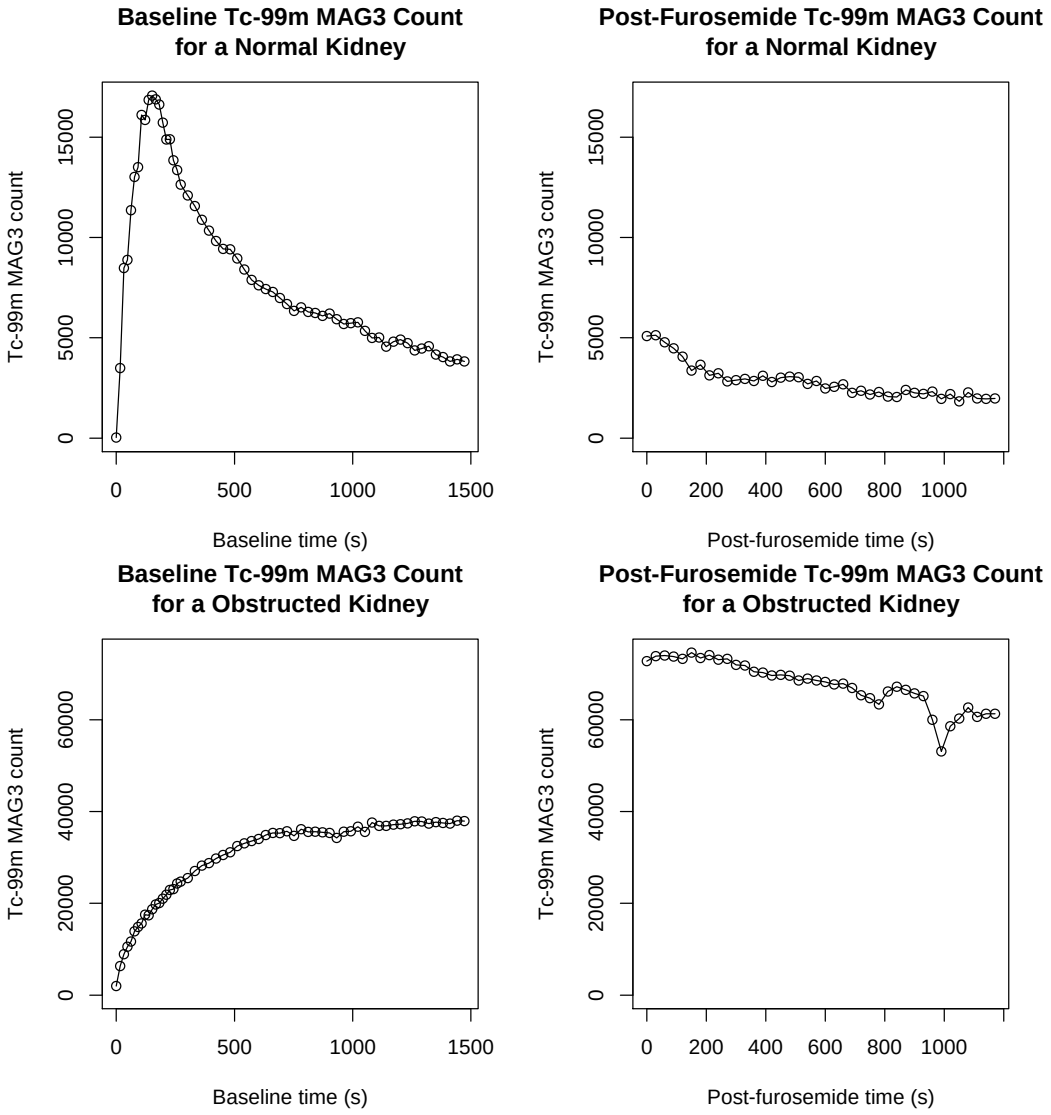

Figure 1.1: An Example of renal curves for a Normal Kidney (top panel) versus an Obstructed Kidney (bottom panel), with the Obstruction Status Determined by a Consensus Rating among 3 Experts

component study.

In summary, renal scan data consist of a baseline renal curve for each kidney. Some patients have received furosemide and there is a second post-furosemide acquisition. When the baseline scans of patients are interpreted by clinicians as being normal, no furosemide acquisition is obtained and post-furosemide data for these kidneys are missing. Figure 1.1 presents two renal curves from a patient when furosemide is given after the baseline scans. renal curve for a normal kidney (non-obstructed) is exempli-

fied by a steep increase in counts in the kidney ROI immediately after the injection of Tc-99m MAG3, followed by a sharp, then steady decrease in counts of the baseline renal curve (top-left figure in Figure 1.1). This declining trend of counts continues in the furosemide renal curve, although the rate of the decline, which signifies Tc-99m MAG3 clearance, decreases (top-right figure in Figure 1.1). Generally, the shape of the curve depends on the rate of uptake of MAG3 from the blood (clearance), the transport across the renal tubular cell, secretion of MAG3 into the renal tubule and drainage from the kidney into the bladder. An obstructed kidney has an abnormal baseline renal curve with ever-increasing photon counts (as exemplified by bottom-left figure of Figure 1.1). Following the injection of furosemide, the photon counts in an obstructed kidney (bottom-right Figure 1.1) has finally starting to decrease, but at this point, the photon count is already considerably higher than that of the normal kidney or even the peak of its own baseline component.

To assist radiologists in limiting their errors and making correct interpretation of MAG3 renal scans in their diagnoses, a study was conducted at Emory, with a goal to develop methodologies for decision support tools. This study consists of data from a large number of patients who were referred to the nuclear medicine clinic by with suspected kidney obstruction. Three experts defined as nuclear medicine physicians who each had more than 20 years of experience in full-time academic nuclear medicine, multiple publications in the area of renal nuclear medicine, and who had presented renal nuclear medicine educational sessions at national radiology and nuclear medicine meetings, were recruited to the study.

In this retrospective study, each expert independently scored each kidney for the presence of obstruction on a 5-point scale: 1, not obstructed; 2, probably not obstructed; 3, equivocal; 4, probably obstructed; and 5, obstructed. As a retrospective study, some patients did not receive furosemide because the baseline scan was judged to be normal by the practicing clinicians. In all, there were renogram data and expert

score data for 164 kidneys. The 164 sets of renal curves came from 76 patients and 81 distinct study dates. Ten kidneys lacked the post-furosemide renogram data.

1.3 Notations

To fix ideas, we let $\mathbf{Y}_{(m\times n)} = \begin{pmatrix} \mathbf{Y}_1 & \cdots & \mathbf{Y}_n \end{pmatrix}$ be the renal curve data, consisting of m time points for n individuals, where $\mathbf{Y}_i (i = 1, \cdots, n)$ is the ith individual's renal curve data taken at m specific time points. Furthermore, in the context of the renal study we define Δ_i as an indicator function, with $\Delta_i = 1$ when $\mathbf{Y}_i^{(2)}$ was observed and $\Delta_i = 0$ when $\mathbf{Y}_i^{(2)}$ was missing, where $\mathbf{Y}_{i(m_1\times 1)}^{(1)}$ represents baseline portion of $\mathbf{Y}_i$ and $\mathbf{Y}_{i((m_2)\times 1)}^{(2)}$ represent the post-furosemide portion ($m_1 + m_2 = m$). We write $\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} Y_{i1}^{(1)} & \cdots & Y_{im_1}^{(1)} & Y_{i1}^{(2)} & \cdots & Y_{im_2}^{(2)} \end{pmatrix}^T$. $\Delta = \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{pmatrix}$ as the vector of indicators for whether $\mathbf{Y}_i^{(2)}$ was observed for each kidney. Finally, we let $\mathbf{Y}_{obs}$ be the observed renal curve data, and $\mathbf{Y}_{mis} = \{\mathbf{Y}_i^{(2)} : \Delta_i = 0\}$ be the unobserved renal curve data.

1.4 Literature Review

1.4.1 Functional Data Analysis

Statistical analysis of data consisting of functions or surfaces is referred to as functional data analysis Ramsay and Dalzell (1991). The renal curves presented in Section 1.2 is an example of functional data. Functional data also include samples of hazard functions (Chiou and Müller, 2009) and density functions (Kneip and Utikal, 2001) and have wide-reaching areas of application, such as econometrics, education, genetics, evolutionary biology, chemometrics, medicine, and many others (Ramsay and

Silverman, 2002). Instead of trying to exhaustively define functional data analysis (FDA) by a set of methods and techniques, Ramsay and Silverman (2002) instead presented four aspects common to functional data, all of which can be seen in the renal curve data.

- First, functional data are continuously defined. This certainly applies to the renal curve data. While Tc-99m MAG3 counts were only measured at discrete time points, kidneys of patients who underwent the renal scan procedure never stopped filtering Tc-99m MAG3 from the bloodstream, nor did urine stop draining from the kidney via the ureters into the bladder. Consequently, the level of Tc-99m MAG3 in the kidneys is constantly changing, and at a steady pace, this means that the underlying renal curves are indeed continuously defined.

- Second, the individual datum is the whole function, rather than its value at any particular point. While we can safely assume that renal curves from different patients, or even from different kidneys, are independent of one another, different values within the same renal curve are certainly correlated to each other.

- Third, sometimes data are functions of time, but there is nothing special about time as a variable. The renal curves are indeed functions of time because they are measured at time intervals. Since the purpose of renal scans is to help physicians diagnose renal obstruction by monitoring how fast tracer transits the kidneys, it is sensible to consider the curves as functions of time.

- Finally, the data do not have to be smooth, but the analysis will often rely on smoothness of the data or other regularity (Ramsay and Silverman, 2002).

We can consider functional data as paths of a stochastic process. Suppose the continuous-time process $Y = \{Y(t), t \in [0, T]\}$ is defined on the space $(\Omega, \mathcal{A}, P)$.

Let Z be a random variable such as outcome. Assuming $\mathbb{E}\left[\int_0^T Y^2(t)\mathrm{d}t\right] < \infty$, one simple linear functional regression model for relating Z to $Y(t)$ is

$$Z = \int_0^T \psi(t)Y(t)\mathrm{d}t + \epsilon \tag{1.1}$$

where $\psi(t)$ is a square integrable function defined on $[0, T]$ and ϵ is a random variable with $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2$ and $\epsilon \perp\!\!\!\perp Y$.

In the context of renal studies, we can think of Z as a continuously observed response of interest, $Y(t)$ as the observed renal curve, and $\psi(t)$ as the function associating the observed curve with the latent renal obstruction status, and ϵ as the part of latent renal obstruction status due to chance. Many important research papers in the field of functional data are devoted to the estimation of (1.1) for predicting Z from $Y(t)$. In practice, the integral in equation (1.1) is often approximated by $\sum_{j=1}^{d} Y(t_j)a(t_j)$, where the curve Y is discretized at points $t_1, \cdots, t_d$. Such direct estimation of the regression coefficient function using least square criterion leads to the ill-conditioned regression problem of having many predictors with high degree of collinearity (Cardot et al., 2003). One solution proposed by Aguilera et al. (1997) and Cardot et al. (1999), known as principal component regression (PCR), utilizes elements derived from principal component analysis (PCA) of Y. PCR, however, does not come without issues of its own: when choosing principal components, one must choose between performance of the model and robustness of the model. An alternative to functional PCR is an extension of partial least squares (PLS) regression to incorporate a functional predictor, as proposed by Preda and Saporta (Preda and Saporta, 2005; Preda et al., 2010). Frank and Friedman (1993) explored PLS and PCR and gave an unifying approach of both and of ridge regression (RR) since all three constrain the coefficient vector in a linear regression model to some subspace. Cardot et al. (2003) argued these methods did not really take into consideration the

functional nature of the data.

Several other authors (Müller, 2005; Müller and Stadtmüller, 2005; James, 2002) have extended equation (1.1) to accomodate various types of outcomes including binary, counts, etc. via generalized functional linear model (GFLM). In GLM, random response variable Z (in the renal study, Z may be consensus of obstruction) with distribution

$$p(z;\eta,\phi) = \exp\left(\frac{z\eta - b(\eta)}{a(\phi)} - c(z,\phi)\right)$$

with predictor $\mathbf{Y}$ of finite dimensions and relationship

$$g(\mu) = \upsilon_0 + \boldsymbol{\upsilon}_1^T \mathbf{Y} \tag{1.2}$$

where $\mu = \mathbb{E}(Z;\eta,\phi) = b'(\eta)$ and $g(.)$ is the link function (McCullagh and Nelder, 1989). Examples of link functions include identity link for Gaussian response and logistic link for binomial response.

Just as GLM provides a flexible framework for relating response and predictor variables of finite dimensions (McCullagh and Nelder, 1989; James, 2002), it can be extended to handle functional predictors, which may contain different numbers of observations for each individual and be measured at different time points. This class of models can be referred to as generalized functional linear models.

When the predictor $Y(t), t \in \mathcal{C}$ is a random curve and corresponds to a square integrable stochastic process on some compact set $\mathcal{C}$ in $\mathbb{R}$, equation (1.2) cannot be applied directly. However, by replacing the summation over finite dimensional space with an integral over infinite dimensional space, we can generalize the mean model:

$$g(\mu) = \upsilon_0 + \int_{\mathcal{C}} Y(t)u(t)\mathrm{d}\omega(t) \tag{1.3}$$

where $u(.)$ is square integrable on $\mathcal{C}$, and $\mathrm{d}\omega$ is a real measure on $\mathcal{C}$, and υ_0 is fixed.

Unfortunately, $Y(t)$ is never observed at an infinite set of times in practice. Similar to the case with equation (1.1), simply replacing the integral with a summation over the observed time points may necessitate fitting an extremely high dimensional vector of coefficients, which can result in very large, potentially infinite, variance terms. This method also cannot handle individuals with different numbers of observations or individuals with observations taken from different sets of time pointsJames (2002). Instead, James (2002) proposed to use natural cubic splines (Silverman, 1985; Green and Silverman, 1993) to model the predictor with the assumption that $Y(t)$ can be modeled as a smooth curve from a given functional family, although Fourier transforms, orthogonal polynomial bases or any other finite dimensional basis can be substituted. Using $s(t)$ to represent the q-dimensional spline basis at time t, with $q-2$ knots when representing a q-dimensional natural cubic spline basis. $Y(t)$ can be reparameterized as

$$Y(t) = \mathbf{s}(t)^T\boldsymbol{\gamma}, \text{where } \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \Gamma) \tag{1.4}$$

where $\boldsymbol{\gamma}$ is the q-dimensional spline coefficients for the predictor with mean and variance being $\boldsymbol{\mu}_\gamma$ and Γ respectively (James, 2002). Combining equations (1.3) and (1.4) results in a mean model

$$g(\mu_i) = \upsilon_0 + \int_{\mathcal{C}} u(t)\mathbf{s}(t)^T\boldsymbol{\gamma}_i \mathrm{d}\omega(t) = \upsilon_0 + \boldsymbol{\upsilon}_1^T\boldsymbol{\gamma}_i$$

where $\boldsymbol{\upsilon}_1 = \int_{\mathcal{C}} u(t)\mathbf{s}(t)\mathrm{d}\omega(t)$. With the further assumption that at each time t, we observe $y(t)$ instead of $Y(t)$ where, $y(t) = Y(t) + \epsilon(t)$, where $\epsilon(t)$ is modeled as a zero-mean stationary Gaussian process and represents measurement error or other factors that would cause the observations to deviate from the spline fit. Let vectors of observations and measurement errors for individual i at times $t_{i1}, \cdots, t_{in_i}$ be represented by $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_i$ and let $S_i = (s(t_{i1}), \cdots, s(t_{in_i}))^T$ be the spline basis matrix

corresponding to $\mathbf{y}_i$, then the FGLM can be written as

$$p(z_i;\eta_i,\phi) = \exp\left(\frac{z_i\eta_i - b(\eta_i)}{a(\phi)} - c(z_i,\phi)\right),$$
$$g(\mu_i) = \upsilon_0 + \boldsymbol{\upsilon}_1^T\boldsymbol{\gamma}_i, \text{where } \boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \Gamma),$$
$$\mathbf{y}_i = S_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), i = 1, \cdots, N$$

where N represents the number of response and predictor pairs that are observed- James (2002).

James (2002) used an Estimation-Maximization (EM) algorithm (Dempster et al., 1977; Laird and Ware, 1982) to fit the FGLM. Depending on the distributions of z_i, there might not be a closed form equation for calculating the expected values and variance of the $\boldsymbol{\gamma}_i$'s given $\mathbf{y}_i$ and current estimates of other parameters. Therefore, James proposed to use Monte Carlo approach for calculating the expected values and variances (James, 2002).

While there is extensive literature regarding FDA, not much work has been done when missing data is present. Therefore one focus of this dissertation is to develop methods for handling missing data in FDA. We provide a brief summary of missing data below.

1.4.2 Missing Data

Researchers in public health and clinical research are often faced with datasets with incomplete data. There are many possible causes that may contribute to missing data. The subject may refuse to disclose certain sensitive information such as income or drug usage. The respondent may feel that none of the choices apply to him or her. The cause may also be purely accidental: the equipment used to measure and record the data may have suffered mechanical failure during data collection and suddenly stopped working.

Depending on the mechanism that led to the missing data, different methods can be applied to analyze the data. The missing-data mechanism is characterized by the conditional distribution of Δ given Y. In the context of renal studies, Δ is the vector of missing-data indicators, and Y is the renal curve data, with the observed components of Y defined as Y_{obs} and the missing components as Y_{mis}. We denote the parameters of interest as θ and other unknown parameters as ϕ. If missingness depends on the missing values in the data Y, then the mechanism is called "missing not at random" (MNAR). If missingness depends only on the observed components of Y and not on the missing components of Y,

$$f(\Delta|Y,\phi) = f(\Delta|Y_{obs},\phi) \text{ for all } Y_{mis},\phi$$

then the missing-data mechanism is "missing at random" (MAR). MAR assumption is less restrictive than "missing completely at random" (MCAR), which is

$$f(\Delta|Y,\phi) = f(\Delta|\phi) \text{ for all } Y,\phi,$$

when the missingness is assumed to not depend on any components of Y (Little and Rubin, 2002).

For maximum likelihood inference, the missing-data mechanism is ignorable if the data are MAR and the parameters θ and ϕ are distinct, that is, the joint parameter space of (θ,ϕ), $\Omega_{\theta,\phi}$, is the product of the parameter space of θ, Ω_θ, and the parameter space of ϕ, Ω_ϕ. Little and Rubin (2002) stated that MAR is typically viewed as the more important condition because if the data are MAR but does not have distinct parameters θ and ϕ, then inferences based on ignorable likelihood are still valid but not fully efficient. Similarly for Bayes inference,

$$p(\theta,\phi|Y_{obs},\Delta) \propto p(\theta,\phi)f(Y_{obs},\Delta|\theta,\phi)$$

when the data are MAR and parameters θ and ϕ are *a priori* independent, that is $p(\theta, \phi) = p(\theta)p(\phi)$, then

$$\begin{aligned} p(\theta, \phi | Y_{obs}, \Delta) &\propto [p(\theta)L(Y_{obs}|\theta)][p(\phi)L(\Delta|Y_{obs}, \phi)] \\ &\propto p(Y_{obs}|\theta)p(\Delta|Y_{obs}, \phi) \end{aligned}$$

the missing-data mechanism is ignorable in Bayesian inference and inferences about θ can be based on $p(\theta|Y_{obs})$. This is a stronger definition of ignorable than the one for ML inference because for θ and ϕ to have independent priors requires their parameter spaces to be distinct (Little and Rubin, 2002).

Complications can arise from the process that created missing data (Little and Rubin, 2002), but if the missing-data mechanism is ignorable, then valid inferences can still be made based on the data. For example, let $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$ denote the density of the joint distribution of Y_{obs} and Y_{mis}, then

$$f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta) \mathrm{d}Y_{mis}$$

is the marginal density of Y_{obs} obtained by integrating out Y_{mis}. Let $L_{ign}(\theta|Y_{obs})$ be any function of θ based on Y_{obs} such that

$$L_{ign}(\theta|Y_{obs}) \propto f(Y_{obs}|\theta) \tag{1.5}$$

inferences about θ based on $L_{ign}(\theta|Y_{obs})$ are valid as long as the mechanism leading to incomplete data are ignorable. Bayes inferences for θ with prior $p(\theta)$ can be then based on the posterior distribution

$$p(\theta|Y_{obs}) \propto p(\theta)L_{ign}(\theta|Y_{obs})$$

while maximum likelihood (ML) estimates can be obtained by maximizing (1.5) with respect to θ.

More generally, if we include the the distribution of the missing-data indicator Δ in our model, then the density of the joint distribution of Δ and Y can be expressed as

$$f(Y, \Delta|\theta, \phi) = f(Y|\theta)f(\Delta|Y, \phi)$$

that is, the product of the density of Y and the conditional distribution of Δ given Y. The actual observed data (Y_{obs}, Δ) will therefore have the density

$$f(Y_{obs}, \Delta|\theta, \phi) = \int f(Y_{obs}, Y_{mis}|\theta)f(\Delta|Y_{obs}, Y_{mis}, \phi)\mathrm{d}Y_{mis}$$

under MAR, this reduces to

$$\begin{aligned} f(Y_{obs}, \Delta|\theta, \phi) &= f(\Delta|Y_{obs}, \phi) \int f(Y_{obs}, Y_{mis}|\theta)\mathrm{d}Y_{mis} \\ &= f(\Delta|Y_{obs}, \phi)f(Y_{obs}|\theta) \end{aligned}$$

Little and Rubin (2002) grouped methods for analyzing partially missing data into four non-mutually exclusive categories:

Procedures based on completely recorded units, which is the easiest of the four and simply involves discarding the incomplete units and analyze only units with complete data. These methods are usually not very efficient and only appropriate when the data is MCAR, otherwise they will lead to biases.

Weighting procedures, which weight sampled units by their design weights adjusted for nonresponse, which are inversely proportional to the product of their probability of selection and an estimate of their probability of response.

Imputation-based procedures involves filling in the missing data followed by stan-

dard complete data analyzing methods. Common imputation methods include hot deck imputation, mean imputation, and regression imputation (Little and Rubin, 2002). Hot deck imputation replaces missing values with values from observed units in the sample. Mean imputation uses the average of observed values from a variable in the sample to replace missing values of that variable. Regression imputation replaces missing value for a variable of a unit by the predicted value based on a regression on the known variables for that unit.

Model-based procedures, which are characterized by having a model for the observed data as the basis for parameter estimations and inferences. Model-based procedures are flexible, not ad hoc, and incorporate uncertainty from incompleteness of the data.

Multiple imputation (MI), which is both imputation-based and model-based, was originally designed for complex surveys used to create public-use datasets, but has been proven to be useful in other settings as well (Harel and Zhou, 2007; Little and Rubin, 2002). Single imputation methods enable the user to use complete-data procedures on the imputed dataset, but the procedures often do not yield statistically valid results because of the underlying assumption of single imputation that the imputed value is the true value (Harel and Zhou, 2007). This limitation of single imputation methods led to underestimation of the variance, which in turn affects confidence intervals and statistical tests (Harel and Zhou, 2007).

MI consists of three stages: the imputation stage where missing data are imputed; the analysis stage where complete-case methods are used to analyze each complete dataset separately; finally, the combining stage where results from all datasets are combined to produce one set of result with adjustments for uncertainty from observed and missing data. Since quality of imputation is crucial, the imputation model should be as general and objective as possible and should always be at least as rich as the analysis model (Harel and Zhou, 2007; Schafer, 2003; Collins et al., 2001), otherwise,

the analysis results may be biased Meng (1994).

The MI procedures (Rubin, 1987) begins with M imputed independent versions of the missing data from the posterior predictive distribution $p(Y_{mis}|Y_{obs}, \Delta)$ under the joint model for Y and Δ. Under the ignorable assumption, this simplifies to $p(Y_{mis}|Y_{obs})$. Under Bayesian framework, let $p(Y|\theta)$ denote the model for the complete data with unknown parameter of interest θ and prior $p(\theta)$, then the posterior distribution of θ is

$$p(\theta|Y_{obs}) \propto p(\theta) \int p(Y_{obs}, Y_{mis}|\theta) \mathrm{d}Y_{mis} \tag{1.6}$$

and the posterior predictive distribution for Y_{mis} is

$$p(Y_{mis}|Y_{obs}) = \int p(Y_{mis}|Y_{obs}, \theta) p(\theta|Y_{obs}) \mathrm{d}\theta$$

For the imputation step, we simply repeat $m = 1, \cdots, M$ times where we first draw $\theta^{(m)}$ from $p(\theta|Y_{obs})$ then draw $Y_{mis}^{(m)}$ from $p(Y_{mis}|Y_{obs}, \theta^{(m)})$. By redrawing $\theta^{(m)}$ each time, we ensure that our imputation of $Y_{mis}^{(m)}$ is proper. Markov Chain Monte Carlo technique may be needed if (1.6) is complex (Schafer, 1997).

Following standard complete-data methods, we obtain estimates $(\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(M)})$ and squared standard errors $(U^{(1)}, \cdots, U^{(M)})$. We can now combine them following Rubin's rules (Rubin, 1987) where the overall estimate of θ

$$\bar{\theta} = M^{-1} \sum \hat{\theta}^{(m)}$$

and the estimated total variance of θ

$$T = (1 + M^{-1})B + \bar{U}$$

where $\bar{U} = M^{-1} \sum U^{(m)}$ is the within-imputation variance and $B = (M-1)^{-1} \sum (\hat{\theta}^{(m)} - \bar{\theta})^2$ is the between-imputation variance. Confidence intervals and tests are based on the

Student's t approximation $(\bar{\theta}-\theta)/\sqrt{T} \sim t_v$ where $v = (M-1)[T/1+M^{-1})B]^2$ is the degrees of freedom.

