The Pennsylvania State University

The Graduate School

THE RECENT HISTORY FUNCTIONAL LINEAR MODEL AND ITS EXTENSION TO SPARSE LONGITUDINAL DATA

A Dissertation in
Statistics
by
Kion Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010

The dissertation of Kion Kim was reviewed and approved* by the following:

Damla Şentürk
Assistant Professor of Statsitics
Dissertation Advisor, Co-Chair of Committee

Runze Li
Professor of Statistics
Co-Chair of Committee

David R. Hunter
Associate Professor of Statistics

Linda M. Collins
Professor of Human Development and Family Studies and Statistics

Bruce G. Lindsay
Willaman Professor of Statistics and Department Head

*Signatures are on file in the Graduate School.

Abstract

We propose a variant of historical functional linear models for cases where the current response is affected by the predictor process in a window into the past. Different from the rectangular support of functional linear models, the triangular support of the historical functional linear models and the point-wise support of varying coefficient models, the current model has a sliding window support into the past. This idea leads to models that bridge the gap between the varying coefficient models and the functional linear (historic) models for densely measured functional data and longitudinal data. By utilizing one dimensional basis expansions and one dimensional smoothing procedures, the proposed estimation algorithm is shown to have better performance and to be faster than the estimation procedures proposed for historical functional linear models.

We also consider the recent history functional linear models, relating a longitudinal response to a longitudinal predictor. We propose an estimation procedure for recent history functional linear models that is geared towards sparse longitudinal data, where the observation times across subjects are irregular and total number of measurements per subject is small. The proposed estimation procedure builds upon recent developments in literature for estimation of functional linear models with sparse data and utilizes connections between the recent history functional linear models and the varying coefficient models. We establish uniform consistency of the proposed estimators, propose prediction of the response trajectories and derive their asymptotic distribution leading to asymptotic point-wise confidence bands. We include a real data application and simulation studies to demonstrate the efficacy of the proposed methodology.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Functional data refers to data that have dense repetitions on each subject. The basic philosophy of functional data is to think of the dense repetitions as realizations of a stochastic process in time. Functional data is becoming more frequent and functional data analysis is more relevant as it gets easier to collect functional data on a dense grid. Ffunctional data cannot be well-explained by traditional statistical methods due to their special characteristics such as high correlation between the measurements and infinite dimensionality.

Longitudinal data is also encountered frequently in various fields such as behavioral science, biomedical science, economics, and marketing. Longitudinal data consists of repeated observations from multiple subjects that may measured irregularly, i.e., the measurement times for different subjects may be differerent. The most important feature of longitudinal data is the correlation between measurements from the same subject. Repeated measurements for each subject is a common feature of both functional and longitudinal data, where one of the current issues in functional data analysis is to extend the applicability of the models developed for functional data to longitudinal data. The need for a unified theory for longitudinal and functional data is highlighted by Marron (2004), where the differences between the two data types are well summarized in Rice (2004).

Functional linear models are used widely in literature to relate vector- or function-valued predictor and response variables. Müller (2005) summarizes the functional linear models in 4 categories according to the nature of the predictor and response variables included as given in Table 1.1. Here L^2 denotes the space of square-integrable functions on a suitable domain and $\mathbb{R}^d$ denotes the space of d-dimensional real-valued random vectors. In this dissertation, we address the estimation and inference of the functional regression model and its application to longitudinal data. Particularly, our focus will be the case where the predictor and response processes are defined on a common closed time interval, $[0, T]$.

Table 1.1: Classification of the functional linear models

Predictor	Response	Model
L^2	L^2	Functional regression models
L^2	$\mathbb{R}^d$	Generalized functional linear models
$\mathbb{R}^d$	L^2	Functional response models
$\mathbb{R}^d$	$\mathbb{R}^d$	Multivariate multiple regression models

Let $X_i(s), Y_i(t),\ \ s \in [0,T], t \in [0,T]$ be the realizations for the i^{th} subject of the predictor and response stochastic processes $X(s)$ and $Y(t)$, respectively. Assuming that the predictor process affects the response process through a linear model, Bosq (2000), Cardot et al. (1999), Ramsay and Dalzell (1991), and Ramsay and Silverman (2002, 2005) considered the functional linear model given by

$$Y_i(t) = \alpha(t) + \int_0^T \beta(s,t)X_i(s)ds + \varepsilon_i(t), \quad s \in [0,T],\ t \in [0,T], \tag{1.1}$$

where $\alpha(t)$ is the intercept function, $\beta(s,t)$ is the bivariate regression function, and $\varepsilon_i(t)$ is the error function with $E\{\varepsilon_i(t)\} = 0, \text{cov}\{\varepsilon_i(t), \varepsilon_j(t')\} = \sigma_{tt'}$ for $i = j$ and 0 otherwise. In model (1.1), the response at the current time t can be affected by the predictor process from any point in time, including the future values.