As described earlier, complete renal studies consist of two functional data sets, one of which is sometimes missing. Since the reason for patients who are missing the post-furosemide renal curves was due to their baseline renal scan results, the missing-data mechanism is MAR. The renal study consist of measurements taken at 99 time points per kidney. MI methods (and other missing data procedures) that treat each time point as a separate variable will either have to choose between working with large dimensions of variance-covariance matrices or poor performance by discarding information that can be gained from time points that are further apart. Such methods also cannot handle cases where subjects do not have measurements taken at the same time points as others. However, none of these methods account for the *functional* nature of the data.

1.4.3 Missing Data in Functional Data Analysis

There has been limited work dealing with missing data for functional variables. Preda et al. (2010) used the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm to impute missing data by their estimates derived from the approximated principal components and principal factors. However, this method has some limitations. Since NIPALS algorithm is based on PCA, their imputation method does not fully take advantage of the functional nature of the data. The performance of their method also depends on the number of principal components being chosen. They also did not account for the uncertainty associated with imputed data. The authors also stated that their algorithm's performance in the presence of missing data under MAR depends on the distribution of the functional data, the degree of linear dependence between elements of the functional data, as well as the sample size relative to the number of points per curve (Preda et al., 2010). Our goal in this dissertation is

to develop methods that overcome some of the above limitations in functional data analysis.

1.5 Statistical Problems

As we have previously stressed, only limited work has been done in handling missing data in functional data analysis. Within the context of renal studies, baseline renal curves are available, but post-furosemide renal curves may or may not be available. If the post-furosemide renal curves are available, then the radiologist would have a more complete set of tools to diagnose the patient's renal obstruction status. In a nuclear medicine department, time and equipment use are expensive. Two sets of renal scans cost the imaging center almost twice as much as one set. Consequently, if the second set can be reliably imputed from the first set, then the second set can be omitted and the cost to the imaging center for the second set (use of the equipment, cost of furosemide, and extra technologist time) can be avoided. This leads to greater efficiency and substantial cost savings. We intend to develop a method that can impute the missing post-furosemide renal curves based on available renal curve data. The advantages of our method are that it is not *ad hoc*, it is simple to implement, it incorporates the uncertainty associated with imputation, and it imputes the entire curve, which allows the radiologist to extract post-furosemide renal curve data at any given time point or to calculate any summary measures based on post-furosemide renal curve alone or based on both curves. We will describe this method in full detail in Chapter 2.

Another problem of interest in renal studies is to determine the association between renograms and renal obstruction. Towards this goal, we intend to develop another method which simultaneously imputes the missing post-furosemide renal curves and use the available curve data and the imputed curves as predictors and the con-

sensus ratings for renal obstruction as response to build a model. This model can help radiologists identify the nature of association between the renograms and renal obstruction. We will describe this method in further detail in Chapter 3.

In Chapter 3, we fit joint models with different numbers of cubic B-spline basis functions and use deviance information criteria to select the optimal number of basis functions. The selection is computationally intensive because multiple joint models are fitted. It is very challenging to adaptively choose the knot locations for defining the cubic B-spline basis functions. To address these limitations, we investigate basis function selection via functional principal component analysis in Chapter 4. We then compare the performance of the joint model fit with cubic B-spline basis against the joint model fit with alternative basis.

In Chapter 5, we provide directions for some future work.

Chapter 2

Multiple imputation of functional data with application to renal studies

2.1 Introduction

Statistical analyses of functional data have received increased attention in recent years. Functional data consists of curves or surfaces, hazard functions (Chiou and Müller, 2009) and density functions (Kneip and Utikal, 2001), and have wide-reaching areas of application, such as econometrics, education, genetics, evolutionary biology, chemometrics, medicine, and many others (Ramsay and Silverman, 2002). Functional data are often encountered in biomedical studies. They are either collected at specific time points separated by fixed intervals or at convenient time points within a time frame.

We can consider functional data as paths of a stochastic process. Suppose the continuous-time process $Y = \{Y(t), t \in [0, T]\}$ is defined on the space $(\Omega, \mathcal{A}, P)$. Let Z be a random variable such as outcome. Assuming $\mathbb{E}\left[\int_0^T Y^2(t)\mathrm{d}t\right] < \infty$, one simple linear functional regression model for relating Z to $Y(t)$ is

$$Z = \int_0^T \psi(t)Y(t)\mathrm{d}t + \epsilon \tag{2.1}$$

where $\psi(t)$ is a square integrable function defined on $[0, T]$ and ϵ is a random variable with $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2$ and $\epsilon \perp\!\!\!\perp Y$.

In practice, the integral in equation (2.1) is often approximated by $\sum_{j=1}^{d} Y(t_j)a(t_j)$, where the curve Y is discretized at points $t_1, \cdots, t_d$. Such direct estimation of the regression coefficient function using least square criterion leads to the ill-conditioned regression problem of having many predictors with high degree of collinearity (Cardot et al., 2003). One solution proposed by Aguilera et al. (1997) and Cardot et al. (1999), known as principal component regression (PCR), utilizes elements derived from principal component analysis (PCA) of Y. PCR, however, does not come without issues of its own: when choosing principal components, one must choose between performance of the model and robustness of the model.

Several other authors(Müller, 2005; Müller and Stadtmüller, 2005; James, 2002) extended equation (2.1) to accomodate various types of outcomes including binary, counts, etc. via generalized functional linear model (GFLM). Just as GLM provides a flexible framework for relating response and predictor variables of finite dimensions (McCullagh and Nelder, 1989; James, 2002), it can be extended to handle functional predictors, which may contain different numbers of observations for each individual and be measured at different time points.

The works mentioned above are limited to complete observations and missing data are not considered. How to handle missing data poses a notable obstacle in the functional data analysis of biomedical data. Prevalent treatment of missing data is to discard subjects with missing data before analysis, this is undesirable as valuable information contained in the discarded data are lost, which can potentially bias the results.

One example of functional data with missing data can be found in a renal studay, which motivates this work. Renal scans play an important role in determining renal obstruction, a condition that if not treated in a timely manner, will lead to irreversible loss of function of the kidney. Consequently, to preserve renal function, it is important to test for obstruction when there is a clinical suspicion. When a patient known or suspected to have renal obstruction is referred to a nuclear medicine clinic, renal scans are performed to help the radiologist diagnose obstruction in the patient's kidneys. The patient receives an injection of Technetium-99m-mercaptoacetyltriglycine (MAG3), a radioactive tracer. A scintillation camera was used to capture the photons emitted by the MAG3 as it travels from the bloodstream through the kidneys and eventually to the bladder. Multiple frames of data obtained during 24 minutes of image acquisition were recorded on a dedicated computer for subsequent processing. The numbers of photons detected in each kidney in each frame were used to generate a time activity curve or renogram curve. The baseline study for each patient was re-

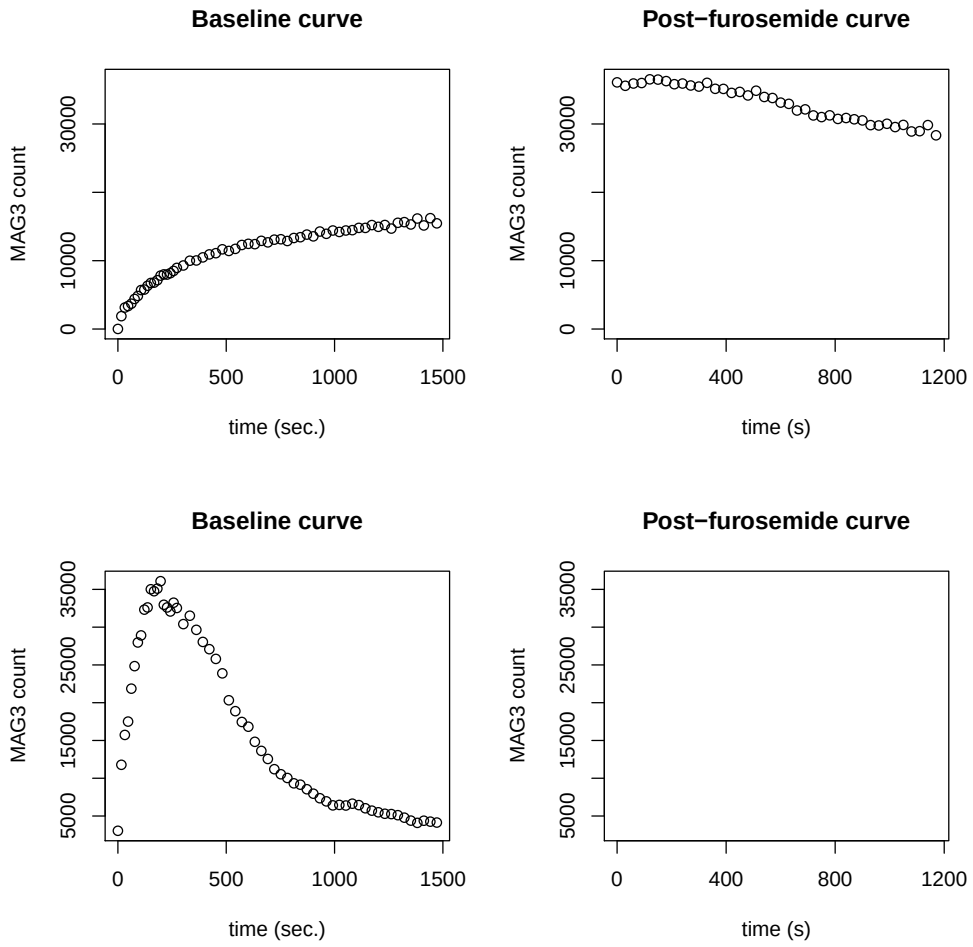

Figure 2.1: Examples of renogram curves for a kidney with both curves and a kidney with only baseline curve

viewed on site to determine if obstruction could be excluded. If obstruction could not be excluded in both kidneys, then a potent diuretic furosemide (Lasix) was injected intravenously and an additional set of images was acquired for 20 minutes. The average time between the injections of Tc-99m MAG3 and furosemide was greater than 30 minutes (Taylor et al., 2008b). The furosemide component of the study was processed similarly to the baseline component to form the furosemide renogram. This creates a situation where some patients have kidneys with baseline (first) and post-furosemide (second) curves while others only have kidneys with the first curve (Fig. 2.1).

In this Chapter, we propose two multiple imputation methods that utilize func-

tional linear models (FLMs) for the first and second curves, and use the Bayesian data augmentation algorithm for their implementation. Our first proposed method requires *a priori* specification of the number of basis functions. In our second proposed method, we use a latent factor model for the basis coefficients of the FLM and apply a shrinkage prior (Bhattacharya and Dunson, 2011) on the basis coefficients so that *a priori* selection of number of basis functions in the FLM is not required, as the need for specifying the number of factors *a priori* must balance between missing important factors and wasting computation on overly conservative estimation of number of factors. Both methods account for the functional nature of the renal curve data. Our methods do not require any additional information about the outcome, so the imputed curves are not biased towards a particular outcome. Our methods allow all subjects to be analyzed, which is a huge advantage over complete case analysis. The uncertainty associated with imputed data can be accounted for by applying Rubin's rules post analysis. In the next section we introduce the data structure and describe our imputation methods in detail. In Section 2.3 we assess our methods by two simulation studies. In Section 2.4 we present an application to data from a renal study. A few concluding remarks are given in Section 2.5.

2.2 Methodology

Let $\mathbf{Y}_{i(m_i\times n)} = \begin{pmatrix} Y_1 & \cdots & Y_{m_i} \end{pmatrix}$ be, in the context of renal study, measurements taken at m_i time points for subject i. Let $\mathbf{Y}^{(1)}_{i(m_{i1}\times 1)}$ be the first curve and $\mathbf{Y}^{(2)}_{i((m_{i2})\times 1)}$ be the second curve, $(m_{i1}+m_{i2} = m_i)$. We write $\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} Y_{i1}^{(1)} & \cdots & Y_{im_{i1}}^{(1)} & Y_{i1}^{(2)} & \cdots & Y_{im_{i2}}^{(2)} \end{pmatrix}^T$.

Let $\Delta = \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{pmatrix}$ be the vector of indicator functions with $\Delta_i = 1$ if $\mathbf{Y}_i^{(2)}$ is observed.

Finally, we let $\mathbf{Y}_{obs}$ represent the observed curve data, and $\mathbf{Y}_{mis} = \{\mathbf{Y}_i^{(2)} : \Delta_i = 0\}$ represent the unobserved curve data.

Given we sample an individual curve with error over distinct time points $t_{i1}, \cdots, t_{im_i}$ (James and Hastie, 2001), we use the following model for the curve:

$$Y_{ij}^{(s)} = \theta_i^{(s)}(t_{ij}) + \epsilon_{ij}^{(s)},\ i = 1, \cdots, n,\ j = 1, \cdots, m_{is},\ s = 1, 2 \tag{2.2}$$

where $Y_{ij}^{(s)}$ is the observed value of the s^{th} curve for kidney i at time point t_{ij}, $\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_i^{(1)} \\ \boldsymbol{\epsilon}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \epsilon_{i1}^{(1)} & \cdots & \epsilon_{im_{i1}}^{(1)} & \epsilon_{i1}^{(2)} & \cdots & \epsilon_{im_{i2}}^{(2)} \end{pmatrix}^T$ represents measurement errors, $\theta_i^{(s)}(t_{ij})$ represents the true value of the renal curve for kidney i at time point t_{ij}. We model $\theta_i^{(s)}(t_{ij})$ as

$$\theta_i^{(s)}(t_{ij}) = \sum_{l=1}^{k_s} \beta_{il}^{(s)} b_l^{(s)}(t_{ij}) \tag{2.3}$$

where $\{b_l(t_{ij}), l = 1, \cdots, k_s\}$ are the cubic spline basis functions ($k_1 + k_2 = k$). The number of knots used in the cubic spline basis functions is $k_s - 3$. We employ the cubic spline model for the true curve for the following reasons. First, the coefficients which define the polynomial form of the cubic spline model can be found by solving a system of equations for which stable and fast numerical algorithms already exist (Silverman, 1985). Second, in a cubic spline model, observation at t_{ij} only has a fast-decreasing influence on nearby parts, which makes it favorable to other curve-fitting methods such as polynomial regression (Silverman, 1985). Third, the dependence of the local bandwidth on the density of observed time points in the cubic spline model is intermediate between fixed kernal smoothing and smoothing based on neighboring values, which is desirable because moving from fixed kernel to nearest neighbor methods resulted in overfitting (Silverman, 1984, 1985).

We let $\mathbf{B}_{i(m_i \times k)}$ be the block diagonal matrix that consists of the cubic spline

basis functions of first and second curve for kidney i, specifically, $\mathbf{B}_i = \begin{pmatrix} \mathbf{B}_i^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_i^{(2)} \end{pmatrix}$ where $\mathbf{B}_i^{(1)}$ corresponds to the first curve and $\mathbf{B}_i^{(2)}$ corresponds to the second curve, and

$$\mathbf{B}_i^{(s)} = \begin{pmatrix} b_1^{(s)}(t_{i1}) & \cdots & b_{k_s}^{(s)}(t_{i1}) \\ \vdots & \ddots & \vdots \\ b_1^{(s)}(t_{im_{is}}) & \cdots & b_{k_s}^{(s)}(t_{im_{is}}) \end{pmatrix}, s = 1, 2$$

We write the corresponding coefficients for $\mathbf{B}_i$ as $\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_i^{(1)} \\ \boldsymbol{\beta}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \beta_{i1}^{(1)} & \cdots & \beta_{ik_1}^{(1)} & \beta_{i1}^{(2)} & \cdots & \beta_{ik_2}^{(2)} \end{pmatrix}^T$. Using $\mathbf{B}_i$ and $\boldsymbol{\beta}_i$ defined above, we can rewrite (2.2) and (2.3) as

$$\mathbf{Y}_i = \mathbf{B}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \tag{2.4}$$

2.2.1 Fixed number of knots multiple imputation method (FK)

In our first proposed method, the fixed number of knots multiple imputation method (FK), we assume the optimal number of basis functions used to model the curves is known. We assume the coefficients for these basis functions $\boldsymbol{\beta}_i$ follow a multivariate normal distribution

$$\boldsymbol{\beta}_i \sim N_k(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) \tag{2.5}$$

based on some hyperparameters $\boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}_0^{(1)} \\ \boldsymbol{\beta}_0^{(2)} \end{pmatrix}$ and $\boldsymbol{\Sigma}_\beta = \begin{pmatrix} \boldsymbol{\Sigma}_\beta^{11} & \boldsymbol{\Sigma}_\beta^{12} \\ \boldsymbol{\Sigma}_\beta^{21} & \boldsymbol{\Sigma}_\beta^{22} \end{pmatrix}$.

We also assume the following model for the measurement error,

$$\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \sigma^2 \mathbf{I}_m) \tag{2.6}$$

with σ^2 as its variance, which we assume to be constant across all subjects, and

that $\boldsymbol{\beta}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ since true curves does not impact measurement errors. We assume the hyperparameters $\boldsymbol{\beta}_0$ and σ^2 to have non-informative and improper priors

$$\pi(\boldsymbol{\beta}_0, \sigma^2) \propto \mathbf{1}$$

and the hyperparameter Σ_β to have an Inverse-Wishart prior

$$\boldsymbol{\Sigma}_\beta \sim \mathcal{W}^{-1}(\mathbf{I}_k, k)$$

where $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ has scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν. These assumptions about the prior specifications of the hyperparameters help facilitate our imputation procedure.

When $\mathbf{Y}$ is completely observed, the posterior distribution is

$$\begin{aligned}
&P\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2 | \mathbf{Y}, \mathbf{B}\right) \\
&\propto \sigma^{-nm} |\boldsymbol{\Sigma}_\beta|^{-\frac{n+2k+1}{2}} \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{B}\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mathbf{B}\boldsymbol{\beta}_i)\right.\right. \\
&\left.\left. + \sum_{i=1}^{n}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_\beta^{-1}\right)\right\}\right]
\end{aligned} \tag{2.7}$$

However, for some individuals, only $\mathbf{Y}_i^{(1)}$ is observed, $\mathbf{Y}_i^{(2)}$ is missing, under ignorable missing data mechanism, the posterior distribution is

$$\begin{aligned}
&P\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2 | \mathbf{Y}, \mathbf{B}, \boldsymbol{\Delta}\right) \\
&= P(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2 | \mathbf{Y}_{obs}, \mathbf{B}, \boldsymbol{\Delta}) \\
&\propto \pi(\boldsymbol{\beta}_0, \sigma^2)\pi(\boldsymbol{\Sigma}_\beta) \prod_{i=1}^{n} \Big\{ P(\mathbf{Y}_i | \boldsymbol{\beta}_i, \sigma^2, \mathbf{B}) \pi(\boldsymbol{\beta}_i | \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) I(\Delta_i = 1) \\
&\quad + P(\mathbf{Y}_i^{(1)} | \boldsymbol{\beta}_i^{(1)}, \sigma^2) \pi(\boldsymbol{\beta}_i^{(1)} | \boldsymbol{\beta}_0^{(1)}, \boldsymbol{\Sigma}_\beta^{11}) I(\Delta_i = 0) \Big\}
\end{aligned}$$

where $I(.)$ is the indicator function.

A common approach for multiple imputation is by using the Bayesian data augmentation method (Tanner and Wong, 1987), which is an iterative approach similar to EM algorithm but where the estimation and maximization steps are replaced by imputation (I) and posterior (P) steps. The algorithm we use is described below.

I step we draw $\mathbf{Y}_{mis}$ from its posterior predictive distribution (PPD) $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{B}, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^{2(t)})$. By (2.4) and (2.6),

$$\begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} |\boldsymbol{\beta}_i, \mathbf{B}, \sigma^2 \sim N_m \left(\begin{pmatrix} \mathbf{B}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_i^{(1)} \\ \boldsymbol{\beta}_i^{(2)} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{I}_{m_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m_2} \end{pmatrix} \right) \tag{2.8}$$

from (2.8) we can see that given $\mathbf{B}^{(2)}, \boldsymbol{\beta}_i^{(2)}$, and $\sigma^2, \mathbf{Y}_i^{(2)}$ does not depdend on $\mathbf{Y}_i^{(1)}, \mathbf{B}^{(1)}, \boldsymbol{\beta}_i^{(1)}$, then the posterior predictive distributions for drawing each $\mathbf{Y}_i^{(2)}$ where $\Delta_i = 0$ is

$$\mathbf{Y}_i^{(2)}|\mathbf{B}^{(2)}, \boldsymbol{\beta}_i^{(2)}, \sigma^2 \sim \mathcal{N}_{m_2} \left(\mathbf{B}^{(2)} \boldsymbol{\beta}_i^{(2)}, \sigma^2 I_{m_2} \right) \tag{2.9}$$

P step we treat $(\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ as $\mathbf{Y}$ in (2.7), and because of our assumptions about the prior specifications, we can draw $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n$, $\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_\beta$, and σ^2 from their respective full conditional distributions,

$$\boldsymbol{\beta}_i|. \sim N_k \left(\left\{ \frac{\mathbf{B}^T\mathbf{B}}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1} \left\{ \frac{\mathbf{B}^T\mathbf{Y}_i}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}_0 \right\}, \left\{ \frac{\mathbf{B}^T\mathbf{B}}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1} \right) \tag{2.10}$$

$$\boldsymbol{\beta}_0|. \sim N_k \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}_i, \frac{\boldsymbol{\Sigma}_\beta}{n} \right) \tag{2.11}$$

$$\boldsymbol{\Sigma}_\beta|. \sim \mathcal{W}^{-1} \left(\sum_{i=1}^{n} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)^T + \mathbf{I}_k, n + k \right) \tag{2.12}$$

$$\sigma^2|. \sim Inverse\text{-}Gamma\left(\frac{nm}{2}-1, \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{B}\boldsymbol{\beta}_i)^T(\mathbf{Y}_i-\mathbf{B}\boldsymbol{\beta}_i)\right) \quad (2.13)$$

where . represents conditioning on data and current estimates of all other unknown parameters.

We repeat this iterative procedure until the algorithm converges. After convergence, this iterative procedure yields draws from the joint posterior distribution given $\mathbf{Y}_{obs}, \mathbf{B}_1, \cdots, \mathbf{B}_n, \boldsymbol{\Delta}$. Any draws of $\mathbf{Y}_{mis}$ after the convergence of the algorithm can be combined with $\mathbf{Y}_{obs}$ to form an imputed dataset.

2.2.2 Sparse latent factor multiple imputation method (SLF)

In practice, we do not know the optimal number of basis functions for fitting the true curves. One strategy is to over-specify the number of basis functions to maximize fit, normally, this is computationally intensive and inefficient. To overcome these deficiencies, we use a sparse latent factor model (Montagna et al., 2012) for the subject-specific basis coefficients $\boldsymbol{\beta}_i$,

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \quad (2.14)$$

where $\boldsymbol{\Lambda}$ is a $k \times p$ factor loading matrix, $\boldsymbol{\eta}_i$ is a vector of continuous latent variables for subject i, and $\boldsymbol{\zeta}_i$ is a residual vector uncorrelated with other variables in the model. We assume $\boldsymbol{\zeta}_i$ follows a Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = diag\left(\sigma^2_{\zeta 1}, \sigma^2_{\zeta 2}, \cdots, \sigma^2_{\zeta k}\right)$. The latent variables $\boldsymbol{\eta}_i$ is related to covariates in the following manner

$$\boldsymbol{\eta}_i = \boldsymbol{\kappa}^T\mathbf{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\delta}_i \sim N_p(\mathbf{0}, \mathbf{I}) \quad (2.15)$$

where $\mathbf{x}_i$ is a $r \times 1$ vector of predictors for subject i such as age and sex, $\boldsymbol{\kappa}$ is a $r \times p$ matrix of coefficients and $\boldsymbol{\delta}_i$ is a normally distributed residual vector. We assume the same model for the measurement errors as (2.6).

For the hyperparameters, we assume σ^2 and $\sigma^2_{\zeta 1}, \cdots, \sigma^2_{\zeta k}$ to have priors

$$\sigma^{-2} \sim Gamma\left(a_\sigma, b_\sigma\right) \tag{2.16}$$

$$\sigma^{-2}_{\zeta j} \sim Gamma\left(a_{\zeta j}, b_{\zeta j}\right), j = 1, \cdots, k \tag{2.17}$$

The prior for $\mathbf{\Lambda}$ is a multiplicative gamma process shrinkage prior, with

$$\lambda_{jh}|\psi_{jh}, \tau_h \sim N\left(0, \psi^{-1}_{jh}\tau^{-1}_h\right), \psi_{jh} \sim Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tau_h = \prod_{l=1}^{h} \rho_l, j = 1, \cdots, k, h = 1, \cdots, p$$

$$\rho_1 \sim Gamma(a_1, 1), \rho_l \sim Gamma(a_2, 1), l \geq 2 \tag{2.18}$$

For the coefficents $\boldsymbol{\kappa}$, we assume a Gaussian prior such that for the jth column of $\boldsymbol{\kappa}$ denoted as $\boldsymbol{\kappa}_j$,

$$\boldsymbol{\kappa}_j \sim N\left(\mathbf{0}, \mathbf{\Sigma}_\kappa\right), j = 1, \cdots, p \tag{2.19}$$

where $\mathbf{\Sigma}_\kappa = diag(\omega^{-1}_{1j}, \cdots, \omega^{-1}_{rj})$ and

$$\omega_{lj} \sim Gamma(1/2, 1/2) \tag{2.20}$$

The posterior distribution for when $\mathbf{Y}$ is completely observed becomes

$$\begin{aligned} &P\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \mathbf{\Lambda}, \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n, \boldsymbol{\kappa}, \mathbf{\Sigma}, \sigma^2 | \mathbf{Y}_1, \cdots, \mathbf{Y}_n, \mathbf{B}_1, \cdots, \mathbf{B}_n, \mathbf{x}_1, \cdots, \mathbf{x}_n\right) \\ &\propto \pi(\sigma^2)\pi(\mathbf{\Sigma})\pi(\mathbf{\Lambda})\pi(\boldsymbol{\kappa}) \prod_{i=1}^{n} P\left(\boldsymbol{\eta}_i | \boldsymbol{\kappa}, \mathbf{x}_i\right) P\left(\boldsymbol{\beta}_i | \mathbf{\Lambda}, \boldsymbol{\eta}_i, \mathbf{\Sigma}\right) P\left(\mathbf{Y}_i | \mathbf{B}_i, \boldsymbol{\beta}_i, \sigma^2\right) \end{aligned} \tag{2.21}$$

and under ignorable missing data mechanism, the posterior distribution becomes

$$
\begin{aligned}
& P\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\Lambda}, \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n, \boldsymbol{\kappa}, \boldsymbol{\Sigma}, \sigma^2 | \mathbf{Y}_1, \cdots, \mathbf{Y}_n, \mathbf{B}_1, \cdots, \mathbf{B}_n, \mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\Delta}\right) \\
= & P\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\Lambda}, \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n, \boldsymbol{\kappa}, \boldsymbol{\Sigma}, \sigma^2 | \mathbf{Y}_{obs}, \mathbf{B}_1, \cdots, \mathbf{B}_n, \mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\Delta}\right) \\
\propto & \pi(\sigma^2)\pi(\boldsymbol{\Sigma})\pi(\boldsymbol{\Lambda})\pi(\boldsymbol{\kappa}) \prod_{i=1}^{n} \Big\{ P\left(\boldsymbol{\eta}_i | \boldsymbol{\kappa}, \mathbf{x}_i\right) P\left(\boldsymbol{\beta}_i | \boldsymbol{\Lambda}, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}\right) P\left(\mathbf{Y}_i | \mathbf{B}_i, \boldsymbol{\beta}_i, \sigma^2\right) I(\Delta_i = 1) \\
& + P\left(\boldsymbol{\eta}_i | \boldsymbol{\kappa}, \mathbf{x}_i\right) P\left(\boldsymbol{\beta}_i^{(1)} | \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}^{(1)}\right) P\left(\mathbf{Y}_i^{(1)} | \mathbf{B}_i^{(1)}, \boldsymbol{\beta}_i^{(1)}, \sigma^2\right) I(\Delta_i = 0) \Big\}
\end{aligned}
$$

where $I(.)$ is the indicator function, $\boldsymbol{\Lambda}^{(1)}$ is the upper $k_1 \times p$ portion of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}^{(1)} = diag\left(\sigma_{\zeta 1}^2, \cdots, \sigma_{\zeta k_1}^2\right)$.

The data augmentation procedure for our second proposed method, which we henceforth refer to as sparse latent factor multiple imputation method (SLF), is as follows:

I step we draw $\mathbf{Y}_{mis}$ from its posterior predictive distribution (PPD) $P(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}, \mathbf{B}_1, \cdots, \mathbf{B}_n, \mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\Lambda}, \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n, \boldsymbol{\kappa}, \boldsymbol{\Sigma}, \sigma^2)$. By (2.4), (2.3) and (2.6),

$$
\begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} | \boldsymbol{\beta}_i, \mathbf{B}_i, \sigma^2 \sim N_{m_i} \left(\begin{pmatrix} \mathbf{B}_i^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_i^{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_i^{(1)} \\ \boldsymbol{\beta}_i^{(2)} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{I}_{m_{i1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m_{i2}} \end{pmatrix} \right) \quad (2.22)
$$

from (2.22) we can see that given $\mathbf{B}_i^{(2)}, \boldsymbol{\beta}_i^{(2)}$, and $\sigma^2, \mathbf{Y}_i^{(2)}$ does not depdend on $\mathbf{Y}_i^{(1)}, \mathbf{B}_i^{(1)}, \boldsymbol{\beta}_i^{(1)}$, then the posterior predictive distributions for drawing each $\mathbf{Y}_i^{(2)}$ where $\Delta_i = 0$ is

$$
\mathbf{Y}_i^{(2)} | \mathbf{B}_i^{(2)}, \boldsymbol{\beta}_i^{(2)}, \sigma^2 \sim \mathcal{N}_{m_{i2}} \left(\mathbf{B}_i^{(2)} \boldsymbol{\beta}_i^{(2)}, \sigma^2 I_{m_{i2}} \right) \quad (2.23)
$$

The I step remains the same as in our first proposed method.

P step we treat $(\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ as $\mathbf{Y}$ in (2.21), and first draw $\boldsymbol{\beta}_i$ from its full conditional

posterior distribution

$$\boldsymbol{\beta}_i|. \sim N_k\left(\left\{\frac{\mathbf{B}_i^T\mathbf{B}_i}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\right\}^{-1}\left\{\frac{\mathbf{B}_i^T\mathbf{y}_i}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\eta}_i\right\},\left\{\frac{\mathbf{B}_i^T\mathbf{B}_i}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\right\}^{-1}\right) \tag{2.24}$$

To ensure numerical stability, we marginalize out $\boldsymbol{\beta}_i$,

$$\begin{aligned}\mathbf{y}_i =&\mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\eta}_i+\mathbf{B}_i\boldsymbol{\zeta}_i+\boldsymbol{\epsilon}_i,\boldsymbol{\zeta}_i\sim N(0,\boldsymbol{\Sigma}),\boldsymbol{\epsilon}_i\sim N(0,\sigma^2 I_{m_i})\\ =&\mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\eta}_i+\boldsymbol{\alpha}_i^*,\boldsymbol{\alpha}_i^*\sim N(0,\sigma^2 I_{m_i}+\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i^T)\end{aligned}$$

and draw $\boldsymbol{\eta}_i$ directly from

$$\boldsymbol{\eta}_i|. \sim N_{p^*}\left(\boldsymbol{\Sigma}_{\eta_i}^{-1}\left\{\boldsymbol{\kappa}^T\mathbf{x}_i+\boldsymbol{\Lambda}^T\mathbf{B}_i^T\left(\sigma^2\mathbf{I}_{m_i}+\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i^T\right)^{-1}\mathbf{y}_i\right\},\boldsymbol{\Sigma}_{\eta_i}^{-1}\right) \tag{2.25}$$

where $\boldsymbol{\Sigma}_{\eta_i}=\boldsymbol{\Lambda}^T\mathbf{B}_i^T\left(\sigma^2\mathbf{I}_{m_i}+\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i^T\right)^{-1}\mathbf{B}_i\boldsymbol{\Lambda}+\mathbf{I}_{p^*}$.

Next, we update $\boldsymbol{\Lambda}$ by drawing $\lambda_{jh},\rho_h,\psi_{jh},j=1,\cdots,k,h=1,\cdots,p^*$ from their respective full conditional distributions. We let $\boldsymbol{\lambda}_j$ be the jth row of $\boldsymbol{\Lambda}$, then

$$\boldsymbol{\lambda}_j|. \sim N_{p^*}\left(\left\{\mathbf{D}_j^{-1}+\frac{\boldsymbol{\eta}^T\boldsymbol{\eta}}{\sigma_{\zeta j}^2}\right\}^{-1}\frac{\boldsymbol{\eta}^T\boldsymbol{\beta}_{\cdot j}}{\sigma_{\zeta j}^2},\left\{\mathbf{D}_j^{-1}+\frac{\boldsymbol{\eta}^T\boldsymbol{\eta}}{\sigma_{\zeta j}^2}\right\}^{-1}\right) \tag{2.26}$$

where $\mathbf{D}_j^{-1}=diag(\psi_{j1}\tau_1,\cdots,\psi_{jp}\tau_p)$, $\boldsymbol{\eta}^{\boldsymbol{T}}=[\boldsymbol{\eta}_1,\cdots,\boldsymbol{\eta}_n]$ and $\boldsymbol{\beta}_{\cdot j}=(\beta_{1j},\cdots,\beta_{nj})$, for $j=1,\cdots,k$. Then we draw ψ_{jh} from

$$\psi_{jh}|. \sim Gamma\left(\frac{\nu+1}{2},\frac{\nu}{2}+\frac{\tau_h\lambda_{jh}^2}{2}\right) \tag{2.27}$$

draw ρ_1 from

$$\rho_1|. \sim Gamma\left(a_1+\frac{p^*k}{2},1+\frac{1}{2}\sum_{h=1}^{p^*}\tau_{h,-1}\sum_{j=1}^{k}\psi_{jh}\lambda_{jh}^2\right) \tag{2.28}$$

and ρ_l from

$$\rho_l|. \sim Gamma\left(a_2 + \frac{(p^* - l + 1)k}{2}, 1 + \frac{1}{2}\sum_{h=l}^{p^*}\tau_{h,-l}\sum_{j=1}^{k}\psi_{jh}\lambda_{jh}^2\right) \tag{2.29}$$

for $l \geq 2$, where $\tau_{h,-l} = \prod_{t=1,t\neq l}^{h}\rho_t$ for $l = 1, \cdots, p^*$.

We draw σ^{-2} from

$$\sigma^{-2}|. \sim Gamma\left(a_\sigma + \frac{1}{2}\sum_{i=1}^{n} m_i, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\beta}_0 - \mathbf{B}_i\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\beta}_0 - \mathbf{B}_i\boldsymbol{\beta}_i)\right) \tag{2.30}$$

and draw $\boldsymbol{\Sigma}^{-1}$ by drawing its diagonal elements $\sigma_{\zeta j}^{-2}, j = 1, \cdots, k$ from

$$\sigma_{\zeta j}^{-2}|. \sim Gamma\left(a_{\zeta j} + \frac{n}{2}, b_{\zeta j} + \frac{1}{2}\sum_{i=1}^{n}(\beta_{ij} - \boldsymbol{\lambda}_j\boldsymbol{\eta}_i)^2\right) \tag{2.31}$$

Finally, we update $\boldsymbol{\kappa}$ by first drawing $\omega_{lj}, l = 1, \cdots, r, j = 1, \cdots, p^*$ from

$$\omega_{lj}|. \sim Gamma\left(1, \frac{1}{2}(1 + \kappa_{lj}^2)\right) \tag{2.32}$$

where κ_{lj} is the ljth element of $\boldsymbol{\kappa}$, then we draw $\boldsymbol{\kappa}_j, j = 1, \cdots, p^*$ from

$$\boldsymbol{\kappa}_j|. \sim N\left(\left\{\mathbf{X}^T\mathbf{X} + \mathbf{A}_j^{-1}\right\}^{-1}\mathbf{X}^T\boldsymbol{\eta}_{.j}, \left\{\mathbf{X}^T\mathbf{X} + \mathbf{A}_j^{-1}\right\}^{-1}\right) \tag{2.33}$$

where $\mathbf{X}$ is the matrix of predictors with each row i of $\mathbf{X}$ corresponding to the vector of predictors for subject i, $\mathbf{x}_i^T = (x_{i1}, \cdots, x_{ir}), i = 1, \cdots, n$, $\mathbf{A}_j$ is a diagonal matrix with its diagonal consisting of the vector $(\omega_{1j}^{-1}, \cdots, \omega_{rj}^{-1})$, and $\boldsymbol{\eta}_{.j}$ is the jth column of $\boldsymbol{\eta}, j = 1, \cdots, p^*$. where . represents conditioning on data and current estimates of all other unknown parameters.

The number of important factors in our model, p^*, is assumed to be much less than the number of basis functions, k. We can adaptively select the appropriate number

of important factors, p^*, such that the model does not miss any important factors and is not computationally intensive (Bhattacharya and Dunson, 2011). Starting with a very conservative guess $\tilde{p}$ for p^*, we denote the posterior samples of $\mathbf{\Lambda}$ as $\mathbf{\Lambda}_{\tilde{p}}$. Let $q^{(t)}$ be the number of columns in $\mathbf{\Lambda}_{\tilde{p}}$ having all elements in a small, pre-specified neighborhood of zero at the t^{th} iteration. We define the number of important factors at t^{th} iteration to be $p^{*(t)} = \tilde{p} - q^{(t)}$. Ideally, we would like to reduce $\tilde{p}$ to p^* by discarding the redundant factors or factors corresponding to a column in the loadings having all elements in a small, pre-specified neighborhood of zero.

To accomplish this aim, at the t^{th} iteration, we calculate a probability $P(t) = \exp(\xi_0 + \xi_1 t)$ and draw u_t from standard uniform distribution. We choose ξ_0 and ξ_1 so $p^{*(t)}$ changes about once every 10 iterations at the beginning of the chain but this frequency decreases exponentially fast (Bhattacharya and Dunson, 2011). If $u_t \leq P(t)$, we check the columns of the loadings, $\mathbf{\Lambda}^{(t)}_{p^*}$, to see if any columns are redundant. We discard all redundant columns, but if none of the columns are redundant, we add a column to the loadings and sample parameters from prior distributions to fill in the additional column. If $u_t > P(t)$, we do not make any changes to the number of important factors for the t^{th} iteration.

2.3 Simulation Studies

In order to assess the finite-sample properties of the methods proposed in Section 2.2, we conduct two simulation studies.

We first investigate the performance of our proposed methods under different sample sizes and missing proportions. We choose six combinations of sample sizes n and missing proportions d, namely the (n, d) combinations (150,10%), (150,30%), (200,10%), (200,30%), (500,10%) and (500,30%). For each combination, we consider 500 Monte Carlo simulated datasets. We begin by simulating curves from the model

given in (2.4). We set the number of cubic B-spline basis functions to be 10, and compute the 10 basis functions at 99 time points, which we use as the 99×10 design matrix $\mathbf{B}$ in (2.4) for all subjects. To simulate $\mathbf{B}$'s corresponding coefficients $\boldsymbol{\beta}_i$ from the model given in (2.5), we let $\boldsymbol{\beta}_0 = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$ and $\boldsymbol{\Sigma}_\beta$ be autoregressive with the $(i,j)^{th}$ element $= 4 * 0.5^{|i-j|}$. We simulate measurement errors $\boldsymbol{\epsilon}_i$ from model given in (2.6) with $\sigma^2 = 0.1$. We then simulate $\mathbf{Y}_i,\ i = 1, \cdots, n$ from the model given in (2.4).

To generate the missingness indicator Δ_i such that $\mathbf{Y}_i^{(2)}$ is missing at random (MAR), we use a simple model we describe below:

$$\log\left(\frac{P(\Delta_i = 1)}{1 - P(\Delta_i = 1)}\right) = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{Y}_i^{(1)} \tag{2.34}$$

where $\gamma_0 = 2.4$ or 0.5 and $\boldsymbol{\gamma} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$ so the proportion of $\mathbf{Y}_i^{(2)}$ being missing is around 10% or 30%, respectively.

To impute the missing functional data by SLF, we fit each simulated dataset to SLF as we describe in Section 2.2.2, assuming the missing data mechanism to be ignorable. We over-specify the number of basis to be 20 and initialize $\boldsymbol{\beta}_1^{(0)}, \cdots, \boldsymbol{\beta}_n^{(0)}, \boldsymbol{\eta}_1^{(0)}, \cdots, \boldsymbol{\eta}_n^{(0)}, \boldsymbol{\Lambda}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \sigma^{2(0)}$ from an approximation to their respective posterior distributions, then apply DA algorithm described in Section 2.2.2, with a chain length of 25000 iterations, a burn-in of 5000 and thinning by only collecting every 5^{th} sample. Each sample we collect contains a set of imputed curves from SLF, resulting in up to 4000 sets of imputed curves. We combine observed curves with a set of imputed curves to form an imputed dataset, which can be analyzed as if no curves are missing.

The purpose of imputation is to facilitate further statistical analyses in the presence of missing data. To evaluate our method, we consider the estimation of the sum of points over time in the second curve. The average is calculated for the imputed datasets and the variance of the estimator by applying Rubin's rules. We also obtain

an estimated 95% credible interval of the estimator following Barnard and Rubin (1999). Once we analyze data from each of the 500 MC datasets, we obtain relative bias, root mean square error (RMSE), Monte Carlo standard deviation (MCSD), estimated standard error (estimated SE), and coverage probability (CP).

To compare the performance of SLF to FK, we repeat the analysis following imputation procedures described in section 2.2.1. In order to allow FK to achieve its best performance, we set the number of basis as the true value, which is 10. We apply the same chain length, burn-in and thinning for the DA algorithm, and same combining rules and credible interval estimation strategy for FK as for SLF.

We also compare our proposed methods to three other methods. The first method analyzes the complete data, where all curves are observed. It serves as a bench mark that produces the best results any imputation-based method can possibly achieve. The second method (CC) is a common approach in practice when missing data is present. CC analyzes only the complete cases, where subjects missing second curve are excluded from the analysis. The third method we consider (naive MI) analyzes imputed datasets from a multiple imputation method that does not take into consideration the functional nature of the data. In naive MI, we impute the missing data using a "naive" multiple imputation method where we treat measurement at each time point as a separate variable, and apply a multivariate normal distribution to model $\mathbf{Y}_i$, then impute the missing $\mathbf{Y}_i^{(2)}$ using $\mathbf{Y}_i^{(1)}$. We summarize results from simulations of sample size 150, 200, and 500, with $\mathbf{Y}^{(2)}$ missing proportions of 10% and 30% in Table 2.1.

As expected, bench mark has the smallest relative bias, as well as RMSE, MCSD, and estimated SE. Its coverage probability achieves the nominal level of 95% but deviates slightly as the sample size increases. Analysis of complete cases (CC) has the largest relative bias out of the five methods being compared, and its relative bias increases as missing proportion increases. CC also has largest RMSE and poor

Table 2.1: Simulation results based on 100 MC datasets of sample size 150, 200, or 500, with 10% or 30% missing $\mathbf{Y}^{(2)}$. The true parameter value is 1.

n	Missing	Method	Rel. Bias	RMSE	MCSD	Est. SE	CP
150	10%	bench mark	-0.006	0.112	0.112	0.113	0.95
		CC	0.046	0.125	0.116	0.119	0.96
		naive MI	-0.006	0.133	0.134	0.127	0.94
		FK	-0.009	0.120	0.120	0.119	0.95
		SLF	-0.002	0.126	0.127	0.118	0.94
150	30%	bench mark	-0.006	0.112	0.112	0.113	0.95
		CC	0.136	0.182	0.121	0.132	0.83
		naive MI	0.011	0.149	0.150	0.155	0.91[1]
		FK	0.008	0.142	0.142	0.136	0.94
		SLF	0.024	0.137	0.135	0.131	0.95
200	10%	bench mark	-0.008	0.093	0.093	0.098	0.97
		CC	0.048	0.107	0.096	0.103	0.93
		naive MI	-0.005	0.110	0.110	0.108	0.91
		FK	-0.004	0.103	0.104	0.104	0.95
		SLF	-0.003	0.106	0.106	0.103	0.94
200	30%	bench mark	-0.008	0.093	0.093	0.098	0.97
		CC	0.118	0.155	0.101	0.115	0.90
		naive MI	0.002	0.129	0.129	0.126	0.96
		FK	-0.013	0.121	0.120	0.118	0.94
		SLF	-0.002	0.131	0.132	0.116	0.91
500	10%	bench mark	-0.011	0.060	0.059	0.062	0.98
		CC	0.040	0.072	0.060	0.065	0.95
		naive MI	-0.010	0.063	0.063	0.066	0.97
		FK	-0.012	0.064	0.063	0.065	0.97
		SLF	-0.011	0.065	0.065	0.065	0.97
500	30%	bench mark	-0.011	0.060	0.059	0.062	0.98
		CC	0.115	0.129	0.060	0.073	0.64
		naive MI	-0.013	0.066	0.065	0.074	0.97
		FK	-0.012	0.072	0.072	0.075	0.96
		SLF	-0.004	0.068	0.068	0.073	0.98

[1]Results for naive MI were based on 96 MC datasets, as naive MI algorithm failed to converge in 4 MC datasets. Only 91 of the MC datasets had 95% CI that contained the truth.

coverage probability when missing proportion is high, regardless of sample size. The large relative bias and large RMSE indicate that CC is not an appropriate analysis when missing data is present, especially when the missing proportion is high (30%). SLF and FK have comparable relative biases which are close to relative bias from the "benchmark". RMSE and MCSD from SLF are slightly larger than FK for small

samples because the model in SLF is more complex. Although RMSE, MCSD, and estimated SE of SLF and FK are larger than those of "benchmark", their differences decrease as sample size increases and missing proportion decreases. Naive MI has comparable relative bias to SLF and FK, but larger variability when the sample size is small. In fact, naive MI is numerical unstable when the sample size is small and missing proportion is large, because it requires larger effective sample size than SLF and FK to fit its imputation model.

The simulation studies demonstrate that when sample size was small and missing proportion large, FK and SLF perform better than naive MI. We do not discern large differences in performance between FK and SLF in this set of simulations, possibly because we specify the true number of basis in FK. In the next set of simulations, we compare performances of FK and SLF when the number of basis functions is misspecified by FK.

We begin by simulating 500 Monte Carlo datasets. Within each dataset, we simulate 200 sets of curves from the model given in (2.4) with $k = 16$. We simulate $\boldsymbol{\beta}_i$ from the model given in (2.5), where we let $\boldsymbol{\beta}_0 = \begin{pmatrix} .5 & 2.0 & 1.5 & 2.5 & 2.0 & 2.3 & 1.8 & 1.5 & 1.5 & 1.3 & 1.0 & 1.3 \end{pmatrix}$ and $\boldsymbol{\Sigma}_\beta$ was autoregressive with the $(i,j)^{th}$ element $= 4 * 0.5^{|i-j|}$. We evaluate the 16 basis functions at 99 time points, which we use as the 99×16 design matrix $\mathbf{B}$ in (2.4) for all subjects. We simulate measurement errors $\boldsymbol{\epsilon}_i$ from model given in (2.6) with $\sigma^2 = 0.1$. We then simulate $\mathbf{Y}_i,\ i = 1, \cdots, n$ from the model given in (2.4). For the missingness indicator, we use (2.34) where $\gamma_0 = -0.1$ and $\boldsymbol{\gamma} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$ so the missing proportion of $\mathbf{Y}_i^{(2)}$ is around 30%.

We analyze the data after we apply FK to impute the missing curves, where we restrict $k = 10$, so the number of basis functions in FK is misspecified. As comparison, we also analyze data imputed by SLF, CC, and bench mark, following the same procedures in our previous set of simulations. We present the results in Table 2.2.

Table 2.2: Simulation results from 500 Monte Carlo datasets of sample size 200 with 30% missing $\mathbf{Y}^{(2)}$

Method	**Rel. Bias**	**RMSE**	**MCSD**	**Est. SE**	**CP**
bench mark	-0.007	3.555	3.549	3.424	0.93
CC	0.063	4.880	4.238	4.079	0.89
FK	0.004	4.621	4.623	4.305	0.92
SLF	0.010	4.696	4.685	4.044	0.91

In Table 2.2, the performance of the bench mark and of CC are as expected based on what we observe in our previous set of simulations. Bench mark has the smallest relative bias and variability. CC has the largest relative bias and RMSE, as well as the poorest coverage probability. Overall, FK and SLF are similar. These results do not reveal any deficiency in misspecification of number of basis in FK compared to SLF.

To summarize the differences between imputed curves and true curves, we compute two additional statistics based on observations at each time point from the same set of simulations. The point-wise root mean squared error (RMSE2) is obtained by taking the square root of the mean difference squared between truth and observed or imputed values averaged across time points and across subjects. The mean absolute error (MAE) is obtained by taking the mean absolute difference between truth and observed or imputed values averaged across time points and across subjects. We present the results in Table 2.3. Bench mark has the smallest RMSE2 and MAE as we expect. FK and SLF have lower RMSE2 and MAE than naive MI, suggesting that curves imputed via FK and SLF are closer to the true curves than curves imputed via naive MI. FK and SLF have lower RMSE2 and MAE than CC, likely due CC having a smaller sample size. Contrary to what we expect, even though the number of basis functions are under-specified in FK, FK has smaller RMSE2 and MAE than SLF.

Our simulation studies demonstrate analysis using datasets obtained via multiple imputation methods that took advantage of the functional nature of the data are superior to analysis using only complete cases or imputed datasets that does not take

Table 2.3: Additional simulation results based on 500 MC datasets of sample size 200, with 30% missing $\mathbf{Y}^{(2)}$

	RMSE2	**MAE**
benchmark	0.113	0.097
CC	0.173	0.143
naive MI	0.184	0.157
FK	0.146	0.126
SLF	0.157	0.134

advantage of the functional nature of the data. We also demonstrate that when the number of basis functions is misspecified, FK still has comparable performance to SLF, which is designed to adaptively select the number of basis functions. However, the simulated true curves are smooth so under-specification of the number of basis functions is not severely penalized.

2.4 Renal Study

Our work is motivated by a renal study conducted in the Division of Nuclear Medicine at Emory University, aimed at improving renal image interpretations by radiologists. The renal study data consists of renal curves from 116 patients totaling 229 kidneys, 20 of which (8.7%) only had the first curve. Renal curves from kidneys of the same patient are considered to be independent of each other. The observations were taken at the same 59 time points for the first curve and the same 40 time points for the second curve, with 15 to 30 second intervals between the time points, across the subjects. The observations range from 0 to 197,295, which means covariance matrix calculations based on the data would be computationally expensive. To lower computational cost and ensure numerical stability, we transform the data by taking the logarithm of the observations.

It is of interest to impute the missing second curves to facilitate analyses of renal study data using all subjects. We impute the missing curves using FK, specifying the

number of basis functions $k = 10$, and SLF, specifying the number of basis functions $k = 20$ then adaptively select the number of latent factors. We impute missing curves via naive MI to investigate whether incorporation of functional nature of the data is pertinent for the imputation of the missing curves. To compare the performance of FK, SLF, and naive MI, we construct a simple summary measure statistic of the second curve by taking the average of the 40 observations for each subject. We also compare the analysis results using the imputed datasets against CC to examine whether imputation is necessary.