For predictive purposes, Malfait and Ramsay (2003) tailored the model in (1.1) to be applicable to situations where the response process $Y(t)$ depends only on the predictor $X(s)$ at times $s \leq t$ as

$$Y_i(t) = \alpha(t) + \int_{s_0(t)}^{t} \beta(s,t) X_i(s) ds + \varepsilon_i(t), \quad s \in [s_0(t), t],\ t \in [0, T], \tag{1.2}$$

where $s_0(t) = max(0, t - \delta)$ and $0 \leq \delta \leq T$. Here, δ denotes the time lag back in time where the predictor process starts affecting the response process. In this model, only the past of the predictor process affects the response at the current time t. Yet another dependence structure was proposed by Hastie and Tibshirani (1993), Staniswalis and Lee (1998) via the varying coefficient models,

$$Y_i(t) = \alpha(t) + X_i(t)\beta(t) + \varepsilon_i(t), \quad t \in [0, T].$$

Different from the above two models, in the varying coefficient models, the current response $Y_i(t)$ is affected only by the current value of the predictor $X_i(t)$. This model can be thought of as a special case of the historical functional linear model.

In this dissertation, we focus on a variant of the historical functional linear model which generalizes the varying coefficient model. In this new model, we allow the current response values to be affected by only the recent past of the predictor process,

$$Y_i(t) = \alpha(t) + \int_{t-\delta_1}^{t-\delta_2} X_i(s)\beta(s,t) ds + \varepsilon_i(t), \quad t \in [\delta_1, T],\ 0 \leq \delta_2 \leq \delta_1. \tag{1.3}$$

Here δ_2, particularly useful for prediction models, allows for a lag for the predictor process to start affecting the response, while δ_1 denotes another lag, beyond which the predictor process does not affect the response. We call this model the recent history functional linear model. This particular model is different from the functional linear model with a rectangular support (1.1), the general historical functional linear model with a triangular support considering the whole past (1.2) and from the varying coefficient model, which has a point support. Instead, the model in (1.3) has a sliding window support, in which the predictor process has an effect on the current response.

To illustrate the cases where model (1.3) would be reasonable, we analyze the relationship between the measurements of rainfall and the flow level of Curlew creek in Florida. A plausible way to model the fluctuations of the flow level would be to assume that the measurements of rainfall from the recent past are affecting the flow level of the river at a given time point. Note that, for this data set, it would not make sense to think that rainfall from future affects the current flow level, nor does one need to consider the whole past of rainfall. It is also not plausible to assume that the current rainfall affects the river flow immediately and hence the proposed model provides a nice generalization and specification of the previously proposed models.

This dissertation is organized in the following manner. In Chapter 2, we give selective literature review on functional linear models, historical functional linear models, and varying coefficient models, which are closely related with the recent history functional linear models. In Chapter 3, we propose an estimation method for the recent history functional linear models for functional data based on one-dimensional basis expansions. In Chapter 4, another estimation method for the recent history functional linear model is proposed for longitudinal data. The efficacy of the method is shown via simulations. In Chapter 5, we address estimation and inference in the recent history functional linear model for sparse longitudinal data via covariance function estimation and basis expansions. We also establish consistency of the estimators and normality of the predicted response trajectories relying on the normality assumption of the principal component scores of the predictor processes. We conclude the dissertation by discussing possible future research projects in Chapter 6.

Chapter 2

Literature Review

In this chapter we review the models that are related to the recent history functional linear model. As the recent history functional linear model is a variant of functional regression model as given in Table 1.1, the focus of this review is the functional linear model with functional response. The functional linear model with scalar response, however, is also extensively reviewed in this chapter, since estimation procedures proposed for it are closely related with those proposed for the functional linear model with functional response. In addition, note that the recent history functional linear model can be seen as a generalization of the historical functional linear model providing more flexibility by introducing an additional lag parameter δ_2. We, therefore, review two estimation methods proposed in literature for the historical functional linear models. We also discuss about the estimation methods for the varying coefficient models, since the estimation procedures proposed for the recent history functional linear models build on those of the varying coefficient models as an intermediate step.