Table 2.4: Estimate and SD of average of observations from second renal curve.

	Estimate	**SE**
CC	8.945	0.097
naive MI	8.838	0.099
FK	8.866	0.096
SLF	8.873	0.095

Table 2.4 shows estimates from FK and SLF are closest to each other. FK and SLF also have smaller SE than naive MI and CC, suggesting the incorporation of smoothness of functional data increases efficiency of the imputation methods. CC estimate is farthest away from all other estimates, demonstrating that CC can be inappropriate for functional data with ignorable missing data mechanism, even when the missing proportion is around 10%.

2.5 Discussion

We have proposed two multiple imputation methods for imputing missing functional data. Both methods incorporated the functional nature of the data, and both performed better than naive MI method that did not take the smoothness in the functional data into consideration. In contrast, CC had the largest relative biases and variability therefore was not advised when the missing data mechanism is ignorable.

Of the two proposed methods, FK was simpler to implement and less computationally intensive than SLF but required *a priori* specification of the number of basis functions. SLF, on the other hand, performed well as long as the number of basis functions specified in SLF was sufficiently large, as SLF adaptively selects the number of latent factors associated with the basis coefficients, effectively shrinks the number of basis functions for each subject to lighten the burden on its computational cost. We expect FK to perform poorly in scenarios where the true curves are more rugged or wavy, which more severely penalize underestimation of the optimal number of basis functions.

We assumed renal curves of kidneys from the same patient to be independent for simplicity, but this assumption may not be accurate. Kidneys from the same patient share the same covariates, such as age and sex of the patient. If patients sought help and received treatment adequately and promptly, obstruction in both kidneys should be rare. Further investigation into models that combine renal curves from kidneys of the same patient is warranted. Such models would be further complicated by the fact that some patients have had one of their kidneys removed.

Our imputation models ignored the renal obstruction status, as we assumed this did not provide any additional information to the curves. This means the imputed curves are not biased toward a specific renal obstruction outcome and can be used to assist radiologists in determining renal obstruction of new patients. However, if the imputed renal curves are used in analyses that seek associations between renal curves and renal obstruction, then renal obstruction status need to be included in the imputation model to avoid uncongeniality between imputation model and analysis model.

Chapter 3

Handling missing data in generalized functional linear models with application to renal studies

3.1 Introduction

Functional data are often encountered in biomedical studies, and are either collected at specific time points separated by fixed intervals or at convenient time points within a time frame. A notable obstacle in the functional data analysis of biomedical data is how to handle missing data. One particular example is renal study data, which consist of measurements taken at fixed timed points within a twenty-four minutes time frame, sometimes supplemented by additional measurements, also taken at fixed time points, within a twenty minute time frame with an approximate thirty minute gap. The resulting renal curve data is used by radiologists in their diagnosis of renal obstruction. The necessity for the additional measurements is determined by a clinician based on the first set of measurements.

Consider the problem of estimating the association between functional data and a binary outcome in the presence of missing functional data. Some existing approaches already incorporate curves, or functional data, as predictors. When the outcome is continuous, Cardot et al. (1999, 2003) used a functional principal component regression approach and a penalized B-spline approach to model the outcome, but failed to accommodate for missing data in their models. When the outcome is categorical or binary, Ferraty and Vieu (2003) proposed a nonparametric curves discrimination model using kernel estimator but did not accommodate for missing data. James (2002) proposed a FGLM model with a logit link and used EM algorithm to fit his model. James's approach required the number of basis functions to be specified *a priori* and results in a trade-off between accuracy and efficiency of estimation of regression coefficients.

In this Chapter, we propose a functional generalized linear model approach to model the association between functional predictors and a binary outcome. The functional predictors are modeled by cubic B-splines. To fit our model, we use the Bayesian data augmentation (Tanner and Wong, 1987) algorithm. Our approach uses

all available data, not just data from complete cases or from summary measures, in building the model, and is easy to implement. We provide details of our approach in the next section. In Section 3.3 we report results from simulation studies we conducted to assess our approach. In Section 3.4 we present an application to data from renal studies. A few concluding remarks are given in Section 3.5.

3.2 Methodology

3.2.1 Data Structure

Let $\mathbf{Y}_{(m\times n)} = \begin{pmatrix} \mathbf{Y}_1 & \cdots & \mathbf{Y}_n \end{pmatrix}$ be the measurements taken at m time points for n individuals. Let $\mathbf{Y}_i$ be the element of $\mathbf{Y}$ and $\mathbf{Y}_i$ denote the vector of measurement of the i^{th} subject. Let $\mathbf{Y}^{(1)}_{i(m_1\times 1)}$ be the first curve and $\mathbf{Y}^{(2)}_{i((m_2)\times 1)}$ be the second curve, $(m_1 + m_2 = m)$. We write $\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} Y_{i1}^{(1)} & \cdots & Y_{im_1}^{(1)} & Y_{i1}^{(2)} & \cdots & Y_{im_2}^{(2)} \end{pmatrix}^T$. Let $\Delta = \begin{pmatrix} \Delta_1 & \cdots & \Delta_n \end{pmatrix}^T$ be the vector of indicator functions with $\Delta_i = 1$ if $\mathbf{Y}_i^{(2)}$ was observed. We let $\mathbf{Y}_{obs}$ represent the observed curve data, and $\mathbf{Y}_{mis} = \{\mathbf{Y}_i^{(2)} : \Delta_i = 0\}$ represent the unobserved curve data. Finally, we let W_i be the observed renal obstruction status, the binary response, for individual i.

3.2.2 Model

Given we sample an individual curve with error over distinct time points $t_{i1}, \cdots, t_{im_i}$ (James and Hastie, 2001), we use the following model to represent the curve data:

$$Y_{ij}^{(s)} = \theta_i^{(s)}(t_j) + \epsilon_{ij}^{(s)},\ i = 1, \cdots, n,\ j = 1, \cdots, m_s,\ s = 1, 2, \tag{3.1}$$

where $Y_{ij}^{(s)}$ is the observed value of the s^{th} curve for kidney i at time point t_j, $\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_i^{(1)} \\ \boldsymbol{\epsilon}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \epsilon_{i1}^{(1)} & \cdots & \epsilon_{im_1}^{(1)} & \epsilon_{i1}^{(2)} & \cdots & \epsilon_{im_2}^{(2)} \end{pmatrix}^T$ represents measurement errors, $\theta_i^{(s)}(t_j)$ represents the true value at time point t_j. We model $\theta_i^{(s)}(t_j)$ as

$$\theta_i^{(s)}(t_j) = \sum_{l=1}^{k/2} \beta_{il}^{(s)} b_l^{(s)}(t_j) = \mathbf{b}^{(s)T}(t_j)\boldsymbol{\beta}_i^{(s)}, \tag{3.2}$$

where $\mathbf{b}^{(s)}(t_j) = \begin{pmatrix} b_1^{(s)}(t_j) & \cdots & b_{k/2}^{(s)}(t_j) \end{pmatrix}^T$, $s = \{1, 2\}$ are the cubic spline basis functions. We choose $\frac{k}{2}$ basis functions for each curve, for cubic spline basis with intercept, $\frac{k}{2} - 4$ is the number of knots. We employ the cubic spline model for the true curve for the following reasons. First, the coefficients which define the polynomial form of the cubic spline model can be found by solving a system of equations for which stable and fast numerical algorithms already exist (Silverman, 1985). Second, in a cubic spline model, observation at t_j only has a fast-decreasing influence on nearby parts, which makes it favorable to other curve-fitting methods such as polynomial regression (Silverman, 1985). Finally, the dependence of the local bandwidth on the density of observed time points in the cubic spline model is intermediate between fixed kernal smoothing and smoothing based on neighboring values, which is desirable because moving from fixed kernel to nearest neighbor methods resulted in overfitting (Silverman, 1984, 1985).

We let $\mathbf{B}_{(m\times k)}$ be the block diagonal matrix that consists of the cubic spline basis functions of first and second curve for a kidney, specifically, $\mathbf{B} = \begin{pmatrix} \mathbf{B}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{(2)} \end{pmatrix}$ where $\mathbf{B}^{(1)}$ corresponds to the first curve and $\mathbf{B}^{(2)}$ corresponds to the second curve, and

$$\mathbf{B}^{(s)} = \begin{pmatrix} b_1^{(s)}(t_1) & \cdots & b_{k/2}^{(s)}(t_1) \\ \vdots & \ddots & \vdots \\ b_1^{(s)}(t_{m_s}) & \cdots & b_{k/2}^{(s)}(t_{m_s}) \end{pmatrix}, s = 1, 2. \tag{3.3}$$

We denote by $\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_i^{(1)} \\ \boldsymbol{\beta}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \beta_{i1}^{(1)} & \cdots & \beta_{i\frac{k}{2}}^{(1)} & \beta_{i1}^{(2)} & \cdots & \beta_{i\frac{k}{2}}^{(2)} \end{pmatrix}^T$ the corresponding vector of spline coefficients for $\mathbf{B}$. We assume the true curve for the i^{th} kidney to follow a multivariate normal distribution

$$\boldsymbol{\beta}_i \sim N_k(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) \tag{3.4}$$

based on some hyperparameters $\boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}_0^{(1)} \\ \boldsymbol{\beta}_0^{(2)} \end{pmatrix}$ and $\boldsymbol{\Sigma}_\beta = \begin{pmatrix} \boldsymbol{\Sigma}_\beta^{11} & \boldsymbol{\Sigma}_\beta^{12} \\ \boldsymbol{\Sigma}_\beta^{21} & \boldsymbol{\Sigma}_\beta^{22} \end{pmatrix}$.

We also assume the following model for the measurement error,

$$\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \sigma^2 \mathbf{I}_m) \tag{3.5}$$

with σ^2 as its variance, which we assume to be constant across all kidneys, and that $\boldsymbol{\beta}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ since true curves should not impact measurement errors. We assume the hyperparameters $\boldsymbol{\beta}_0$ and σ^2 to have non-informative and improper priors $\pi(\boldsymbol{\beta}_0, \sigma^2) \propto \mathbf{1}$ and the hyperparameter Σ_β to have an Inverse-Wishart prior $\boldsymbol{\Sigma}_\beta \sim \mathcal{W}^{-1}(\mathbf{I}_k, k)$ where $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ has scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν. These assumptions about the prior specifications of the hyperparameters help facilitate our imputation procedure.

We let $Z_i (i = 1, \cdots, n)$ be a latent variable associated with true renal obstruction status for individual i and let $\mathbf{Z}_{(n\times 1)} = \begin{pmatrix} Z_1 & \cdots & Z_n \end{pmatrix}$ be the vector of latent variable Z_i for all n subjects. In lieu of the unknown true renal obstruction, we use consensus ratings of three expert readers as our outcome, with $W_i = 1$ for obstruction and $W_i = 0$ for non-obstruction. We assume a probit model for W_i given Z_i such that

$$W_i = \begin{cases} 1 & if \quad Z_i > 0 \\ 0 & if \quad Z_i \leq 0 \end{cases}, \tag{3.6}$$

with Z_i following Gaussian distribution $Z_i \sim \mathcal{N}(\mu_i, 1)$ where the mean μ_i of Z_i is

modeled by the functional model

$$\mu_i = \alpha_0 + \int a^{(1)}(t)\theta_i^{(1)}(t)\mathrm{d}t + \int a^{(2)}(t)\theta_i^{(2)}(t)\mathrm{d}t \tag{3.7}$$

and where $\int a^{(s)T}(t)\theta_i^{(s)}(t)\mathrm{d}t$ represent coefficient of association integrated over the first ($s = 1$) or second ($s = 2$) curve. We use orthonormal cubic spline basis functions $\mathbf{b}^{(s)}(t)$ we used in (3.2) to model the coefficient $a^{(s)}(t)$ as $a^{(s)}(t) = \boldsymbol{\alpha}^{(s)T}\mathbf{b}^{(s)}(t)$, where $\boldsymbol{\alpha}^{(s)}$ are the coefficients corresponding to $\mathbf{b}^{(s)}(t)$. Since $\mathbf{b}^{(s)}(t)$ are orthonormal,

$$\int a^{(s)}(t)\theta_i^{(s)}(t)\mathrm{d}t = \int \boldsymbol{\alpha}^{(s)T}\mathbf{b}^{(s)}(t)\mathbf{b}^{(s)T}(t)\boldsymbol{\beta}_i^{(s)}\mathrm{d}t = \boldsymbol{\alpha}^{(s)T}\left(\int \mathbf{b}^{(s)}(t)\mathbf{b}^{(s)T}(t)\mathrm{d}t\right)\boldsymbol{\beta}_i^{(s)} = \boldsymbol{\alpha}^{(s)T}\boldsymbol{\beta}_i^{(s)},$$

then from (3.7) we have

$$\mu_i = \alpha_0 + \boldsymbol{\alpha}^{(1)T}\boldsymbol{\beta}_i^{(1)} + \boldsymbol{\alpha}^{(2)T}\boldsymbol{\beta}_i^{(2)} = \alpha_0 + \boldsymbol{\alpha}^T\boldsymbol{\beta}_i, \tag{3.8}$$

where $\boldsymbol{\alpha}$ is a column vector of length k and $\boldsymbol{\alpha}^T = \begin{pmatrix}\boldsymbol{\alpha}^{(1)T} & \boldsymbol{\alpha}^{(2)T}\end{pmatrix}$.

3.2.3 Likelihood

When $\mathbf{Y}$ is completely observed, the posterior distribution is

$$\begin{aligned}
&P(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2, \alpha_0, \boldsymbol{\alpha}, \mathbf{Z}|\mathbf{Y}, \mathbf{W}, \mathbf{X}) \\
\propto\ &\pi(\boldsymbol{\beta}_0, \sigma^2)\pi(\boldsymbol{\Sigma}_\beta)\prod_{i=1}^{n}\left\{P(\mathbf{Y}_i|\boldsymbol{\beta}_i, \sigma^2, \mathbf{X})\pi(\boldsymbol{\beta}_i|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)P(Z_i|\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_i)P(W_i|Z_i)\right\} \\
\propto\ &\sigma^{-nm}|\boldsymbol{\Sigma}_\beta|^{-\frac{2n+2k+1}{2}}\prod_{i=1}^{n}\left\{\left(1 - \int_0^{\alpha_0+\boldsymbol{\alpha}^T\boldsymbol{\beta}_i} e^{-\frac{t^2}{2}}\mathrm{d}t\right)^{-1} I(W_i = 1) + \left(\int_0^{\alpha_0+\boldsymbol{\alpha}^T\boldsymbol{\beta}_i} e^{-\frac{t^2}{2}}\mathrm{d}t\right)^{-1} I(W_i = 0)\right\} \\
&\times \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}(Z_i - \alpha_0 - \boldsymbol{\alpha}^T\boldsymbol{\beta}_i)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)\right.\right. \\
&\left.\left.+ \sum_{i=1}^{n}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_\beta^{-1}\right)\right\}\right]
\end{aligned} \tag{3.9}$$

where $I(.)$ is the indicator function.

When for some individuals ($\Delta_i = 0$), only $\mathbf{Y}_i^{(1)}$ was observed, $\mathbf{Y}_i^{(2)}$ was missing. Under ignorable missing data mechanism, the posterior distribution is

$$
\begin{aligned}
& P(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2, \alpha_0, \boldsymbol{\alpha}, \mathbf{Z} | \mathbf{Y}, \mathbf{W}, \mathbf{X}, \boldsymbol{\Delta}) \\
&= P(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2 | \mathbf{Y}_{obs}, \mathbf{X}, \boldsymbol{\Delta}) \\
&\propto \pi(\boldsymbol{\beta}_0, \sigma^2)\pi(\boldsymbol{\Sigma}_\beta) \prod_{i=1}^{n} \Bigg[\left\{ P(\mathbf{Y}_i | \boldsymbol{\beta}_i, \sigma^2, \mathbf{X}) \pi(\boldsymbol{\beta}_i | \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) P(Z_i | \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_i) P(W_i | Z_i) \right\}^{\Delta_i} \\
&\quad \times \left\{ P(\mathbf{Y}_i^{(1)} | \boldsymbol{\beta}_i^{(1)}, \sigma^2) \pi(\boldsymbol{\beta}_i^{(1)} | \boldsymbol{\beta}_0^{(1)}, \boldsymbol{\Sigma}_\beta^{11}) P(Z_i | \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_i) P(W_i | Z_i) \right\}^{1-\Delta_i} \Bigg] \\
&\propto \sigma^{-nm_1 - m_2 \sum_{i=1}^n \Delta_i} |\boldsymbol{\Sigma}_\beta|^{-\frac{\sum_{i=1}^n \Delta_i + n + 2k + 1}{2}} |\boldsymbol{\Sigma}_\beta^{11}|^{-\frac{\sum_{i=1}^n 1 - \Delta_i}{2}} \\
&\quad \times \prod_{i=1}^{n} \left\{ \left(1 - \int_0^{\alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{\beta}_i} e^{-\frac{t^2}{2}} \mathrm{d}t \right)^{-1} I(W_i = 1) + \left(\int_0^{\alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{\beta}_i} e^{-\frac{t^2}{2}} \mathrm{d}t \right)^{-1} I(W_i = 0) \right\} \\
&\quad \times \exp \Bigg[-\frac{1}{2} \Bigg\{ \sum_{i=1}^{n} (Z_i - \alpha_0 - \boldsymbol{\alpha}^T \boldsymbol{\beta}_i)^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} \Delta_i (\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T (\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i) \\
&\quad + \frac{1}{\sigma^2} \sum_{i=1}^{n} (1 - \Delta_i) \left(\mathbf{Y}_i^{(1)} - \mathbf{X}^{(1)} \boldsymbol{\beta}_i^{(1)} \right)^T \left(\mathbf{Y}_i^{(1)} - \mathbf{X}^{(1)} \boldsymbol{\beta}_i^{(1)} \right) + \sum_{i=1}^{n} \Delta_i (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0) \\
&\quad + \sum_{i=1}^{n} (1 - \Delta_i) \left(\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_0^{(1)} \right)^T \boldsymbol{\Sigma}_\beta^{11-1} \left(\boldsymbol{\beta}_i^{(1)} - \boldsymbol{\beta}_0^{(1)} \right) + \mathrm{Tr} \left(\boldsymbol{\Sigma}_\beta^{-1} \right) \Bigg\} \Bigg] . \qquad (3.10)
\end{aligned}
$$

3.2.4 Markov Chain Monte Carlo

Posterior distribution of Σ_β based on (3.10) does not have a closed form. However, if $\mathbf{Y}_{mis}$ are observed, then the observed data likelihood no longer involve a mixture of distributions, instead, (3.9) can be used. Therefore, we adopt a Bayesian data augmentation (DA) algorithm (Tanner and Wong, 1987) to simulate posterior distributions in expression (3.10). First, we initialize $\alpha_0^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}_1^{(0)}, \cdots, \boldsymbol{\beta}_n^{(0)}, \boldsymbol{\beta}_0^{(0)}, \boldsymbol{\Sigma}_\beta^{(0)}$, and $\sigma^{2(0)}$ from an approximation to their respective posterior distributions. Then, we use an iterative approach where each iteration consists of an imputation (I) and a posterior (P) step described below:

I step we draw $\mathbf{Y}_{mis}$ from the density

$P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\mathbf{W},\mathbf{X},\mathbf{Z},\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_n,\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_\beta,\sigma^2\right)$. By (3.1), (3.2) and (3.5),

$$\begin{pmatrix}\mathbf{Y}_i^{(1)}\\ \mathbf{Y}_i^{(2)}\end{pmatrix}|\boldsymbol{\beta}_i,\mathbf{X},\sigma^2 \sim N_m\left(\begin{pmatrix}\mathbf{X}^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^{(2)}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_i^{(1)}\\ \boldsymbol{\beta}_i^{(2)}\end{pmatrix},\sigma^2\begin{pmatrix}\mathbf{I}_{m_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{m_2}\end{pmatrix}\right) \tag{3.11}$$

given $\mathbf{X}^{(2)},\boldsymbol{\beta}_i^{(2)}$, and $\sigma^2,\mathbf{Y}_i^{(2)}$ does not depend on $\mathbf{Y}_i^{(1)},\mathbf{X}^{(1)},\boldsymbol{\beta}_i^{(1)}$ by (3.11), then the posterior predictive distributions of $\mathbf{Y}_i^{(2)}$ where $\Delta_i = 0$ is

$$\mathbf{Y}_i^{(2)}|\mathbf{X}^{(2)},\boldsymbol{\beta}_i^{(2)},\sigma^2 \sim \mathcal{N}_{m_2}\left(\mathbf{X}^{(2)}\boldsymbol{\beta}_i^{(2)},\sigma^2 I_{m_2}\right). \tag{3.12}$$

Since $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j|\boldsymbol{\beta}_i,\boldsymbol{\beta}_j \forall i \neq j$, for each i where $\Delta_i = 0$, we can use (3.12) to draw $\mathbf{Y}_i^{(2)}$, then $\mathbf{Y}_{mis} = \{\mathbf{Y}_i^{(2)} : \Delta_i = 0\}$.

P step we draw $\mathbf{Z},\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_n,\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_\beta$, and σ^2 from the density their respective full-conditional densities given data and current estimates of all other unknown parameters. By treating $(\mathbf{Y}_{obs},\mathbf{Y}_{mis})$ as $\mathbf{Y}$ in (3.9), the full conditional likelihood for the parameters α_0, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}_i$ are as follows: Let $\boldsymbol{\alpha}^* = \begin{pmatrix}\alpha_0\\ \boldsymbol{\alpha}\end{pmatrix}$ and $\boldsymbol{\beta}_i^* = \begin{pmatrix}1\\ \boldsymbol{\beta}_i\end{pmatrix}$, then

$$\boldsymbol{\alpha}^*|. \sim \mathcal{N}_{k+1}\left(\left\{\sum_{i=1}^n \boldsymbol{\beta}_i^*\boldsymbol{\beta}_i^{*T}\right\}^{-1}\sum_{i=1}^n \boldsymbol{\beta}_i^* Z_i, \left\{\sum_{i=1}^n \boldsymbol{\beta}_i^*\boldsymbol{\beta}_i^{*T}\right\}^{-1}\right) \tag{3.13}$$

$$\mathbf{S}_\beta = \left\{\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \boldsymbol{\Sigma}_\beta^{-1}\right\}^{-1}$$

$$\boldsymbol{\beta}_i|. \sim \mathcal{N}_k\left(\mathbf{S}_\beta\left\{\frac{\mathbf{X}^T\mathbf{Y}_i}{\sigma^2} + \boldsymbol{\alpha}(Z_i - \alpha_0) + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}_0\right\},\mathbf{S}_\beta\right) \tag{3.14}$$

where . represents all other parameters and data. Full conditional likelihood for

the latent variable $\mathbf{Z}$ and other parameters and hyperparameters can be found in the Appendix.

We repeat this iterative procedure until the algorithm converges. After convergence, algorithm yields draws from the joint posterior distribution of $(\mathbf{Y}_{mis}, \mathbf{Z}, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta, \sigma^2)$ given $\mathbf{Y}_{obs}$. Any draws of $\mathbf{Y}_{mis}$ after the convergence of the algorithm can be combined with $\mathbf{Y}_{obs}$ to form an imputed dataset.

3.2.5 Model Selection

Often the optimal basis functions that best approximate functional predictors are unknown. One approach is to set the basis functions *a priori* then fit the joint model. This approach requires the analyst to have prior knowledge on the characteristics and distributions of the functional predictors. An alternative approach is to fit joint models with different basis functions and apply selection criteria to choose the fitted joint model with the best basis functions. One selection criterion, the deviance information criterion (DIC), is proposed by Spiegelhalter et al. (2002) as a measure of model assessment and comparison. Celeux et al. (2006) then extend DIC to Bayesian hierarchical models with missing data. In the case of basis function selection, we fit joint models with varying number of cubic B-spline basis functions, then estimate their DIC via the following formula,

$$DIC = -4E_z \left[\log f\left(y, z|\theta\right)|y\right] + 2E_z \left[\log f\left(y, z|E_\theta\left[\theta|y, z\right]\right)|y\right],$$

where $\log f\left(y, z|\theta\right)$ is the log-likelihood given observed data y, missing data z, and parameters θ. We consider latent variables $Z_1, \cdots, Z_n$ as missing data for the purpose of DIC calculations. Once we have DIC estimates from the fitted models, the model with the minimum DIC is chosen as the best model.

3.3 Simulation Study

We conduct simulation studies in order to assess the finite-sample properties of the method we proposed in Section 3.2. We demonstrate the viability of using DIC as a criterion for selecting the number of basis functions in the first set of simulations.

We begin data generation by simulating the underlying true curves for 500 subjects using the model given in (3.2). In order to allow for under-specification and over-specification of the number of basis functions in candidate models, we use cubic B-spline basis functions with intercept and 2 interior knot for the first and second curve, resulting in a total number of basis functions of 12. We evaluate the basis functions at 40 time points ($m_1 = m_2 = 20$), which we use as the 40×12 design matrix $\mathbf{B}$ in (3.1) for all subjects. To simulate the coefficients $\boldsymbol{\beta}_i$ corresponding to the basis functions from the model given in (3.4), we let $\boldsymbol{\beta}_0$ be the vector $(\,0.70\ 3.41\ 1.25\ 2.02\ 1.40\ 1.62\ 2.43\ 1.93\ 1.45\ 1.34\ 1.78\ 1.01\,)$, and let $\boldsymbol{\Sigma}_\beta$ be a 12×12 compound-symmetric matrix with the diagonal elements = 5.49 and off-diagonal elements = 2.3. We simulate the noise $\boldsymbol{\epsilon}_i$ from the model given in (3.5) with $\sigma^2 = 0.16$. Then we simulate the curves $\mathbf{Y}$ from the model given in (3.1) using $\mathbf{B}$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$. To simulate $W_i (i = 1, \cdots, n)$ from the model given in (3.6), we first simulate Z_i using the model given in (3.8), where $\alpha_0 = -0.87$ and $\boldsymbol{\alpha} = (\,.14\ -.18\ -.39\ .38\ .31\ -.28\ .16\ .12\ .14\ -.16\ -.08\ -.42\,)$.

One simple model to simulate the missingness indicator Δ_i such that $\mathbf{Y}_i^{(2)}$ is missing at random (MAR) is

$$\log\left(\frac{P(\Delta_i = 1)}{1 - P(\Delta_i = 1)}\right) = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{Y}_i^{(1)} \tag{3.15}$$

where γ_0 and $\boldsymbol{\gamma}$ are fixed, and in the case where $\boldsymbol{\gamma} = \mathbf{0}$, $\mathbf{Y}_i^{(2)}$ is missing completely at random (MCAR). We let $\begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}$ be the true value for $\boldsymbol{\gamma}$, and let $\gamma_0 = -0.6$ so the proportion of $\Delta_i = 0$ is around 30%. We set $\mathbf{Y}_i^{(2)}$ to be missing whenever $\Delta_i = 0$.

To study the impact of misspecifying the number of basis functions on the estimation of deviance information criterion (DIC), we consider seven models. In the first model, we set the number of basis as the true value, which is 12. In the second and third models, we under-specify the number of basis as 8 or 10. In the remaining models, we over-specify the number of basis as 14, 16, 18, or 20. We use the joint-modeling approach described in Section 3.2 to fit the data to each of the seven models, assuming the missing data mechanism to be ignorable. In the DA algorithm of our joint-modeling approach, we run each MCMC chain for 15000 iterations and keep every 5^{th} sample after 10000 iterations to ensure convergence of the parameters and minimize autocorrelation between consecutive samples.

We present mean DIC estimates from 110 MC datasets in Table 3.1. Misspecifications of the number of basis functions result in increases in mean DIC. The increase in mean DIC occurs both when the number of basis functions is under-specified and when it is over-specified. In each of the MC dataset, the minimum DIC estimate originates from the model with the true number of basis functions, which is 12. Basis function selection via minimum DIC results in the model with 12 basis functions being correctly selected as the optimal model in every MC dataset. This indicates minimum DIC is a good selection criterion for choosing the number of basis functions in our proposed method.

Table 3.1: Mean DIC from 110 MC datasets of sample size 500, with 30% of $\mathbf{Y}^{(2)}$ missing on average. The true number of basis functions is 12.

	Number of Basis Functions						
	8	10	12	14	16	18	20
DIC	-34596.81	-37787.27	-52545.34	-36888.70	-30941.17	-21523.13	-12369.496

The true number of basis functions is selected by minimum DIC in each MC dataset.

We conduct additional simulation studies in order to assess the finite-sample properties of the method proposed in Section 3.2 for samples of size 500 or 1000, with approximately 30% of second curves missing.

We begin data generation by simulating the underlying true curves for 500 subjects using the model given in (3.2). We use cubic B-spline basis functions with intercept and 0 interior knot for the first and second curve, resulting in a total number of basis functions of 8. We evaluate the basis functions at 40 time points ($m_1 = m_2 = 20$), which we use as the 40×8 design matrix $\mathbf{B}$ in (3.1) for all subjects. To simulate the coefficients $\boldsymbol{\beta}_i$ corresponding to the basis functions from the model given in (3.4), we let $\boldsymbol{\beta}_0$ be the vector $(\,0.70\ 3.41\ 1.25\ 2.02\ 2.43\ 1.93\ 1.45\ 1.34\,)$, and let $\boldsymbol{\Sigma}_\beta$ be a 8×8 compound-symmetric matrix with the diagonal elements = 5.49 and off-diagonal elements = 2.3. We simulate the noise $\boldsymbol{\epsilon}_i$ from the model given in (3.5) with $\sigma^2 = 0.16$. Then we simulate the curves $\mathbf{Y}$ from the model given in (3.1) using $\mathbf{B}$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$. To simulate $W_i (i = 1, \cdots, n)$ from the model given in (3.6), we first simulate Z_i using the model given in (3.8), where $\alpha_0 = -0.87$ and $\boldsymbol{\alpha} = (\,0.14\ {-0.18}\ {-0.39}\ 0.38\ 0.16\ 0.12\ 0.14\ {-0.16}\,)$. We simulate the missingness indicator Δ_i from the model given in (3.15), where $\gamma_0 = -0.74$ and $\boldsymbol{\gamma} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}$ so the proportion of $\Delta_i = 0$ was around 30%. We set $\mathbf{Y}_i^{(2)}$ to be missing whenever $\Delta_i = 0$.

To assess the performance of estimation via our joint-modeling approach (JM) described in Section 3.2, we fit the joint model assuming the missing data mechanism to be ignorable. Due to the heavy computation involved in basis function selection, we assume the true number of basis functions is selected. The DA algorithm of our joint-modeling approach converged after 2000 iterations, therefore we run each MCMC chain for 5000 iterations and keep every 3^{rd} posterior sample after 2000 iterations to minimize correlation between consecutive posterior samples.

To evaluate the performance of our proposed method, we compare our results against results from two other methods. The first method (JMCC) applies our joint-modeling approach only to subjects with both curves completely observed and excludes subjects with missing curve, then estimates $\boldsymbol{\beta}_i$, α_0, $\boldsymbol{\alpha}$ and $\mathbf{Z}$ though an iterative process involving only the P step of the DA algorithm. We include JMCC in this sim-

ulation because it is more convenient for practitioners than JM. The second method (TSM) is less computationally intensive than JM because it analyzes the data in two separate stages. In the first stage, missing curves are imputed by the fixed number of knots multiple imputation method (FK) described in Chapter 2. In the second stage, each imputed dataset is analyzed in an identical manner to JMCC. The parameter estimates from TSM are obtained by applying Rubin's rules to the parameter estimates of the analyses from the second stage.

We present the parameter estimates and their variance estimates from 500 Monte Carlo (MC) datasets, with sample size 500 and missing proportion around 30%, for the three methods described above in Table 3.2. JM has the smallest relative bias for most of the parameters. MCSD and estimated SE from JM are also smaller than or comparable to JMCC and TSM for all the parameters in the table. Relative biases of $\boldsymbol{\alpha}$ in JMCC are comparable to JM, despite the exclusion of 30% of the sample. However, MCSD and estimated SE are larger for $\boldsymbol{\alpha}^{(1)} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \end{pmatrix}$ in JMCC compared to JM. Relative biases of $\boldsymbol{\beta}_0$ are much larger in JMCC compared to both JM and TSM because of the smaller sample size of JMCC and its $\boldsymbol{\beta}_0^{(1)}$ estimates are biased because the second curves are MAR. The coverage probability for $\boldsymbol{\beta}_0$ in JMCC is also very poor. The large positive relative bias and poor coverage probability we observe in JMCC for $\boldsymbol{\beta}_0$ are expected because of the exclusion of 30% of the sample from JMCC. Due to our simulation settings, $\boldsymbol{\beta}_i$ from excluded subjects are generally below $\boldsymbol{\beta}_i$ from subjects included in JMCC, which results in a positive bias in $\boldsymbol{\beta}_0$ estimates from JMCC. The relative biases for $\boldsymbol{\alpha}^{(2)} = \begin{pmatrix} \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 \end{pmatrix}$ in TSM are more than twice as large as in JM and JMCC, this can be contributed to the fact the model used to impute the missing curves in TSM is uncongenial (Meng, 1994) to the model used for the analysis in TSM.

To evaluate the performance of estimation via JM, JMCC, and TSM under a large sample size, we repeat the simulation with a sample size of 1000. We present

Table 3.2: Simulation results from 500 MC datasets of sample size 500, with 30.1% of $\mathbf{Y}^{(2)}$ missing on average.

		Joint Modeling				Complete Case Analysis				Two-Stage Modeling			
Par	Truth	Rel Bias	MCSD[1]	Est SE[2]	CP[3]	Rel Bias	MCSD[1]	Est SE[2]	CP[3]	Rel Bias	MCSD[1]	Est SE[2]	CP[3]
α_0	-0.87	0.070	0.178	0.172	0.940	0.073	0.231	0.217	0.926	-0.103	0.152	0.160	0.902
α_1	0.14	0.074	0.044	0.043	0.936	0.071	0.050	0.049	0.948	0.097	0.089	0.050	0.964
α_2	-0.18	0.082	0.045	0.044	0.936	0.080	0.050	0.049	0.950	-0.056	0.094	0.050	0.958
α_3	-0.39	0.078	0.057	0.053	0.892	0.071	0.062	0.058	0.904	-0.030	0.077	0.055	0.944
α_4	0.38	0.086	0.057	0.052	0.892	0.082	0.067	0.063	0.914	0.016	0.067	0.052	0.952
α_5	0.16	0.079	0.053	0.049	0.928	0.085	0.053	0.049	0.928	-0.233	0.053	0.051	0.920
α_6	0.12	0.058	0.050	0.048	0.938	0.048	0.050	0.048	0.938	-0.270	0.035	0.048	0.958
α_7	0.14	0.087	0.049	0.048	0.930	0.093	0.050	0.048	0.936	-0.231	0.057	0.050	0.942
α_8	-0.16	0.100	0.049	0.047	0.932	0.108	0.049	0.048	0.934	-0.226	0.043	0.048	0.930
β_{01}	0.70	-0.005	0.118	0.113	0.936	0.581	0.136	0.132	0.128	-0.005	0.118	0.113	0.938
β_{02}	3.41	-0.0005	0.107	0.115	0.964	0.122	0.123	0.134	0.114	-0.0005	0.107	0.116	0.968
β_{03}	1.25	0.0003	0.105	0.110	0.956	0.333	0.125	0.127	0.094	0.0003	0.105	0.110	0.962
β_{04}	2.02	0.0000	0.105	0.108	0.948	0.499	0.098	0.105	0.000	0.0000	0.106	0.108	0.948
β_{05}	2.43	0.014	0.154	0.139	0.908	0.173	0.130	0.128	0.074	0.0005	0.140	0.140	0.944
β_{06}	1.93	-0.022	0.153	0.136	0.916	0.215	0.126	0.126	0.088	0.001	0.135	0.138	0.962
β_{07}	1.45	0.010	0.144	0.135	0.942	0.285	0.122	0.127	0.090	-0.006	0.139	0.138	0.956
β_{08}	1.34	-0.033	0.151	0.135	0.916	0.307	0.122	0.126	0.074	-0.0001	0.135	0.135	0.952
σ^2	0.16	0.001	0.003	0.003	0.922	0.0003	0.004	0.003	0.930	0.001	0.003	0.003	0.922

[1]standard deviation of the 500 parameter estimates from the MC datasets

[2]mean of the 500 parameter standard deviation estimates from the MC datasets

[3]for proposed method and for complete data analysis, the 95% CIs from each MC dataset were constructed using 2.5^{th} and 97.5^{th} quantiles of posterior distributions of the parameters of interest as boundaries; for imputed dataset analysis, the 95% CIs were constructed based on t_R-distributions following Rubin and Schenker (1986)

Table 3.3: Simulation results from 500 MC datasets of sample size 1000, with 30.0% of $\mathbf{Y}^{(2)}$ missing on average.

		Joint Modeling				Complete Case Analysis				Two-Stage Modeling			
Par	Truth	Rel Bias	MCSD[1]	Est SE[2]	CP[3]	Rel Bias	MCSD[1]	Est SE[2]	CP[3]	Rel Bias	MCSD[1]	Est SE[2]	CP[3]
α_0	-0.87	-0.015	0.119	0.112	0.928	0.058	0.158	0.147	0.918	-0.109	0.104	0.106	0.872
α_1	0.14	0.055	0.030	0.029	0.948	0.033	0.031	0.033	0.954	0.039	0.028	0.028	0.954
α_2	-0.18	0.072	0.031	0.030	0.932	0.036	0.035	0.033	0.930	-0.101	0.028	0.028	0.912
α_3	-0.39	0.032	0.036	0.036	0.936	0.032	0.040	0.039	0.944	-0.061	0.031	0.033	0.892
α_4	0.38	0.016	0.035	0.035	0.956	0.043	0.043	0.043	0.950	-0.011	0.032	0.033	0.956
α_5	0.16	0.028	0.036	0.034	0.928	0.020	0.035	0.033	0.922	-0.274	0.025	0.032	0.782
α_6	0.12	0.061	0.033	0.033	0.950	0.070	0.032	0.033	0.950	-0.253	0.023	0.032	0.926
α_7	0.14	0.043	0.034	0.033	0.936	0.034	0.034	0.033	0.936	-0.277	0.024	0.032	0.842
α_8	-0.16	0.034	0.035	0.033	0.930	0.029	0.034	0.033	0.944	-0.275	0.024	0.032	0.752
β_{01}	0.70	0.005	0.075	0.088	0.974	0.595	0.087	0.105	0.014	0.005	0.075	0.088	0.976
β_{02}	3.41	0.001	0.076	0.096	0.990	0.123	0.088	0.110	0.010	0.001	0.076	0.095	0.992
β_{03}	1.25	-0.002	0.076	0.082	0.952	0.332	0.089	0.096	0.006	-0.002	0.076	0.082	0.954
β_{04}	2.02	-0.001	0.074	0.079	0.958	0.497	0.070	0.082	0.000	-0.001	0.074	0.079	0.960
β_{05}	2.43	-0.024	0.094	0.100	0.914	0.173	0.086	0.096	0.002	0.001	0.098	0.103	0.950
β_{06}	1.93	-0.025	0.092	0.096	0.924	0.215	0.084	0.093	0.000	-0.001	0.097	0.100	0.954
β_{07}	1.45	-0.026	0.097	0.097	0.932	0.287	0.090	0.094	0.006	-0.002	0.099	0.100	0.950
β_{08}	1.34	-0.025	0.095	0.097	0.948	0.313	0.088	0.094	0.004	0.003	0.099	0.099	0.936
σ^2	0.16	-0.001	0.002	0.002	0.942	-0.003	0.002	0.002	0.948	-0.0004	0.002	0.002	0.962

[1]standard deviation of the 500 parameter estimates from the MC datasets

[2]mean of the 500 parameter standard deviation estimates from the MC datasets

[3]for proposed method and for complete data analysis, the 95% CIs from each MC dataset were constructed using 2.5^{th} and 97.5^{th} quantiles of posterior distributions of the parameters of interest as boundaries; for imputed dataset analysis, the 95% CIs were constructed based on t_R-distributions following Rubin and Schenker (1986)

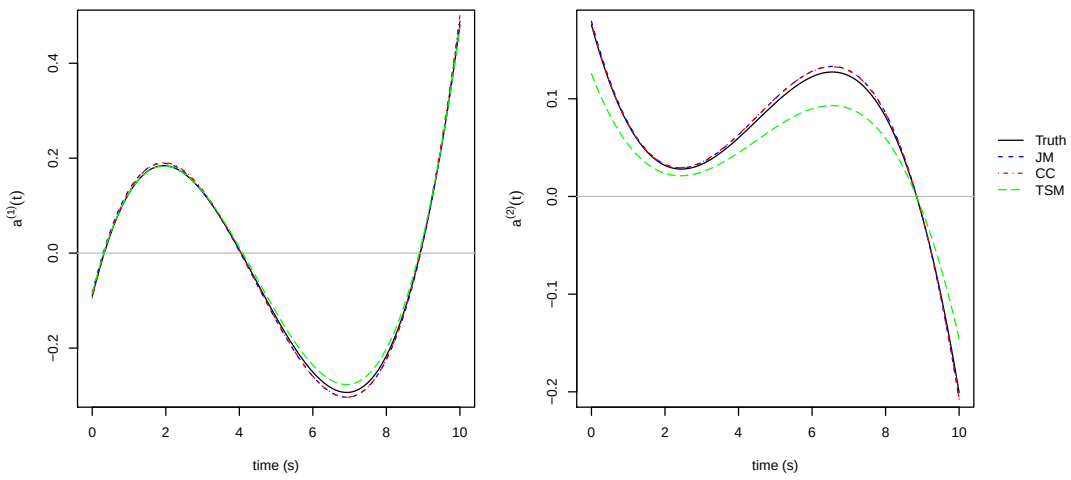

Figure 3.1: True and estimated functional coefficients of the first and second curves.

the parameter estimates and their variance estimates from 500 Monte Carlo (MC) datasets, with sample size 1000 and missing proportion around 30%, for JM, JMCC, and TSM in Table 3.3. The performance of JM and of JMCC are as expected based on what we observe in Table 3.2. The relative biases of α_0 and $\boldsymbol{\alpha}$ from JM and from JMCC are smaller compared to those found in Table 3.2 as a result of increasing sample size, but the relative biases for $\boldsymbol{\alpha}^{(2)}$ in TSM are larger. The relative biases of $\boldsymbol{\alpha}^{(2)}$ in TSM also remains larger than those of JM and JMCC. This may be indication that the effect of uncongeniality between imputation model and analysis model we observe in TSM cannot be compensated by larger sample size.

We can easily obtain regression coefficients $a^{(s)}(t)$ on the association between the curves and the binary outcome from our simulations using $\boldsymbol{\alpha}$ obtained from JM, JMCC, and TSM. We present the $a^{(s)}(t)$ estimates in Figure 3.1. The true coefficients are also plotted in the figure. The figure shows functional coefficient estimates between JM and JMCC are close to each other and to the truth. Functional coefficient estimates for TSM are further from the truth, especially for the second curve.

To quantify the deviance of the $a^{(s)}(t)$ estimates from the true $a^{(s)}(t)$, we obtain mean integrated squared errors (MISE) of $a^{(s)}(t)$ from JM, JMCC, and TSM, from

Table 3.4: Mean integrated squared errors from 500 MC datasets of sample size 500 and 1000, with 30% of $\mathbf{Y}^{(2)}$ missing on average.

	n = 500			n = 1000		
Function	JM	JMCC	TSM	JM	JMCC	TSM
$a^{(1)}(t)$	0.00233	0.00207	0.00047	0.00042	0.00048	0.00094
$a^{(2)}(t)$	0.00059	0.00067	0.00461	0.00013	0.00012	0.00605
$\theta^{(1)}(t)$	0.00001	1.52971	0.00001	0.00004	1.53319	0.00004
$\theta^{(2)}(t)$	0.00505	0.68787	0.00009	0.00832	0.69717	0.00003

the same set of simulations, and present them in Table 3.4. The MISE for $\theta^{(s)}(t)$ are also presented in Table 3.4 to show the deviance of the $\theta^{(s)}(t)$ estimates from JM, JMCC, and TSM. TSM has the largest MISE for $a^{(2)}(t)$. Furthermore, increasing the sample size only makes TSM's MISE of $a^{(s)}(t)$ larger. This trend is a reverse of what is observed in MISE of $a^{(s)}(t)$ from JM and JMCC, which decreases as the effective sample size increases. Uncongeniality between the imputation model and the analysis model in TSM is the likely reason behind TSM's large MISE of $a^{(2)}(t)$, suggesting TSM is inappropriate for the analysis. MISE of $a^{(s)}(t)$ from JMCC is comparable to MISE of $a^{(s)}(t)$ from JM, but MISE of $\theta^{(s)}(t)$ from JMCC is larger than MISE of $\theta^{(s)}(t)$ from JM and does not seem to decrease as sample size increases. This suggests exclusion of 30% of the sample in JMCC leads to biased estimates of $\theta^{(s)}(t)$. But for this simulation setting, the estimates of $a^{(s)}(t)$ from JMCC do not seem to be affected by exclusion of 30% of the sample.

To assess the performance of our proposed method under simulation settings that mimic what we observe in the motivating data, we conduct another set of simulations. We begin data generation by simulating the underlying true curves for 500 subjects using the model given in (3.2). We set the number of cubic B-spline basis functions for the simulated datasets to be 10. We evaluate the basis functions at 40 time points ($m_1 = m_2 = 20$), which we use as the 40×10 design matrix $\mathbf{B}$ in (3.1) for all subjects. To simulate the coefficients $\boldsymbol{\beta}_i$ corresponding to the basis functions from the model

given in (3.4), we let $\boldsymbol{\beta}_0$ be the vector (6 13 9 7 4 10 10 7 5 3), and let $\boldsymbol{\Sigma}_\beta$ be a 10×10 autoregressive covariance matrix with the diagonal elements (16 81 49 25 9 100 144 81 36 9) and $\rho = 0.8$. We simulate the noise $\boldsymbol{\epsilon}_i$ from the model given in (3.5) with $\sigma^2 = 2$. Then we simulate the curves $\mathbf{Y}$ from the model given in (3.1) using $\mathbf{B}$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$. To simulate $W_i (i = 1, \cdots, n)$ from the model given in (3.6), we first simulate Z_i using the model given in (3.8), where $\alpha_0 = -2.22$ and $\boldsymbol{\alpha} =$ (−.09 −.12 .01 .01 .01 .05 .05 .04 .03 .02).

To assess the performance of estimation via our joint-modeling approach (JM) described in Section 3.2, we fit 4 models, each with different number of basis functions, and use minimum DIC to select the best model. In the first model, we set the number of basis as the true value, which is 10. In the second model, we under-specify the number of basis as 8. In the remaining two models, we over-specify the number of basis as 12 or 14. We fit the data to each of the four models, assuming the missing data mechanism to be ignorable. In the DA algorithm of our joint-modeling approach, we run each MCMC chain for 15000 iterations and keep every 5^{th} sample after 10000 iterations to ensure convergence of the parameters and minimize autocorrelation between consecutive samples.

To evaluate the performance of our proposed method, we compare our results against results from JMCC and TSM, which are introduced and described in our previous simulations. We fit 4 models each for JMCC and TSM, varying the number of basis functions from 8, 10, 12, and 14, and use minimum DIC to select the number of basis for each method separately. As another comparison method, we apply the data to the method proposed by Ferrarty and Vieu (2006) implemented by Febrero-Bande and Oviedo de la Fuente (2012) in the R package 'fda.usc', which only uses complete cases. We present selection of the number of basis functions in JM, JMCC, and TSM from our simulation in Table 3.5.

For our proposed method, minimum DIC chooses 10 basis in 472 (94.4%) MC datasets, 12 basis in 13 (2.6%) MC datasets, and 14 basis in 15 (3.0%) MC datasets.

Table 3.5: Proportion of time each number of basis functions is selected by minimum DIC in JM, JMCC, and TSM. The true number of basis functions is 10.

	Number of Basis Functions			
	8	10	12	14
JM	0.0%	94.4%	2.6%	3.0%
JMCC	100.0%	0.0%	0.0%	0.0%
TSM[1]	100.0%	0.0%	0.0%	0.0%

[1]DIC for TSM is based on its analysis model.

For JMCC, minimum DIC chooses 8 basis in all 500 MC datasets. This demonstrates that in the presence of missing data, minimum DIC is a reasonable method for selecting the number of cubic B-spline basis functions in our joint-modeling approach. The fact that minimum DIC consistently select the model that under-specify the number of basis functions in JMCC may be indication that some features of the curves are not captured by JMCC when 30% of the data are excluded.

As a main focus of the analysis, we want to visualize and compare the regression coefficients $a^{(s)}(t)$ on the association between the curves and the binary outcome from our simulations. We obtain $a^{(s)}(t)$ using $\boldsymbol{\alpha}$ obtained from JM, JMCC, and TSM. When the number of basis deviate from the true value, it provides outlying estimates for $a^{(s)}(t)$, the regression coefficient between curves and the outcome. The mean $a^{(s)}(t)$ estimates are biased due to the outlying curves. To obtain $a^{(s)}(t)$ estimates more representative of the true $a^{(s)}(t)$, we used the median $a^{(s)}(t)$ estimates from the 500 MC datasets. We present the $a^{(s)}(t)$ estimates in Figure 3.2. The true coefficients are also plotted in the figure.

As expected, the estimates for JM are close to the true $a^{(s)}(t)$. However, the exclusion of 30% of the sample and under-specifying the number of basis functions by 2 in JMCC does not seem to increase bias of $a^{(s)}(t)$ estimates from JMCC compared to JM. Under-specifying the number of basis functions by 2 in TSM also does not seem to increase bias of its $a^{(s)}(t)$ estimates. Estimates from Ferraty and Vieu's method,

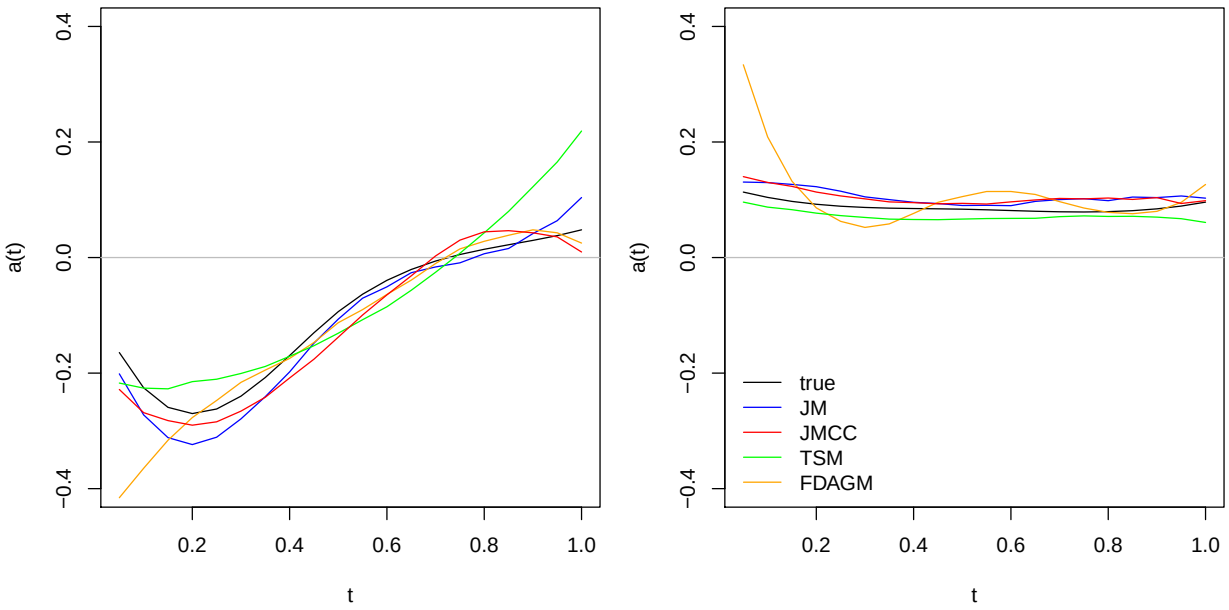

Figure 3.2: True and estimated functional coefficients of the first and second curves from our proposed joint-modeling approach (JM), joint-modeling approach on complete cases (JMCC), two-stage modeling approach (TSM), and Ferraty and Vieu's method (FDAGM).

which also excludes 30% of the sample, are further away from the true coefficients compared to estimates from JM and JMCC.