2.1 Functional Linear Models with Scalar Response

For n independent subjects, the functional linear model with scalar response is defined as

$$Y_i = \alpha + \int_{\tau} \beta(s) X_i(s) ds + \varepsilon_i, \ i = 1, \ldots, n, \tag{2.1}$$

where $X_i(t)$ and Y_i are the i^{th} realizations of the predictor process and the scalar response, and α and $\beta(t)$ are the intercept and one-dimensional regression function of interest. Here, τ denotes the domain for the predictor process and the regression function $\beta(t)$ and is usually taken as a closed time interval. The term ε_i represents the i^{th} realization of the i.i.d. random error with zero mean and constant variance σ_ε^2, and is assumed to be independent of $X_i(t)$. The functional linear model given in (2.1) can be thought of as a generalization of the univariate multiple regression to one with infinitely many covariates.

One naive approach in estimating the regression function, $\beta(t)$, is to discretize the function and apply the least squares method considering the model (2.1) as a gigantic multiple regression model. There are, however, three potential problems in this approach. The first one is that the estimated coefficients have extremely large or infinite variance due to too many coefficients to estimate in the model. Secondly, the functional nature of the regression function, such as smoothness, continuity, and differentiability, cannot be preserved with this approach since covariates are exchangeable in a multiple regression model, i.e., exchanging order of the covariates will not affect the estimation results. Lastly, from a practical point of view, it may not be straightforward to discretize the trajectory in the case where the observations are not taken on a common set of grid points, and this problem is intensified when one wants to apply the method to longitudinal data.

More precisely, the first point can be formalized in the following way. For this, let H be a real separable Hilbert space of square integrable functions defined on a suitable domain. Let us define the cross- and auto-covariance functions, $G_{XY}(t) = \text{cov}\{X(t), Y\}$ and $G_X(s,t) = \text{cov}\{X(s), X(t)\}$, respectively. Then, from (2.1), it is easy to show that the covariance functions, $G_{XY}(t)$, $G_X(s,t)$, and regression function $\beta(t)$ are linked as

$$G_{XY}(t) = \int_{\tau} G_X(s,t)\beta(s)ds = (A_G\beta)(t), \tag{2.2}$$

where A_G is a linear integral operator with the auto-covariance function $G_X(s,t)$ as its kernel. In infinite dimensional Hilbert space, the inverse of the linear integral operator, A_G, however, does not exist in general and even if it does, its inverse is not bounded. As a consequence, the estimation of $\beta(t)$ in (2.1) is an ill-posed problem. Further discussions on this can be found in Müller (2005), Dauxois et al. (1982), and Cardot et al. (1999, 2003) and references therein.

To address those problems, various regularization methods including basis expansion, truncation of series expansion, penalized spline, and thresholding techniques are employed in literature. In the following, we review some of the estimation methods proposed for the functional linear model with scalar response. In the following discussion, we assume that the predictor and response processes are all centered, i.e., $E\{X(t)\} = 0$, for all $t \in \tau$, and $E(Y) = 0$ without loss of generality. For the basis expansion approach, Ramsay and Silverman (2005), Cardot et al. (2003), Marx and Eiler (1999), Crambes et al. (2009) proposed estimation procedures based on least squares or penalized least squares method by minimizing

$$Q\{\beta(t)\} = \sum_{i=1}^{n} \left\{ Y_i - \int_{\tau} \beta(s)X_i(s)ds \right\}^2 + \rho f(\beta), \tag{2.3}$$

where $f(\cdot)$ is a penalty function given as a functional that maps the regression function $\beta(t)$ to a real number. Here ρ is a tuning parameter that determines how much penalty will be applied to the sum of squares of error to control the smoothness of the regression function $\beta(t)$. The least squares method minimizes a special case of (2.3) without the penalty term, $f(\beta)$. Depending on how the regression function and predictor processes are expanded, there are two major classes of estimators proposed in literature: 1) those using predetermined basis functions, 2) those using basis functions estimated from the data, namely the eigenbasis.