Table 3.6: Median integrated squared errors from 500 MC datasets of sample size 500, with 30% of $\mathbf{Y}^{(2)}$ missing on average

Function	JM	JMCC	TSM	FDAGM
$a^{(1)}(t)$	0.46	0.27	0.20	0.71
$a^{(2)}(t)$	0.28	0.12	0.06	0.39
$\theta^{(1)}(t)$	4.51	24.67	14.05	-
$\theta^{(2)}(t)$	9.50	12.78	10.63	-

To quantify the deviance of the $a^{(s)}(t)$ estimates from the true $a^{(s)}(t)$, we obtain median integrated squared errors of $a^{(s)}(t)$ from JM, JMCC, TSM, and FDAGM, from the simulation results, and present them in Table 3.6. We also present median integrated squared errors of $\theta^{(s)}(t)$ from JM, JMCC, and TSM in Table 3.6 to quantify the deviance of the estimation of the true curves. Estimates on $\theta^{(s)}(t)$ from FDAGM cannot be obtained. Median integrated squared errors of $a^{(s)}(t)$ is smaller in JM than

FDAGM, indicating JM to have better performance compared to FDAGM. JMCC and TSM have smaller median integrated squared errors of $a^{(s)}(t)$ compared to JM. However, both JMCC and TSM have larger median integrated squared errors of $\theta^{(s)}(t)$ than JM, suggesting JM's estimation of the true curves is more accurate.

3.4 Renal Study

Our work is motivated by a renal study conducted in the Division of Nuclear Medicine at Emory University, aimed at improving renal image interpretations by radiologists. The renal study data consists of renal curves and renal obstruction diagnosis of 163 kidneys from 77 patients. Renal curves from kidneys of the same patient are considered to be independent of each other. The observations were taken at the same 59 time points for the first curve and the same 40 time points for the second curve, with 15 to 30 second intervals between the time points, across the subjects. Two kidneys with extreme curves are excluded from our study. Eight of the kidneys (4.9%) are missing the second curve. The observations range from 0 to 197,295, which means covariance matrix calculations based on the data is computationally expensive. To lower computational cost and ensure numerical stability, we transform the data by dividing the observations by 1000.

Our interest is to find the association between renal curves and renal obstruction status. The outcome of renal obstruction diagnosis is binary, with 0 being non-obstructed and 1 being obstructed. Thus our parameters of interest are the intercept α_0, and the functional coefficients $a^{(1)}(t)$ and $a^{(2)}(t)$, which are associated with the first and second renal curves, respectively.

To estimate the parameters of interest, we apply our joint-modeling method (JM) to the renal study data, and utilize DIC to select the number of basis functions for the renal curves. As shown in Figure 3.3, DICs indicate $k = 8$ to be the appropriate

Table 3.7: Parameter estimates from analysis of renal study data via joint modeling approach (JM), joint modeling approach on complete cases only (JMCC), and two-stage modeling approach (TSM)

	JM		JMCC		TSM			JM		JMCC		TSM	
Par	Est	SD	Est	SD	Est	SD	Par	Est	SD	Est	SD	Est	SD
α_0	-2.305	0.634	-1.930	0.486	-1.976	0.994	σ^2	1.876	0.022	2.902	0.035	3.109	0.037
α_1	0.058	0.282	0.415	0.398	0.573	0.909	β_{01}	-0.013	0.528	-0.103	0.540	-0.009	0.541
α_2	-0.323	0.352	-0.751	0.516	-0.953	1.157	β_{02}	-0.043	0.681	0.033	0.721	-0.010	0.717
α_3	0.637	0.687	1.403	1.016	1.968	2.511	β_{03}	-0.039	0.481	0.084	0.506	-0.005	0.496
α_4	-0.571	0.897	-1.381	1.224	-2.051	3.035	β_{04}	-0.020	0.270	0.008	0.287	-0.004	0.282
α_5	0.005	0.100	-0.002	0.122	-0.040	0.152	β_{05}	-0.024	1.087	0.023	1.136	-0.168	1.099
α_6	-0.081	0.233	-0.053	0.277	0.012	0.406	β_{06}	-0.001	0.813	0.018	0.835	-0.136	0.811
α_7	0.281	0.343	0.446	0.555	0.446	0.725	β_{07}	0.004	0.591	0.016	0.605	-0.099	0.585
α_8	0.059	0.558	-0.204	0.782	-0.243	1.165	β_{08}	0.002	0.343	0.011	0.349	-0.062	0.338

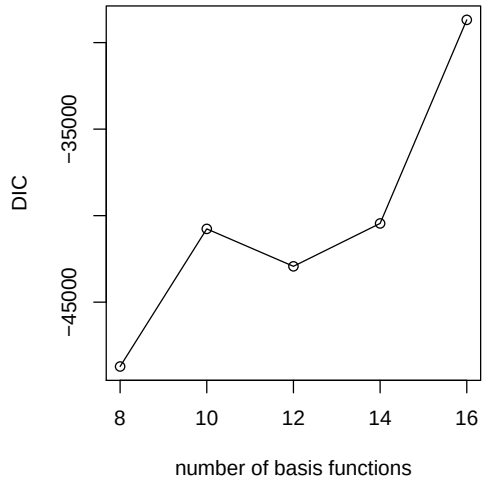

Figure 3.3: Deviance information criteria (DIC) of joint models with different number of cubic B-spline basis functions for renal study data, where the number of basis functions for first and second curve are equal.

number of basis functions for the renal curves. We then apply joint-modeling method to only the complete cases (JMCC), and two-stage modeling method (TSM) to renal study data for comparison, assuming the same 8 basis functions as we do in our joint-model selected by DIC. Table 3.7 show parameter estimates of α_0, as well as estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_0$, and σ^2, along with their standard deviations, from the analysis. Compared to JMCC and TSM, JM has the smallest standard deviations of each parameter of interest. The parameter estimates for α_0, $\boldsymbol{\alpha}$ and σ^2 for JMCC and TSM are closer to each other than estimates from JM, suggesting the two methods have similar performance. When we examine the parameter estimates for $\boldsymbol{\beta}_0$, which are coefficients corresponding to the basis functions from centered curves, we notice that for JM, the $\boldsymbol{\beta}_0$ estimates are all close to zero as we expected. However, $\boldsymbol{\beta}_0$ estimates from JMCC are larger, maybe due to excluding incomplete cases from the analysis which may bias the results. $\boldsymbol{\beta}_0$ estimates associated with the second curve from TSM are larger than estimates from JM and JMCC, which may be attributed to uncongeniality between imputation model and analysis model in TSM. These observations regarding $\boldsymbol{\beta}_0$ estimates have led us to believe that parameter estimates of $\boldsymbol{\alpha}$, which are used to

compute $a^{(1)}(t)$ and $a^{(2)}(t)$, and of α_0 may be biased in JMCC and TSM.

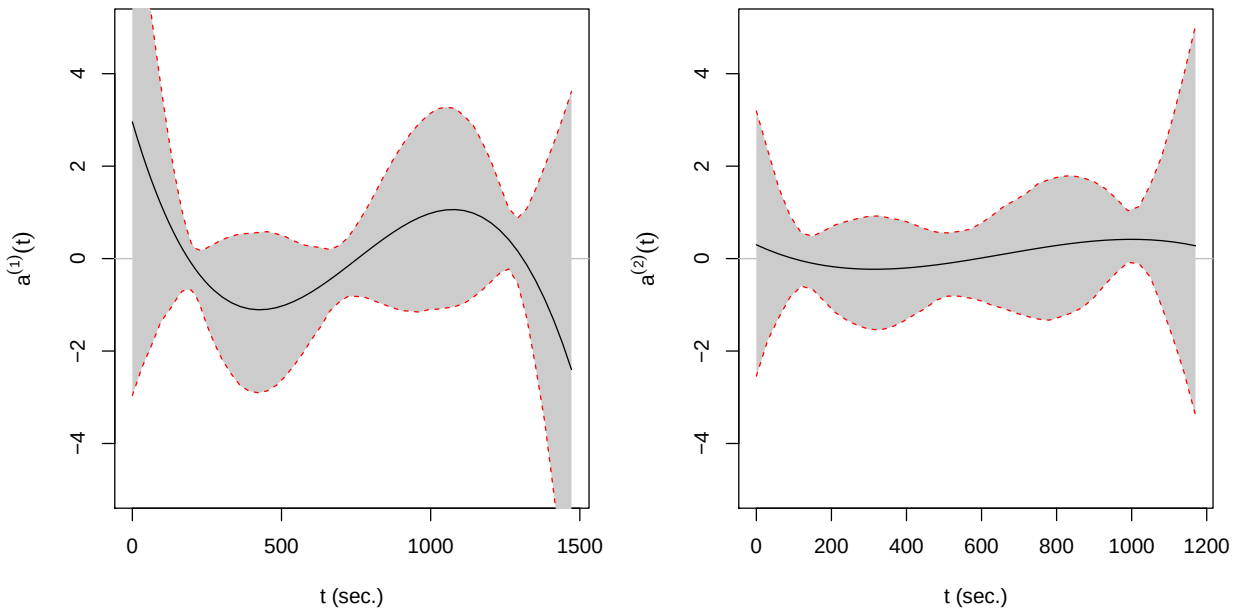

Figure 3.4: Functional coefficients and 95% credible intervals of the coefficients of the association between renal curves and kidney obstruction from joint modeling approach (JM).

As a main focus of the analysis, we want to visualize the functional coefficients $a^{(1)}(t)$ and $a^{(2)}(t)$ from JM. We present the $a^{(1)}(t)$ and $a^{(2)}(t)$ estimates in Figure 3.4. Here the dashed lines are the 95% credible intervals of the estimates. The figure shows that the coefficients are cubic in shape and have larger magnitude for the first curve and smaller magnitude for the second curve but are not significant.

We also want to compare the functional coefficients $a^{(1)}(t)$ and $a^{(2)}(t)$ from JM to those from JMCC and TSM. As an additional comparison method, we apply the renal studies data to the method proposed by Ferrarty and Vieu (2006) implemented by Febrero-Bande and Oviedo de la Fuente (2012) in the R package 'fda.usc'. The estimated functional coefficients using JM, JMCC, TSM, and Ferrarty and Vieu's method (FDAGM) are shown in Figure 3.5. Functional coefficients of the four methods presented are all similar in pattern and magnitude, suggesting they had reasonable results. The functional coefficients estimated by JMCC and TSM are close in

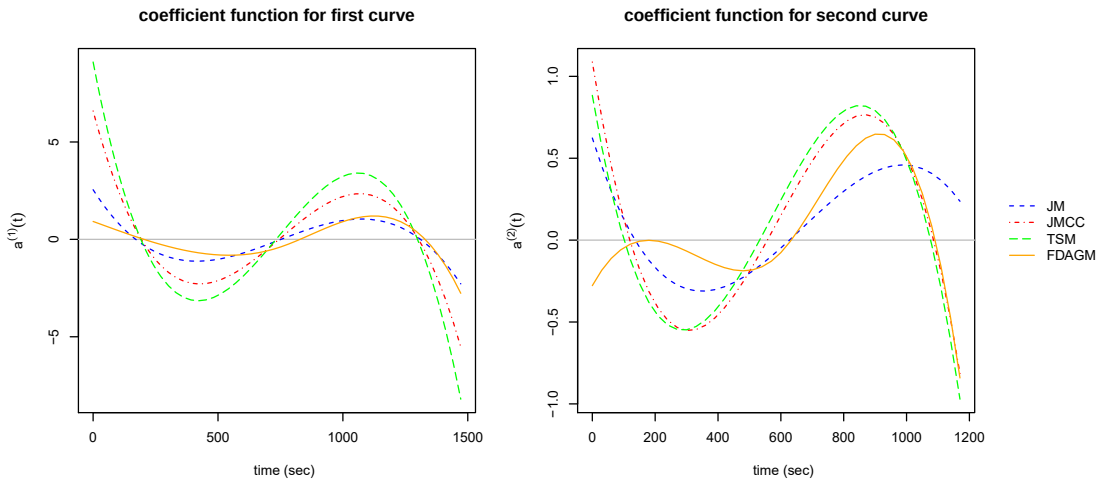

Figure 3.5: Functional coefficients of the association between renal curves and kidney obstruction from joint modeling approach (JM), joint modeling approach on complete cases only (JMCC), two-stage modeling approach (TSM), and a FGLM method by Ferrarty and Vieu (FDAGM).

both first curve and second curve, however, these estimates may be biased because estimates of $\boldsymbol{\alpha}$ may be biased. Functional coefficients estimated by JM are closer to FDAGM for both curves, indicating the two methods have similar performances.

3.5 Discussion

A notable obstacle in the functional data analysis of biomedical data is how to handle missing data. In this Chapter, we provided a joint-modeling approach that estimates the association between functional predictors and a binary outcome while incorporating information from all available data. We conducted simulation studies that compared our proposed method to a two-stage alternative and another alternative where only data from complete cases were considered. In the first simulation study, we fit our method using different number of basis functions and assuming the knots were equally spaced, then evaluated deviance information criteria (DIC) from the fit models. We correctly chose the model with true basis functions when we selected the model with the minimum DIC. In the second simulation study, we assumed the

true number of basis functions and location of knots have been selected and applied joint-modeling approach, two-stage approach, and analysis using only complete cases with the true number of basis functions. In the third simulation study, we evaluated the performance of our proposed method against two-stage approach, and two methods using only complete cases under simulation setting that mimicked the renal study data. This simulation demonstrated with missing data selection of basis based on minimum DIC had some issues. It chose the correct value 94.4% of the time, but other times it was away from that value. The simulation results showed our method had smaller relative biases and mean integrated squared errors than the methods in comparison. The performance of our method improved as we increased the effective sample size. We also applied our method and the comparison methods to the motivating renal study data as an application. In our simulation studies, we included the basis functions we used to simulate the data as a candidate, thus we were able to accurately select the optimal basis functions most of the time. In reality, basis function selection is much more complex. Minimum DIC may be a reasonable method for basis selection in our joint-modeling approach in the presence of missing data, but the sensitivity of results to the choice of basis functions need further investigation in a data analysis. We will explore an alternative approach of basis function selection in the next Chapter.

Chapter 4

Handling missing data in generalized functional linear models through functional principal component analysis with application to renal studies

4.1 Introduction

Functional data analysis has become popular in recent years with the accumulation of large data sets from rapid growth of technology. The nature of the functional data analysis is quite complex and poses unique challenges (Wang et al., 2015). In the Chapter 3, we proposed a joint-modeling approach capable of handling missing data in functional regression, where the response is binary and the predictors are two functional curves, and in some cases one is missing. We employ a Bayesian hierarchical model for jointly modeling the functional curves that are measured with error and the binary outcome, in which the association between the noise-free curves and the outcome is of interest. We model functional curves using cubic B-spline basis functions and use deviance information criterion (DIC) to select number of basis functions. Many authors have used spline basis functions in analyzing functional data (Ramsay and Silverman, 2002). For example, James (2002) used a generalized linear model for linking the outcome and functional covariate which is expressed in terms of a known number of cubic splines. In these aforementioned methods, a critical challenge is to appropriately determine the number of basis function and locations of the knots. This problem remains an open question.

When too many basis functions are selected, the model may overfit the data and the computational cost maybe expensive. Yet when few basis functions are selected, they may not represent the functional data well (Bhattacharya and Dunson, 2011). The functional predictors in the model have infinite dimensions intrinsically, and the optimal combinations of basis functions are unknown. The possible combinations of basis functions that can represent the functional predictors are infinite, so an exhaustive search is not practical. Although splines have been widely applied to model functional data in functional data analysis, there is no guarantee that the cubic splines or equidistant knots are optimal in every type of functional data, so natural basis or B-spline basis of different degrees or unequal spacing between its knots may

be more appropriate in other cases (Yao and Lee, 2008). This places a burden on the analyst to choose the appropriate basis functions for the data. Another challenge lies in how to compare between models with different basis functions and what criteria to facilitate our model selection. Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC), a measure of model assessment and comparison which can be extended to Bayesian hierarchical models with missing data (Celeux et al., 2006). Our experience with our data as well as others (Gelman et al., 2014) suggest that these data driven methods may provide some guidance but they do not always perform well.

Alternatively, many authors have used functional principal component analysis (FPCA) for functional data analysis to explain major source of variation in a sample of random curves (Shang, 2011). FPCA is a popular tool for dimension reduction in functional data and the top few FPCS are usually chosen to explain the variation in functional data. FPCA's popularity is partly attributed to its ability to facilitate conversion of the inherently infinite-dimensional functional data to a finite-dimensional score vector (Wang et al., 2015). In FPCA, the functional data are viewed as realizations of a L^2 stochastic process $\theta(\cdot)$ defined on an interval $(0, T)$ with mean $\mu_\theta(t) = E(\theta(t))$ and covariance $cov(\theta(t_1), \theta(t_2))$. A countable sequence of uncorrelated random variables known as functional principal component scores, or simply scores, can be used to express the underlying stochastic process with mild assumptions. In practical applications, the scores are often truncated to a finite vector, which can then be used in multivariate data analysis.

There are many methods for constructing FPCs in functional data analysis (Wang et al., 2015), some of which are conveniently implemented by an R package (Goldsmith et al., 2016). Yao et al. (2005a) first smooth the covariance function of the functional data, then perform eigendecomposition of the smoothed covariance function. Xiao et al. (2013, 2016) achieve covariance smoothing via a sandwich smoother, which is a

fast penalized spline bivariate smoother. Di et al. (2009) and Goldsmith et al. (2013) use a different method for covariance smoothing, by penalized splines via a mixed model. Huang et al. (2008) approximates the functional data matrix using penalized rank one approximation, then apply singular value decomposition on the rank one matrix to obtain the right singular vectors. These methods and the availability of software provide convenient tools for data practitioners to perform FPCA. In addition, FPCA can be considered as a data-driven method for estimation of basis functions, which we investigate through simulation and a renal study.

Specifically, we apply FPCA as a means for choosing the basis to represent the subject-specific variability among the curves. Each curve is approximated by a linear combination of the top few functional principal components (FPCs), the subject-invariant functions to express functional curves, that explain most of the variation in functional curves (say 95%). We employ the FPCs to address the association between noise-free curves and the binary outcome via a functional generalized linear model (FGLM). This allows the simultaneous estimation of association between the true curve and the binary outcome with the allowance made for the accommodation for missing data in the second curve. Compared to the FGLM that uses cubic B-spline basis functions in Chapter 3, this approach chooses a very small number of FPCs that represents the variability in the functional predictors. It eliminates the need to select basis functions through fitting multiple models, significantly reducing the computational cost of our joint-modeling approach. One additional advantage of this method is that the analysis may also provide insight to the pattern of functional covariates which can be interpreted via FPCs. In Section 4.2 we introduce the FPCAs and our modeling framework. In Section 4.3 we present a simulation study for comparing the choice of cubic B-spline basis to FPCs in our joint-modeling approach. In Section 4.4, we provide a comprehensive analysis of our motivated data set, renal study, using FPCAs and compare the results to Chapter 3 method. Finally, we conclude with a

discussion in Section 4.5.

4.2 Methodology

4.2.1 Data Structure and Model

Let $\mathbf{Y}_{(m\times n)} = \begin{pmatrix} \mathbf{Y}_1 & \cdots & \mathbf{Y}_n \end{pmatrix}$ be the measurements taken at m time points for n individuals. Let $\mathbf{Y}_i$ be the element of $\mathbf{Y}$ and $\mathbf{Y}_i$ denote the vector of measurement of the i^{th} subject. Let $\mathbf{Y}^{(1)}_{i(m_1\times 1)}$ be the first curve and $\mathbf{Y}^{(2)}_{i((m_2)\times 1)}$ be the second curve after, $(m_1 + m_2 = m)$. We write $\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^{(1)} \\ \mathbf{Y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} Y_{i1}^{(1)} & \cdots & Y_{im_1}^{(1)} & Y_{i1}^{(2)} & \cdots & Y_{im_2}^{(2)} \end{pmatrix}^T$. Let $\Delta = \begin{pmatrix} \Delta_1 & \cdots & \Delta_n \end{pmatrix}^T$ be the vector of indicator functions with $\Delta_i = 1$ if $\mathbf{Y}_i^{(2)}$ was observed. We let $\mathbf{Y}_{obs}$ represent the observed curves, and $\mathbf{Y}_{mis} = \{\mathbf{Y}_i^{(2)} : \Delta_i = 0\}$ represent the unobserved curve data. Finally, we let W_i be the binary outcome for individual i, and $\mathbf{W}_{(n\times 1)} = \begin{pmatrix} W_1 & \cdots & W_n \end{pmatrix}$.

Given we sample an individual curve with error over distinct time points $t_{i1}, \cdots, t_{im_i}$ (James and Hastie, 2001), we use the following model to represent the curve data:

$$Y_{ij}^{(s)} = \theta_i^{(s)}(t_j) + \epsilon_{ij}^{(s)},\ i = 1, \cdots, n,\ j = 1, \cdots, m_s,\ s = 1, 2, \tag{4.1}$$

where $Y_{ij}^{(s)}$ is the observed value of the s^{th} curve for subject i at time point t_j, $\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_i^{(1)} \\ \boldsymbol{\epsilon}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \epsilon_{i1}^{(1)} & \cdots & \epsilon_{im_1}^{(1)} & \epsilon_{i1}^{(2)} & \cdots & \epsilon_{im_2}^{(2)} \end{pmatrix}^T$ represents measurement errors, $\theta_i^{(s)}(t_j)$ represents the true value of curve for subject i at time point t_j. A common approach models $\theta_i^{(s)}(t_j)$ as

$$\theta_i^{(s)}(t_j) = \sum_{l=1}^{k_s} \beta_{il}^{(s)} b_l^{(s)}(t_j) = \mathbf{b}^{(s)T}(t_j)\boldsymbol{\beta}_i^{(s)}, \tag{4.2}$$

where $\mathbf{b}^{(s)}(t_j) = \left(b_1^{(s)}(t_j) \cdots b_{k_s}^{(s)}(t_j) \right)^T$ are the k_s cubic spline basis functions with equally-spaced knots, $s = \{1, 2\}$. The choice of the number of basis functions and the locations of the knots are critical components of this method.

To overcome the difficulties in deciding the basis functions for $\theta_i^{(s)}(t_j)$, we employ functional principal component analysis (FPCA) to select a small number of FPCs that can explain sufficient variability in the curves. We consider $\theta_i^{(s)}(t_j)$ as realizations of a L^2 stochastic process $\theta(\cdot)$ defined on an interval $(0, T)$ with mean $\mu_\theta(t) = E(\theta(t))$ and covariance $cov(\theta(t_1), \theta(t_2))$. We let $\mu_\theta^{(s)}(t_j)$ be the mean curve at time point t_j, by Karhunen-Loeve Theorem, we can express the centered curve $\theta_i^{(s)}(t_j) - \mu_\theta^{(s)}(t_j)$ in terms of FPCs, $\theta_i^{(s)}(t_j) = \mu_\theta^{(s)}(t_j) + \sum_{l=1}^{\infty} \xi_{il}^{(s)} \phi_l^{(s)}(t_j)$, where $\xi_{il}^{(s)}$ is the l^{th} score for subject i and $\phi_l^{(s)}(t_j)$ is the l^{th} FPC evaluated at time t_j, $s = \{1, 2\}$.

When $\sum_{l=k_s^*+1}^{\infty} \xi_{il}^{(s)} \phi_l^{(s)}(t_j)$ explain a negligible amount of variability in the curves, we can approximate $\theta_i^{(s)}(t_j)$ by $\theta_{i,k_s^*}^{(s)}(t_j)$, which is expressed using only the first k_s^* FPCs:

$$\theta_i^{(s)}(t_j) \approx \theta_{i,k_s^*}^{(s)}(t_j) = \mu_\theta^{(s)}(t_j) + \sum_{l=1}^{k_s^*} \xi_{il}^{(s)} \phi_l^{(s)}(t_j) = \mu_\theta^{(s)}(t_j) + \boldsymbol{\phi}_{k_s^*}^{(s)T}(t_j) \boldsymbol{\xi}_i^{(s)}, \tag{4.3}$$

where $\boldsymbol{\phi}_{k_s^*}^{(s)T}(t_j) = \left(\phi_1^{(s)}(t_j) \cdots \phi_{k_s^*}^{(s)}(t_j) \right)^T$ are the first k_s^* FPCs, and $\boldsymbol{\xi}_i^{(s)}$ is the vector of corresponding scores for subject i, $s = \{1, 2\}$.

When $Y_{ij}^{(s)}$ is already centered, that is, $\mu_\theta^{(s)}(t_j) = 0$, then (4.3) can be seen as a variation of (4.2), with a different set of basis functions, where each basis function is a linear combination of spline basis functions. We can employ existing FPCA methods to obtain top few FPCs, $\boldsymbol{\phi}_{k_1^*}^{(1)}(t)$ and $\boldsymbol{\phi}_{k_2^*}^{(2)}(t)$, that explains 95% of the variability in the curves. Three methods from the R package "refund" can be applied to our motivating data with ease. The first method (FACE) smooths the covariance function in the functional data via a sandwich smoother then estimates the FPCs and scores by eigendecomposition (Xiao et al., 2013, 2016). The second method (SC) smooths the

covariance function by penalized splines via a mixed model before eigendecomposition (Di et al., 2009; Goldsmith et al., 2013). The third method (SSVD) approximates the functional data matrix using penalized rank one approximation, then apply singular value decomposition on the rank one matrix to obtain the right singular vectors, which are its FPCs (Huang et al., 2008).

We let $\mathbf{\Phi}_{(m\times k)}$ be the block diagonal matrix that consists of the FPCs of first and second curve for a kidney, specifically, $\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}^{(2)} \end{pmatrix}$ where $\mathbf{\Phi}^{(1)}$ corresponds to the first curve and $\mathbf{\Phi}^{(2)}$ corresponds to the second curve, and

$$\mathbf{\Phi}^{(s)} = \begin{pmatrix} \phi_1^{(s)}(t_1) & \cdots & \phi_{k_s^*}^{(s)}(t_1) \\ \vdots & \ddots & \vdots \\ \phi_1^{(s)}(t_{m_s}) & \cdots & \phi_{k_s^*}^{(s)}(t_{m_s}) \end{pmatrix}, s = 1, 2.$$

We denote by $\boldsymbol{\xi}_i = \begin{pmatrix} \boldsymbol{\xi}_i^{(1)} \\ \boldsymbol{\xi}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \xi_{i1}^{(1)} & \cdots & \xi_{ik_1^*}^{(1)} & \xi_{i1}^{(2)} & \cdots & \xi_{ik_2^*}^{(2)} \end{pmatrix}^T$ the corresponding vector of scores for $\mathbf{\Phi}$. We assumed the true curve for the i^{th} kidney to follow a multivariate normal distribution

$$\boldsymbol{\xi}_i \sim N_k(\boldsymbol{\xi}_0, \mathbf{\Sigma}_\xi)$$

based on some hyperparameters $\boldsymbol{\xi}_0 = \begin{pmatrix} \boldsymbol{\xi}_0^{(1)} \\ \boldsymbol{\xi}_0^{(2)} \end{pmatrix}$ and $\mathbf{\Sigma}_\xi = \begin{pmatrix} \mathbf{\Sigma}_\xi^{11} & \mathbf{\Sigma}_\xi^{12} \\ \mathbf{\Sigma}_\xi^{21} & \mathbf{\Sigma}_\xi^{22} \end{pmatrix}$.

We also assumed the following model for the measurement error,

$$\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \sigma^2 \mathbf{I}_m)$$

with σ^2 as its variance, which we assumed to be constant across all kidneys, and that $\boldsymbol{\xi}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ since true curves should not impact measurement errors. We assumed the

hyperparameters $\boldsymbol{\xi}_0$ and σ^2 to have non-informative and improper priors $\pi(\boldsymbol{\xi}_0, \sigma^2) \propto \mathbf{1}$ and the hyperparameter Σ_ξ to have an Inverse-Wishart prior $\boldsymbol{\Sigma}_\xi \sim \mathcal{W}^{-1}(\mathbf{I}_k, k)$ where $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ has scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν. These assumptions about the prior specifications of the hyperparameters helped facilitate our model-fitting procedure.

In lieu of the unknown true renal obstruction, we use consensus ratings of three expert readers as our outcome, with $W_i = 1$ for obstruction and $W_i = 0$ for non-obstruction. We assumed a probit model for W_i and modeled W_i using latent variable $Z_i (i = 1, \cdots, n)$, which is associated with true renal obstruction status for individual i, by

$$W_i = \begin{cases} 1 & if \quad Z_i > 0 \\ 0 & if \quad Z_i \leq 0 \end{cases}, \tag{4.4}$$

with Z_i following Gaussian distribution $Z_i \sim \mathcal{N}(\mu_i, 1)$ where the mean μ_i of Z_i is modeled by the functional model

$$\mu_i = \alpha_0 + \int a^{(1)}(t)\theta_i^{(1)}(t)\mathrm{d}t + \int a^{(2)}(t)\theta_i^{(2)}(t)\mathrm{d}t \tag{4.5}$$

and where $\int a^{(s)T}(t)\theta_i^{(s)}(t)\mathrm{d}t$ represent coefficient of association integrated over the first ($s = 1$) or second ($s = 2$) curve. We use orthonormal cubic spline basis functions $\boldsymbol{\phi}_{k_s^*}^{(s)}(t)$ we used in (4.2) to model the coefficient $a^{(s)}(t)$ as $a^{(s)}(t) = \boldsymbol{\alpha}^{(s)T}\boldsymbol{\phi}_{k_s^*}^{(s)}(t)$, where $\boldsymbol{\alpha}^{(s)}$ are the coefficients corresponding to $\boldsymbol{\phi}_{k_s^*}^{(s)}(t)$. Since $\boldsymbol{\phi}_{k_s^*}^{(s)}(t)$ are orthonormal,

$$\int a^{(s)}(t)\theta_i^{(s)}(t)\mathrm{d}t = \int \boldsymbol{\alpha}^{(s)T}\boldsymbol{\phi}_{k_s^*}^{(s)}(t)\boldsymbol{\phi}_{k_s^*}^{(s)T}(t)\boldsymbol{\xi}_i^{(s)}\mathrm{d}t = \boldsymbol{\alpha}^{(s)T}\left(\int \boldsymbol{\phi}_{k_s^*}^{(s)}(t)\boldsymbol{\phi}_{k_s^*}^{(s)T}(t)\mathrm{d}t\right)\boldsymbol{\xi}_i^{(s)} = \boldsymbol{\alpha}^{(s)T}\boldsymbol{\xi}_i^{(s)},$$

then from (4.5) we have

$$\mu_i = \alpha_0 + \boldsymbol{\alpha}^{(1)T}\boldsymbol{\xi}_i^{(1)} + \boldsymbol{\alpha}^{(2)T}\boldsymbol{\xi}_i^{(2)} = \alpha_0 + \boldsymbol{\alpha}^T\boldsymbol{\xi}_i, \tag{4.6}$$

where $\boldsymbol{\alpha}$ is a column vector of length k and $\boldsymbol{\alpha}^T = \begin{pmatrix} \boldsymbol{\alpha}^{(1)T} & \boldsymbol{\alpha}^{(2)T} \end{pmatrix}$.

The missing data present in the curves adds complication to the model-fitting. Conventional methods ignore subjects with missing curves, resulting in loss of information and potentially biased estimates. To incorporate kidneys with missing curves, we follow the Bayesian data augmentation (DA) algorithm (Tanner and Wong, 1987) that we adopted to fit our model in Chapter 3. The posterior samples of $\boldsymbol{\alpha}$ can be combined to form our estimates of $\boldsymbol{\alpha}$, which can be interpreted directly in terms of the FPCs, or used in estimating the coefficient functions $a^{(1)}(t)$ and $a^{(2)}(t)$.

4.3 Simulation Study

In this section, we assess the performance of our proposed joint-modeling approach with different choice of basis functions including by FPCA via a simulation study. We first study the impact of misspecifying the number of basis functions when using cubic B-spline basis functions. We then compare results to the case where basis functions are selected by each of three FPCA methods described in Section 4.2.

We begin data generation by simulating the underlying true curves for 500 subjects using the model given in (4.2). We use cubic B-spline basis functions with intercept and 0 interior knot for the first and second curve, resulting in a total number of basis functions of 8. We simulate the curves using the same models from (4.1) and (4.2) given parameters mentioned in Simulaion Studies in Chapter 3. Next, we simulate the outcome $W_i (i = 1, \cdots, n)$ by generating Z_i from $\mathcal{N}(\mu_i, 1)$ using the model given in (4.6) and (4.4). We consider 100 Monte Carlo simulations.

To study the impact of misspecifying the number of basis functions on the estimation of the parameter of interest ($a^{(s)}(t)$ in (4.5)), we consider three cases. In the first case (CBS8), we set the number of basis as the true value, which is 8. In the second case (CBS10), we misspecify the number of basis as 10. In the third case (CBS12),

we highly over-specify the number of basis as 12. We use the posterior samples from the three models to obtain our $a^{(s)}(t)$ estimates. These three $a^{(s)}(t)$ estimates, along with the true curves, are plotted in Figure 4.1.

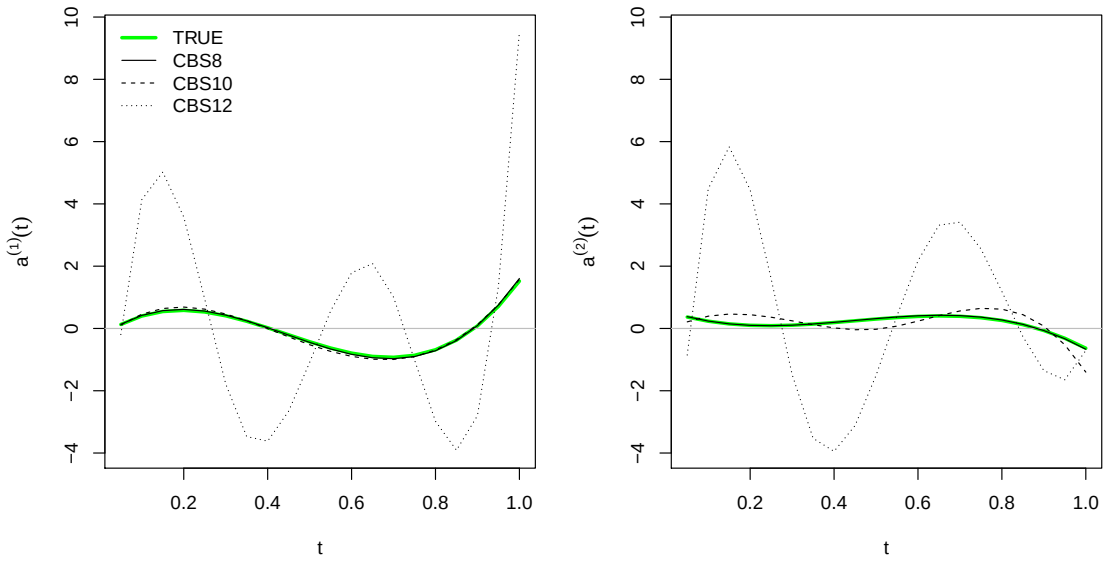

Figure 4.1: True and estimated functional coefficients of the first and second curves by CBS8, CBS10, and CBS12.

Figure 4.1 shows that when the basis functions are fixed at true values, the true and estimated curves are very close, indicating unbiased results. When the basis functions are misspecified, there is a bias, which gets much higher when the specified number of basis functions deviates from the true value. These results emphasize the need for selecting correct number of basis functions for unbiased estimation.

In order to assess the performance of the estimation via FPC basis, we use the same simulation setting. After selecting FPCs that explain at least 95% of variance in functional data, and treating the FPCs as basis functions, we fit the joint model. Three methods (FACE, SC, and SSVD) are used for selecting the FPCs. For FACE and SC, covariance-matrix smoothing is achieved by using 10 spline basis. To display the impact of estimation of $a^{(s)}(t)$, the functional coefficients, we plot the estimated functional coefficients by three methods for FPCA (Figure 4.2) along with the true curve. Biases in $a^{(s)}(t)$ estimates are small when FPCs are used as basis, close to

the case when the true number of basis is specified for cubic B-spline basis. This demonstrates that applying FPCA to select basis functions in our joint-modeling approach is a valid approach with reasonable results.

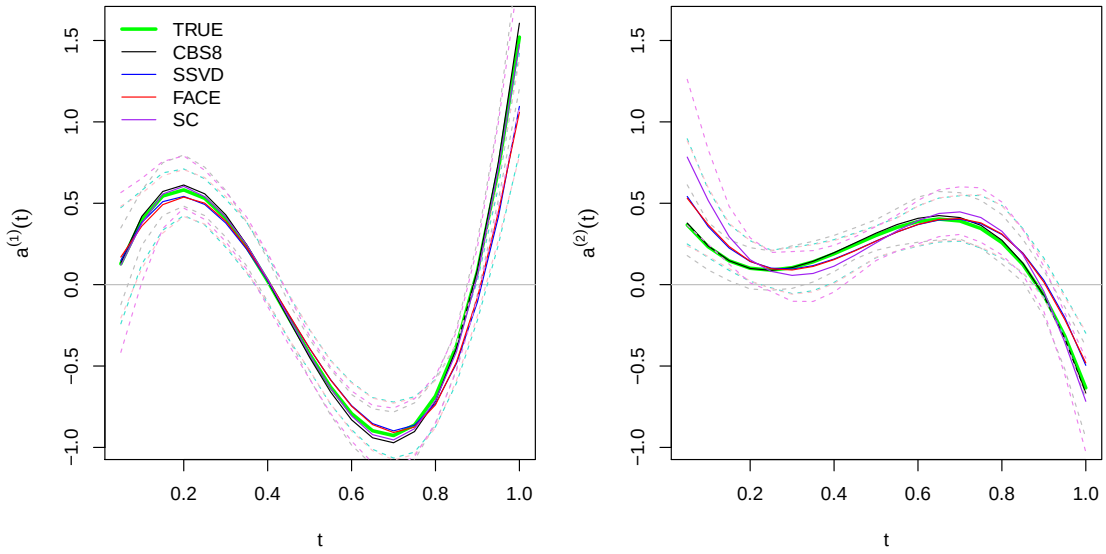

Figure 4.2: True and estimated functional coefficients of the first and second curves. The dashed curves are the 95% credible intervals of the estimated functional coefficients.

4.4 Renal Study

Renal scans play an important role in the determination of kidney obstruction. Without adequate monitoring, patients with known or susceptible obstruction in one or both kidneys could potentially lose the affected kidneys and become reliant on kidney dialysis to maintain their livelihood, putting a high strain on themselves and their loved ones, both physically and financially. When a patient known or suspected to have renal obstruction is referred to a nuclear medicine clinic, renal scans are performed to help the radiologist evaluate possible obstruction in the patients kidneys. Unfortunately, a large percentage of the renal scans performed in the United States are interpreted at sites that perform fewer than 3 studies per week and are inter-

preted by radiologists who have less than 4 months of training in nuclear medicine (IMV, 2003). Lack of training and limited experience coupled with the demands of interpreting a large variety of complex imaging studies at ever faster rates increases the error rate of the diagnosis (Taylor et al., 2008a).

With the aim of improving renal image interpretations by radiologists, the Division of Nuclear Medicine at Emory University conducted a renal study. The renal study data consists of renal curves and renal obstruction diagnosis of 163 kidneys from 77 patients. The observations from the renal curves were counts of a tracer measured at 59 time points for the first curve and 40 time points for the second curve. The time points at which the observations were taken are identical across subjects and are separated by 15 to 30 second intervals. Renal curves from kidneys of the same patient are considered to be independent of each other. Eight of the kidneys (4.9%) are missing the second curve. The observations range from 0 to 197,295, which means covariance matrix calculations based on the data is computationally expensive. To lower computational cost and ensure numerical stability, we apply transformation on the data by dividing the observations by 1000. The outcome of renal obstruction diagnosis is binary, with 0 being non-obstructed and 1 being obstructed. Out of 163 kidneys, 32 (19.6%) are obstructed. Since FPCA is not robust against outliers (Wang et al., 2015), two kidneys with extreme curves are excluded from our study.

The first and second curves for each kidney are shown in Figure 4.3. The first curves are convex in shape and are either strictly increasing or peaks at some time point before 1000 seconds. The second curves are either decreasing or flat. The mean first and second curves for obstructed kidneys and non-obstructed kidneys are shown in Figure 4.4. The figure shows mean first curves are close between non-obstructed kidneys and obstructed kidneys. The figure shows a larger difference in the mean second curves, where the mean curves are decreasing and are separated.

As a descriptive analysis, we first investigate the pattern of relationship between

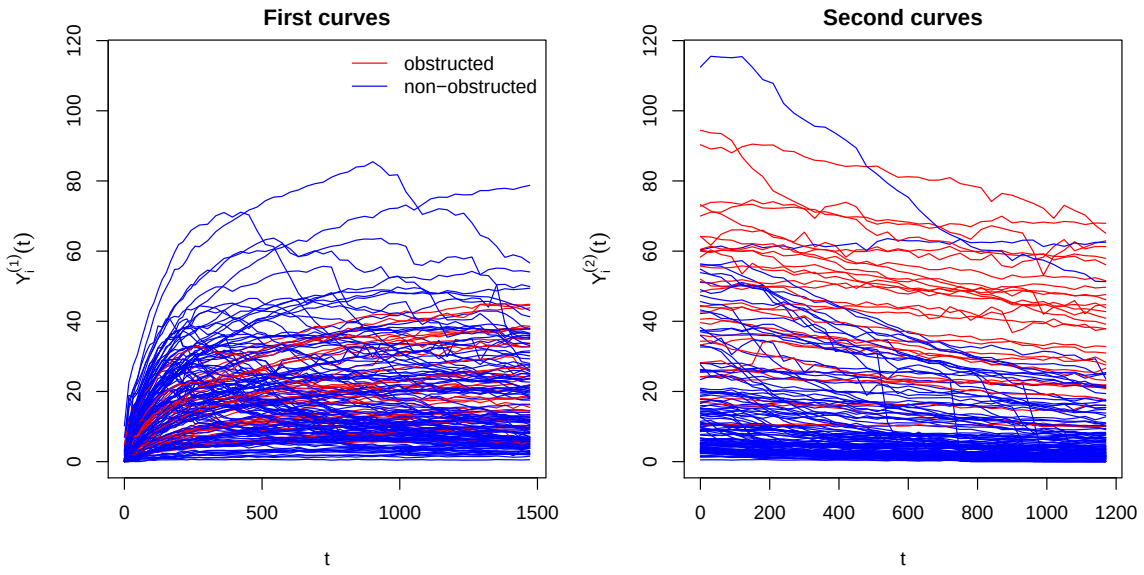

Figure 4.3: First and second renal curves.

the outcome, kidney obstruction, and the functional covariate by a crude analysis. We group the functional covariate data according to time intervals of 20 groups. The average counts are regressed on kidney obstruction in a generalized linear model with Probit link. We then obtain estimated regression coefficients on the average counts and 95% confidence intervals of the estimated regression coefficients, which are presented in Figure 4.5. Coefficient estimates from the regression seem to suggest an oscillating pattern around zero for both curves.

In order to estimate the association between renal curves and kidney obstruction using FPCA-based approach, we first select FPCs that explain at least 95% of the variability in the functional data. For the covariance-matrix smoothing in FACE and SC, we use a large number of basis, 35 to be exact. After enforcing orthonormality on the FPCs, the three methods become equivalent in their choice of FPCs. Therefore, we only use one set of FPCs (from SC) in subsequent analyses.

The functional part of each principal component is illustrated in Figure 4.6, by adding or subtracting a suitable multiple of the l^{th} FPC $\phi_l^{(s)}(t)$ to the mean s^{th} curve $\mu_\theta^{(s)}(t)$, which are defined in equation (4.3) and are estimated from the data. The

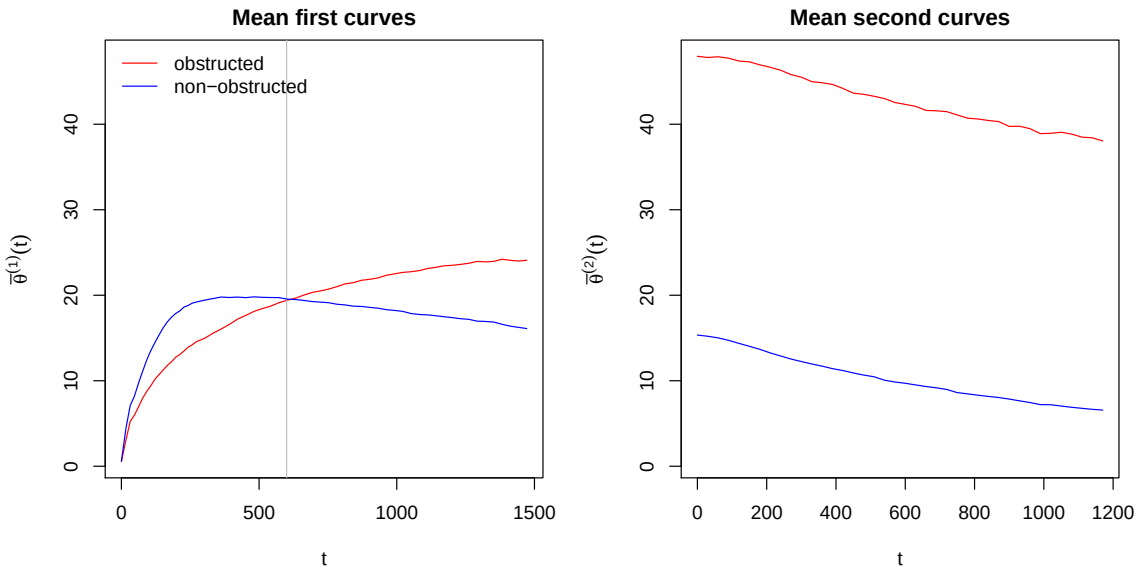

Figure 4.4: Mean first and second curves.

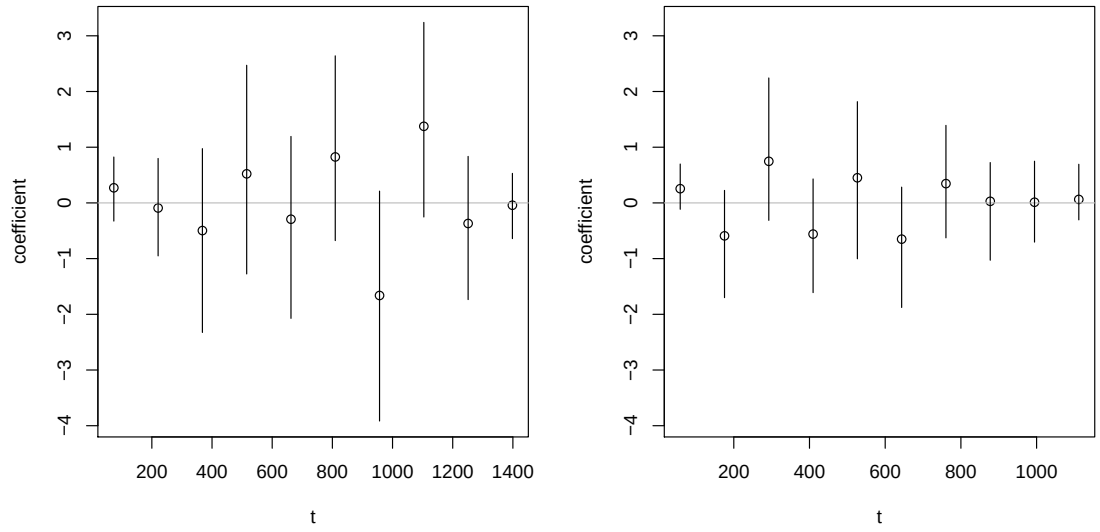

Figure 4.5: Estimated coefficients from a GLM with Probit link and 20 covariates.