Ramsay and Silverman (2005) expanded the regression function $\beta(t)$ and the predictor trajectory $X_i(t)$ on two known sets of basis functions $\phi_k(t),\ k = 1, \ldots, K_1$ and $\psi_l(t),\ l = 1, \ldots, K_2$ as $\beta(t) \approx \sum_{k=1}^{K_1} b_k\phi_k(t)$ and $X_i(t) \approx \sum_{l=1}^{K_2} x_{il}\psi_l(t)$, and approximated the integral in (2.3) by $\int_{\mathcal{T}} \beta(s)X_i(s)ds \approx \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} b_k x_{il} \int_{\mathcal{T}} \phi_k(s)\psi_l(s)ds$. They used a penalty on the second derivative of the regression function by considering the penalty of the form, $b^{\mathrm{T}}R_2 b$, where $b = [b_1, \ldots, b_{K_1}]^{\mathrm{T}}$ and the $(i, j)^{th}$ element of R_m is defined as $\int_{\mathcal{T}} \phi_i^{(m)}(s)\phi_j^{(m)}(s)ds$. Here $g^{(m)}$ denotes the m^{th} derivative of the function g. Expanding the predictor process on a set of basis functions, the observations from different subjects are not required to be measured at common time points across different subjects.

Cardot et al. (2003), on the other hand, expanded the regression function on B-spline basis functions of order q with $K - q$ interior equidistant knots as $\beta(t) \approx \sum_{k=1}^{K} b_k\phi_k(t)$ and approximated the integral in (2.3) as $\sum_{k=1}^{K} b_k \int_{\mathcal{T}} X_i(s)\phi_k(s)ds$. They used the penalty function that is similar with but more flexible than the one proposed by Ramsay and Silverman (2005) by putting penalty of the form, $b^{\mathrm{T}}R_m b$, for $m < q - 1$, and m can be chosen considering the purpose of the analysis. The most common choice is $m = 2$ and the ridge type solution can be obtained when $m = 0$.

Another important class of estimator is built upon functional principal component analysis and is studied by Bosq (2000), Cardot et al. (1999), Cardot et al. (2006), and Hall and Horowitz (2007) among many others. As the functional principal component analysis on the predictor process provides a set of orthonormal eigenfunctions, the main idea of this approach is to use the eigenfunctions as basis functions and express the process as a linear combination of the eigenfunctions. More precisely, let us denote the auto-covariance function of $X(t)$ by $G_X(s,t)$ and the cross-covariance function between $X(t)$ and Y by $G_{XY}(t)$. The auto-covariance function of X, $G_X(s,t)$, can be expressed as

$$G_X(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s) \psi_i(t),$$

where $\lambda_i,\ i \in \mathbb{Z}$, is a sequence of eigenvalues and $\psi_i(s),\ i \in \mathbb{Z}$, is a sequence of eigenfunctions. Here, the eigenvalues are ordered from the largest to smallest, i.e., $\lambda_1 > \lambda_2 > \ldots$, and $\int_{\mathcal{T}} \psi_i(s)\psi_j(s)ds = \delta_{ij}$, where δ_{ij} is 1 if $i = j$ and zero elsewhere.

Then, based on Kahrunen-Loéve expansion (1975), the predictor process $X_i(t)$ can be represented as

$$X_i(t) = \sum_{j=1}^{\infty} \zeta_{ij} \psi_j(t),$$

where $\zeta_{ij}, j \in \mathbb{Z}$, is a sequence of random variables, called principal component scores, that satisfy $E(\zeta_{ij}) = 0$ and $\text{var}(\zeta_{ij}) = \lambda_j$. By truncating the basis at a finite number of basis functions, the inverse problem of the auto-covariance function can be surmounted. Estimated from data, a relatively small number of eigenfunctions can approximate the processes effectively compared to the number of standard basis functions such as B-spline, Fourier, and wavelet basis that would have to be used.

Assuming the predictor processes are discretized on a set of m equidistant points, let $\hat{G}_X(s,t) = \dfrac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t)$ and $\hat{G}_{XY}(t) = \dfrac{1}{n}\sum_{i=1}^{n} X_i(t)Y_i$ be the estimates of $G_X(s,t)$ and $G_{XY}(t)$, for $s,t = 1,\ldots,m$. Then, we can define discretized auto- and cross-covariance functions, $\hat{G}_X$ and $\hat{G}_{XY}$, the $m \times m$ matrix with $G_X(s,t)$ as its $(s,t)^{th}$ element and the m dimensional vector with $G_{XY}(t)$ as its t^{th} element, respectively. Performing standard eigen-decomposition on the auto-covariance function, we can obtain $\hat{\phi}_j$, an m-dimensional eigenvector, that corresponds to the j^{th} largest estimated eigenvalue, $\hat{\lambda}_j$, for $j \leq m$. At the same time, for an individual trajectory, $X_i(t)$, we can estimate $\hat{\zeta}_{ij}$, the j^{th} estimated component score. Then, relying on the first K eigen-components, Cardot (1999), Bosq (2000) proposed an estimator given by