FPCs are included in Appendix B. Now we consider Figure 4.6 in detail. The solid lines denote $\mu_\theta^{(s)}(t)$, the dashed lines correspond to adding $\phi_l^{(s)}(t)$ to $\mu_\theta^{(s)}(t)$, and the fine dotted lines correspond to subtracting $\phi_l^{(s)}(t)$ from $\mu_\theta^{(s)}(t)$. The first principal component for the first curve accounts for 85% of the variability in the observed first curves and is strictly positive. The variability associated with this component increases as the mean curve increases, then plateaus after the mean curve reaches

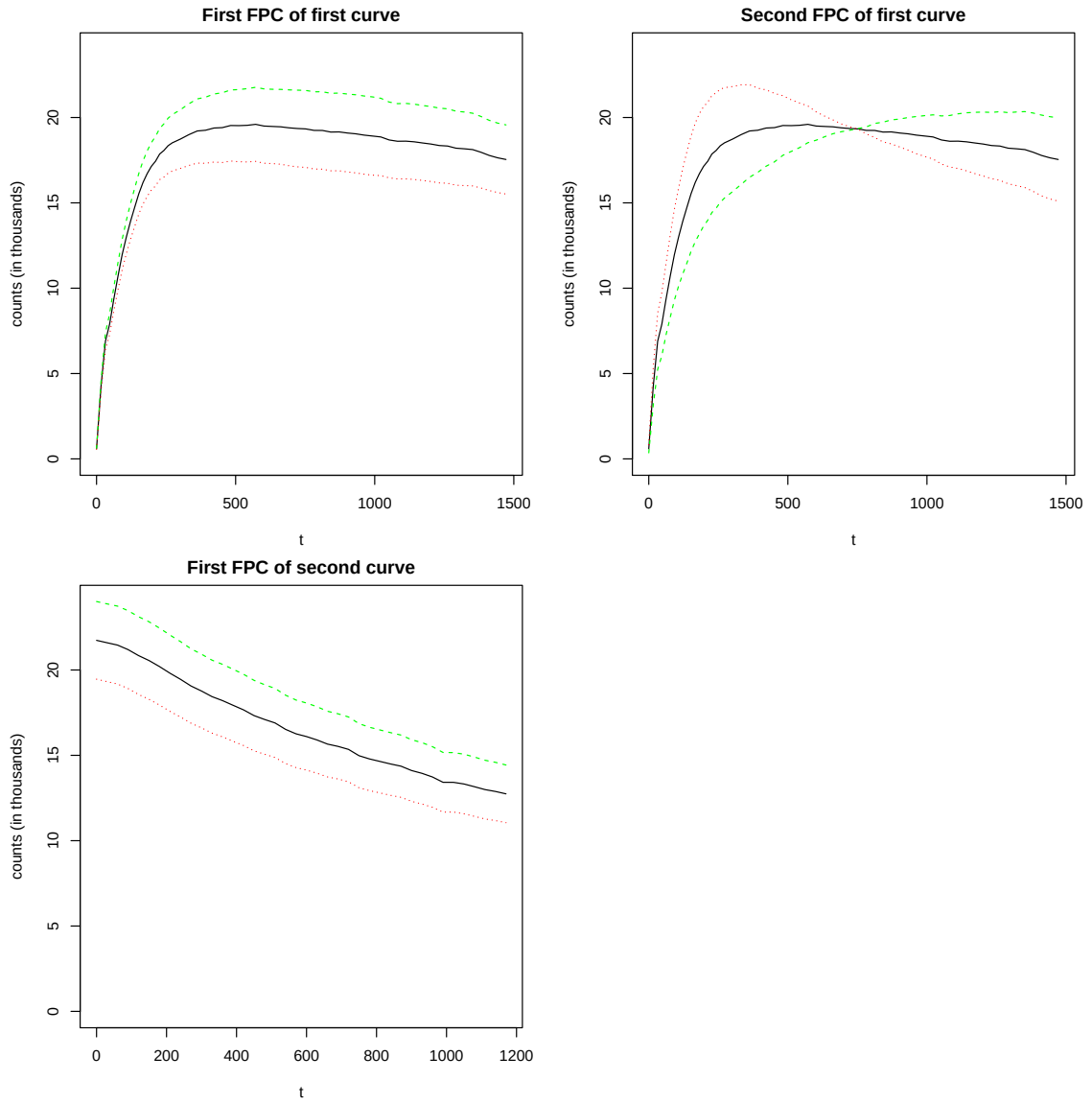

Figure 4.6: The mean first and second renal curves and the effects of adding and subtracting a multiple of each functional principal component from SC. The top panel is for the first curve and the bottom panel is for the second curve. The solid lines denote mean curves, the dashed lines correspond to adding a multiple of a functional principal component the mean curve, and the fine dotted lines correspond to subtracting a multiple of a functional principal component from the mean curve.

its peak. Positive scores on this component is associated with higher than average curve values (counts) in first curve over time. The second component for the first curve accounts for 14% of the variability in the observed first. The second component seems to depict differences between mean curve of first curves for obstructed kidneys

and mean curve of first curves for non-obstructed kidneys. Positive score in this component is associated with lower than average counts until approximately time 700 seconds and higher than average counts after approximately time 700 seconds. The first component for the second curve accounts for 97% of the variability in the observed second curves and is strictly positive and decreasing. Positive score in this component is associated with higher than average counts in second curve over time.

To obtain regression coefficients on the association between the renal curves and kidney obstruction, we fit the data from the renal study using $\boldsymbol{\phi}_2^{(1)}(t)$ and $\phi_1^{(2)}(t)$ obtained above. We present the $a^{(s)}(t)$ estimates in Figure 4.7. The estimated functional coefficients when using FPC basis show a negative association between the first 1400 seconds of the first curve and kidney obstruction. The association between the first curve and kidney obstruction is not significant except between time points 500 seconds and 800 seconds where the difference between mean curves of first curves for obstructed and non-obstructed kidneys are small. The figure also shows a significant, constant and positive association between the second curve and kidney obstruction. This indicates that given similar first curves, an obstructed kidney is more likely to have a higher second curve than a non-obstructed kidney. These findings are consistent with the clinical expectation of nuclear medicine experts at Emory.

For comparison, we also apply cubic B-spline basis to fit our data. We use the same number of basis functions we selected via deviance information criteria (DIC) in Chapter 3. As another comparison to our methods, we apply the functional data analysis by generalized model (FDAGM) method (Febrero-Bande and Oviedo de la Fuente, 2012) to our data. The $a^{(s)}(t)$ estimates from model with FPC basis and from the two comparison models are presented in Figure 4.8. Estimated coefficient functions for cubic B-spline basis model and FDAGM are larger in magnitude compared to model with FPC basis, and are cubic (or quartic) in shape. The patterns we observe in the two comparison models are likely attributed to the small number

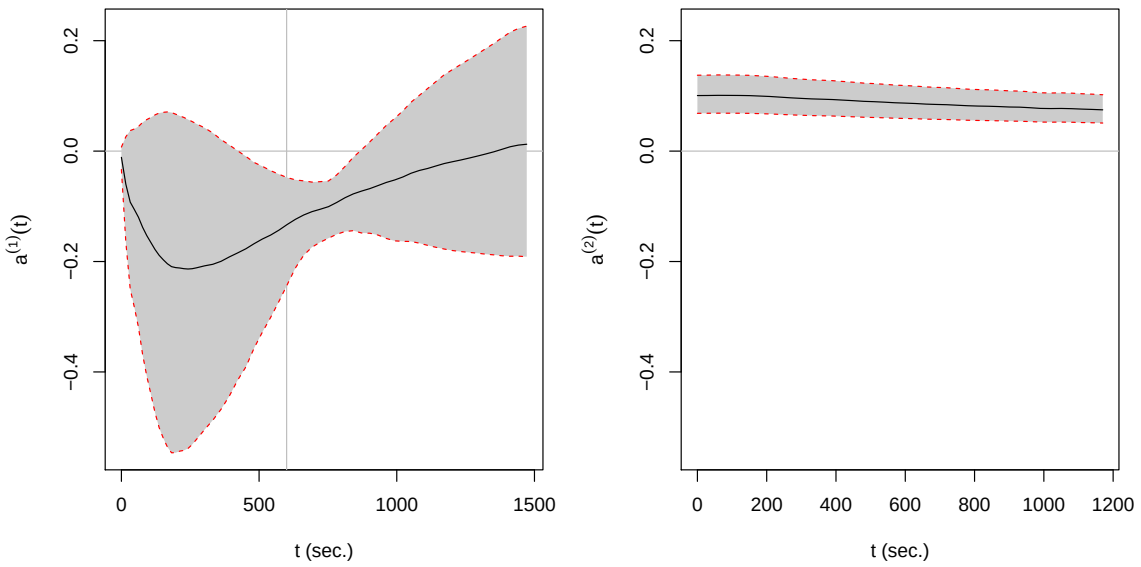

Figure 4.7: Estimated coefficient functions and their 95% credible intervals from joint models using FPC basis from SC.

of cubic B-spline basis functions we use to fit the model.

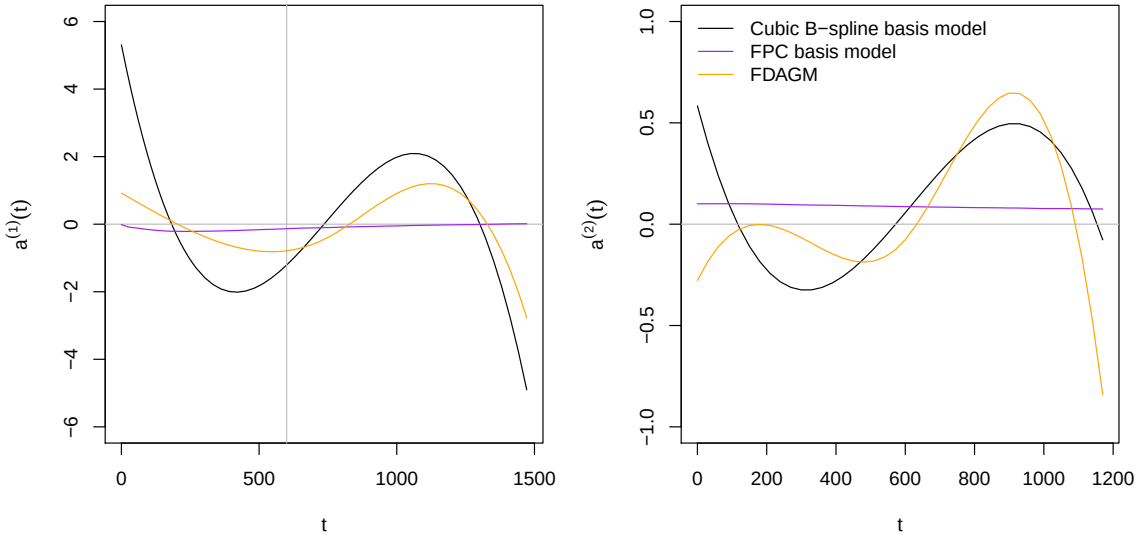

Figure 4.8: Estimated coefficient functions from joint models using cubic B-spline basis, FPC basis, and from FDAGM.

The estimated coefficient functions from our joint models are different in shape and magnitude depending on whether we use cubic B-spline basis or FPC basis to model the curves, raising the question whether cubic B-spline basis and FPC basis

fit the data well. As an assessment for the models, we can plot the predicted vs. observed probabilities for joint models using FPC basis and using cubic B-spline basis to compare the goodness-of-fit of each model. We start by generating the predicted probability of obstruction for each kidney by each of these 2 models. Next, we order the kidneys by their predicted probability and group them into 16 groups of size 10 or more. For each group, the predicted probability is the mean predicted probability of the kidneys in that group, and the observed probability is the proportion of obstruction observed from kidneys in that group.

We display the differences between observed probability of each group and the predicted probability from each of the two models visually in Figure 4.9. We also assess the goodness-of-fit of FDAGM as a comparison. The figure shows that all three models fit the data reasonably well. The predicted probability of obstruction from the model using FPC basis is slightly closer to the observed probability than predicted probabilities of obstruction from model using cubic B-spline basis and from FDAGM. Figure 4.9 suggests that FPC basis fit the data just as well as cubic B-spline basis. The goodness-of-fit assessment indicates functional principal components can be used as alternatives to cubic B-splines for fitting renal curves when estimating the association between the renal curves and kidney obstruction.

4.5 Discussion

A critical challenge in functional data analysis is to appropriately determine the number of basis functions and locations of the knots when modeling functional data in the presence missing data. Functional data have infinite dimensions intrinsically, and splines have been widely applied to model functional data in functional data analysis. The optimal combinations of spline basis functions required to model the functional data are unknown, and when too many basis functions are selected, the

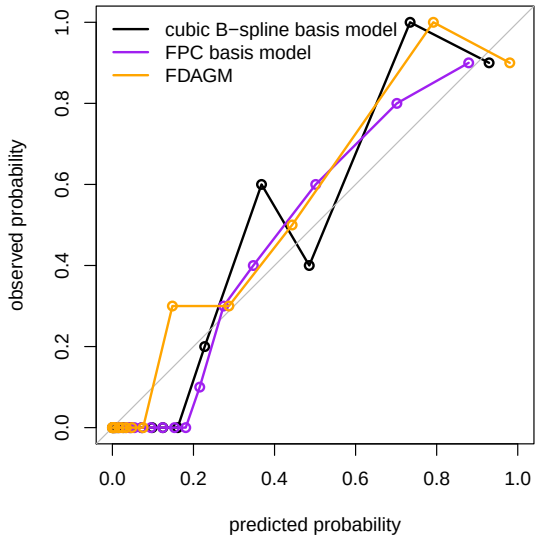

Figure 4.9: Observed vs. predicted probabilities of renal obstruction when using cubic B-spline basis, FPC basis, and by FDAGM.

model may overfit the data and the computational cost maybe expensive. Yet when too few basis functions are selected, they may not represent the functional data well (Bhattacharya and Dunson, 2011). This places a burden on the analyst to choose the appropriate basis functions for the data set he or she works with or may lead to bias in the analysis results. How to compare between models with different basis functions and what criteria to facilitate the model selection also poses a challenge. Although Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC), to assess and compare models, which Celeux et al. (2006) extended to work with Bayesian hierarchical models with missing data, our experience with our data as well as others (Gelman et al., 2014) suggest that these data driven methods may provide some guidance but they do not always perform well. Alternatively, many authors have used functional principal component analysis (FPCA) for functional data analysis to explain major source of variation in a sample of random curves.

In this Chapter, we applied FPCA to select the functional principal components used as basis in our joint modeling approach developed in Chapter 3. Compared to conventional FGLM that uses cubic B-spline basis functions, this approach chooses a

very small number of functional principal components that represents the variability in the functional predictors. It eliminates the need to select basis functions through fitting multiple models, significantly reducing the computational cost of our joint-modeling approach. Our simulation study showed FPCA to be a valid approach for basis selection. Joint models with functional principal component basis produced results comparable to results estimated from fitting the true model. In our data analysis, model with functional principal component basis and model with cubic B-spline basis showed difference in the patterns of their estimated coefficient functions. Our joint model methods based on FPC basis model and cubic B-spline basis model both incorporate subjects with missing functional data, which make them more appropriate for the analysis than the method by Febrero-Bande and Oviedo de la Fuente (2012). However, goodness-of-fit assessment indicated both models to have fit the data reasonably. As was noted in Wang et al. (2015), FPCA is not robust against outliers. Perhaps the presence of undetected outliers in our data contributed to the differences we observed in estimated coefficient functions in our data analysis. Therefore, further investigation of potential outliers in our data is necessary. In general, caution must be used when substituting functional principal components in our joint model approach when outliers are present in the data.

Chapter 5

Future work

In this Chapter, we discuss potential extensions of this dissertation and some directions for future research.

Motivated by a renal study, the methods proposed in Chapters 2, 3, and 4 are only applicable to a special situation, namely informative missing of a second curve. These methods can be extended to multiple curves. The issues such as the order of curves and computational feasibility need to be considered in such modeling.

In this dissertation, we assume the functional data are MAR. As a direction for future research, it would be interesting to incorporate covariate information from the renal study data that were not used in this dissertation, such as age and sex of the patient, for imputing the missing functional data. The inclusion of covariate information in our models may relax the MAR assumption. The relaxation of the MAR assumption allows our proposed methods applicable to situations with general missing-data mechanism.

In Chapter 3 and Chapter 4, we assume probit link between the functional predictors and a binary outcome. We can extend our joint model with other links to accommodate different types of outcomes, such as survival outcome or ordinal categorical data. Extension to ordinal categorical data is straightforward with a cumulative logit link but handling survival data may need further modeling with proportional hazard assumption.

Results in Chapter 2 may be used to impute the second curves, hence facilitating the FPC selection in the presence of missing data. Accommodation of missing data in FPC selection may improve the joint modeling approach for estimating the association between functional predictors and the outcome.

Appendix A

Appendix for Chapter 3

Full conditional likelihoods for additional parameters and hyperparameters:

$$Z_i|W_i = 1, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_i \sim \mathcal{N}\left(\alpha_0 + \boldsymbol{\alpha}^T\boldsymbol{\beta}_i, 1\right) \text{truncated to } (0, \infty)$$

$$Z_i|W_i = 0, \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_i \sim \mathcal{N}\left(\alpha_0 + \boldsymbol{\alpha}^T\boldsymbol{\beta}_i, 1\right) \text{truncated to } (-\infty, 0]$$

$$\boldsymbol{\beta}_0|. \sim \mathcal{N}_k\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}_i, \frac{\boldsymbol{\Sigma}_\beta}{n}\right)$$

$$\boldsymbol{\Sigma}_\beta|. \sim \mathcal{W}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)^T + \mathbf{I}_k, n + k\right)$$

$$\sigma^2|. \sim \textit{Inverse-Gamma}\left(\frac{nm}{2} - 1, \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)\right)$$

where . represents all other parameters and data.

Appendix B

Appendix for Chapter 4

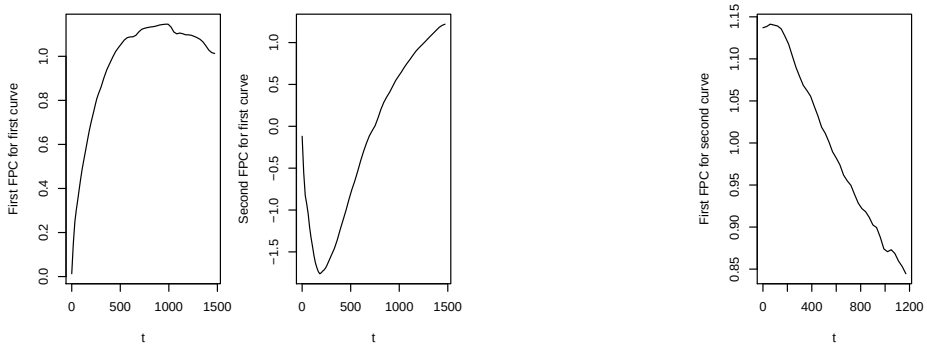

Figure B.1: Functional principal components for first and second curves.

Bibliography

2003 Nuclear Medicine Census Market Summary Report Vol. IV. (2003), Technical report, IMV Medical Information Division, Des Plaines, IL: IMV, Ltd.

Aguilera, A. M., Ocaña, F. a. and Valderrama, M. J. (1997), 'An approximated principal component prediction model for continuous-time stochastic processes', Applied Stochastic Models and Data Analysis **13**(2), 61–72.
URL: *http://doi.wiley.com/10.1002/(SICI)1099-0747(199706)13:2¡61::AID-ASM296¿3.0.CO;2-I*

Bao, J., Manatunga, A., Binongo, J. N. G. and Taylor, A. T. (2011), 'Key variables for interpreting 99mTc-mercaptoacetyltriglycine diuretic scans: Development and validation of a predictive model', American Journal of Roentgenology **197**(2), 325–333.

Barnard, J. and Rubin, D. B. (1999), 'Small-sample degrees of freedom with multiple imputation', Biometrika **86**(4), 948–955.
URL: *http://biomet.oxfordjournals.org/content/86/4/948.abstract*

Bhattacharya, a. and Dunson, D. B. (2011), 'Sparse Bayesian infinite factor models', Biometrika **98**(2), 291–306.

Brick, J. M. and Kalton, G. (1996), 'Handling missing data in survey research.', Statistical Methods in Medical Research **5**(3), 215–238.

Cai, T. T. and Hall, P. (2006), 'Prediction in functional linear regression', Annals of Statistics **34**(5), 2159–2179.
URL: *http://arxiv.org/abs/math/0702650*

Cardot, H., Ferraty, F. and Sarda, P. (1999), 'Functional linear model', Statistics & Probability Letters **45**(1), 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003), 'Spline estimators for the functional linear model', Statistica Sinica **13**, 571–591.
URL: *http://www3.stat.sinica.edu.tw/statistica/oldpdf/A13n31.pdf*

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006), 'Deviance information criteria for missing data models', Bayesian Analysis **1**(4), 651–673.

Chiou, J.-M. and Müller, H.-G. (2009), 'Modeling Hazard Rates as Functional Data for the Analysis of Cohort Lifetables and Mortality Forecasting', Journal of the American Statistical Association **104**(486), 572–585.

Collins, L. M., Schafer, J. L. and Kam, C. M. (2001), 'A comparison of inclusive and restrictive strategies in modern missing data procedures.', Psychological methods **6**(4), 330–351.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38.

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

Fan, J. and Zhang, J.-T. (2000), 'Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data', Journal of the Royal Statistical Society.

Series B (Statistical Methodology) **62**(2), 303–322.

URL: *http://www.jstor.org/stable/3088861*

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012), 'Statistical computing in functional data analysis: the R package fda. usc', Journal of Statistical Software **51**(4), 1–28.

URL: *http://www.jstatsoft.org/v51/i04/paper*

Ferraty, F. and Vieu, P. (2003), 'Curves discrimination: A nonparametric functional approach', Computational Statistics and Data Analysis **44**(1-2), 161–173.

Frank, I. and Friedman, J. (1993), 'A Statistical View of Some Chemomstrics Regression Tools', Technometrics **35**(2), 109–135.

Gelman, A., Hwang, J. and Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', Statistics and Computing **24**(6), 997–1016.

Goldsmith, J., Greven, S. and Crainiceanu, C. (2013), 'Corrected Confidence Bands for Functional Data Using Principal Components', Biometrics **69**(1), 41–51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C. and Reiss, P. T. (2016), refund: Regression with Functional Data. R package version 0.1-16.

URL: *http://CRAN.R-project.org/package=refund*

Green, P. J. and Silverman, B. W. (1993), Nonparametric regression and generalized linear models: a roughness penalty approach, Chapman and Hall, London.

Harel, O. and Zhou, X.-H. (2007), 'Multiple imputation: review of theory, implementation and software.', Statistics in medicine **26**(16), 3057–3077.

Huang, J. Z., Shen, H. and Buja, A. (2008), 'Functional principal components analysis via penalized rank one approximation', Electronic Journal of Statistics

2(March), 678–695.

URL: *http://dx.doi.org/10.1214/08-EJS218\nhttp://arxiv.org/abs/0807.4862*

Hyndman, R. J. and Ullah, Shahid, M. (2007), 'Robust forecasting of mortality and fertility rates: A functional data approach', Computational Statistics and Data Analysis **51**(10), 4942–4956.

James, G. (2002), 'Generalized Linear Models with Functional Predictor Variables', Journal of the Royal Statistical Society, Series B **64**(2), 411–432.

James, G. M. and Hastie, T. J. (2001), 'Functional linear discriminant analysis for irregularly sampled curves', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(3), 533–550.

URL: *http://doi.wiley.com/10.1111/1467-9868.00297*

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

Kneip, A. and Utikal, K. J. (2001), 'Inference for Density Families Using Functional Principal Component Analysis', Journal of the American Statistical Association **96**(454), 519–542.

Laird, N. M. (1988), 'Missing data in longitudinal studies.', Statistics in medicine **7**, 305–15.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/3353609*

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

Little, R. J. A. and Rubin, D. B. (2002), Statistical analysis with missing data, Wiley.

URL: *http://books.google.com/books?id=aYPwAAAAMAAJ*

McCullagh, P. and Nelder, J. A. (1989), Generalized Linear Models, Second Edition.
URL: *http://www.amazon.com/dp/0412317605*

Meng, X.-L. (1994), 'Multiple-Imputation Inferences with Uncongenial Sources of Input', Statistical Science **9**(4), 538–558.
URL: *http://projecteuclid.org/euclid.ss/1177010269*

Montagna, S., Tokdar, S. T., Neelon, B. and Dunson, D. B. (2012), 'Bayesian Latent Factor Regression for Functional and Longitudinal Data', Biometrics **68**(4), 1064–1073.

Müller, H.-G. (2005), 'Functional modelling and classification of longitudinal data', Scandinavian Journal of Statistics **32**(2), 223–240.

Müller, H.-G. and Stadtmüller, U. (2005), 'Generalized functional linear models', The Annals of Statistics **33**(2), 774–805.

Preda, C. and Saporta, G. (2005), 'Clusterwise PLS regression on a stochastic process', Computational Statistics and Data Analysis **49**(1), 99–108.

Preda, C., Saporta, G., Hedi, M. and Hadj, B. E. N. (2010), 'The NIPALS algorithm for missing functional data', Rev. Roumaine Math. Pures Appl. **55**(4), 315–326.

Ramsay, J. and Dalzell, C. (1991), 'Some Tools for Functional Data Analysis', Journal of the Royal Statistical Society. Series B (Methodological) **53**(3), 539–572.

Ramsay, J. O. and Silverman, B. W. (2002), Applied functional data analysis: methods and case studies, Springer-Verlag.

Rubin, D. B. (1987), 'Comment', Journal of the American Statistical Association **82**(398), 543–546.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478461*

Rubin, D. B. and Schenker, N. (1986), 'Multiple imputation for interval estimation from simple random samples with ignorable nonresponse', Journal of the American Statistical Association **81**(394), 366–374.

Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data, first edit edn, Chapman and Hall.

Schafer, J. L. (2003), 'Multiple imputation in multivariate problems when the imputation and analysis models differ', Statistica Neerlandica **57**(1), 19–35.

Shang, H. L. (2011), 'A survey of functional principal component analysis', Computational Statistics & Data Analysis (May).

Silverman, B. W. (1984), 'Spline smoothing: the equivalent variable kernel method', The Annals of Statistics **12**(3), 898–916.

Silverman, B. W. (1985), 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting', Journal of the Royal Statistical Society: Series B (Methodological) **47**(1), 1–52.
URL: *http://www.jstor.org/stable/2345542*

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', Journal of the Royal Statistical Society. Series B: Statistical Methodology **64**(4), 583–616.

Tanner, M. and Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', Journal of the American Statistical Association **82**(398), 528–540.

Taylor, A., Garcia, E. V., Binongo, J. N. G., Manatunga, A., Halkar, R., Folks, R. D. and Dubovsky, E. (2008), 'Diagnostic performance of an expert system for

interpretation of 99mTc MAG3 scans in suspected renal obstruction', J Nucl Med **49**(2), 216–224.

Taylor, A., Manatunga, A. and Garcia, E. V. (2008), 'Decision support systems in diuresis renography', Semin Nucl Med **38**(1), 67–81.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2015), 'Review of Functional Data Analysis', Annu. Rev. Statist. pp. 1–47.
URL: *http://arxiv.org/abs/1507.05135*

Xiao, L., Li, Y. and Ruppert, D. (2013), 'Fast bivariate P -splines: the sandwich smoother', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(3), 577–599.
URL: *http://doi.wiley.com/10.1111/rssb.12007*

Xiao, L., Zipunnikov, V., Ruppert, D. and Crainiceanu, C. (2016), 'Fast covariance estimation for high-dimensional functional data', Statistics and Computing **26**(1-2), 409–421.

Yao, F. and Lee, T. C. M. (2008), 'On knot placement for penalized spline regression', Journal of the Korean Statistical Society **37**(3), 259–267.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a), 'Functional data analysis for sparse longitudinal data', Journal of the American Statistical Association **100**(470), 577–590.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005b), 'Functional linear regression analysis for longitudinal data', The Annals of Statistics **33**(6), 2873–2903.
URL: *http://arxiv.org/abs/math/0603132*