$$\hat{\beta} = \sum_{j=1}^{K} \frac{\hat{G}_{XY}^{\mathrm{T}} \hat{\phi}_j}{\lambda_j} \hat{\phi}_j,$$

where $\hat{\beta}$ is the m dimensional vector of discretized regression function. Cardot (2003) improved the above estimator by smoothing $\hat{\beta}$ to get more smooth estimates and better interpretation. Hall and Horowitz (2007) proposed an estimation method based on Tikhonov regularization by suggesting the estimator of the form

$$\widehat{\beta}_\rho = \left(\hat{G}_X + \rho I\right)^{-1} \hat{G}_{XY}$$

where ρ is a ridge parameter. They studied the convergence rate of the estimator in terms of the eigenvalues of the covariance function. Instead of projecting the covariance function on the space spanned by the eigenfunctions, Cardot and Johannes (2010) employed the projection on the space spanned by finite number of orthonormal basis functions such as trigonometric or wavelet basis functions and used a thresholding method when the inverse of the projected covariance function is too large. Based on empirical covariance functions, $\hat{G}_X$ and $\hat{G}_{XY}$, estimated from discretized processes, those methods cannot be applied to irregular data.

2.2 Functional Linear Model with Functional Response

Functional linear models with functional response are relatively less explored in literature compared to the case of the scalar response. Let $X_i(s), s \in [0, S]$ be the i^{th} realization of the predictor process and $Y_i(t), t \in [0, T]$ be the i^{th} realization of the response process, for $i = 1, \cdots, n$. Assuming the predictor process affects the response process through a linear model, Bosq (2000), Cardot et al. (1999), Ramsay and Dalzell (1991), and Ramsay and Silverman (2005) considered the functional linear model

$$Y_i(t) = \alpha(t) + \int_0^S \beta(s,t) X_i(s) ds + \varepsilon_i(t), \quad s \in [0, S],\ t \in [0, T], \tag{2.4}$$

where $\alpha(t)$ is the intercept function, $\beta(s,t)$ is the bivariate regression function and $\varepsilon_i(t)$ is the error function with $E(\varepsilon_i(t)) = 0, \mathrm{cov}(\varepsilon_i(t), \varepsilon_j(t')) = \sigma_{tt'}$ for $i = j$ and 0 otherwise. Note that (2.4) is a general regression model where the entire predictor values $X(s), s \in [0, S]$, have an impact on the value of response process at t.

Ramsay and Silverman (1997) considered the case where both the predictor, $X(s)$, and the response, $Y(t)$, are observed over the same time period $[0, T]$ and both functions are periodic. The model then can be written as

$$Y_i(t) = \alpha(t) + \int_0^T \beta(s,t) X_i(s) ds + \varepsilon_i(t), \quad s \in [0, T],\ t \in [0, T]. \tag{2.5}$$

By changing the domain of the predictor process $[0, S]$ in general functional linear models into the time domain $[0, T]$, this model now became a functional regression model where past, present, and future of the predictor process can have an effect on the response. They discussed the identifiability and estimability of the regression function $\beta(s,t)$, pointing out that the estimation of $\beta(s,t)$ involves infinite number of parameters. Again, we need a regularization method to obtain interpretable regression estimators. The basis expansion, functional principal component analysis, and roughness penalty are the frequently used techniques for regularization.

The main idea for basis expansion methods for the functional linear model with functional response is to expand the regression function $\beta(s,t)$ in terms of K known two-dimensional basis functions $\phi_k(s,t)$ such as $\beta(s,t) \approx \sum_{k=1}^{K} b_k \phi_k(s,t)$. For basis functions, one can use either known basis functions that do not depend on the data or those derived from the given data. Ramsay and Silverman (2005) utilized the expansion of the regression function $\beta(s,t)$ on tensor product basis, which is a product of univariate basis functions and can serve as a basis system for higher dimensional approximation provided that the support is rectangular. For a given time point (s,t), let $\phi(s) = [\phi_1(s), \cdots, \phi_{K_1}(s)]^{\mathrm{T}}$ and $\eta(t) = [\eta_1(t), \cdots, \eta_{K_2}(t)]^{\mathrm{T}}$ be two basis systems with K_1 and K_2 bases respectively. A double expansion of the regression function yields

$$\beta(s,t) \approx \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} b_{kl} \phi_k(s) \eta_l(t) = \phi(s)^{\mathrm{T}} B \eta(t),$$

where B is a $K_1 \times K_2$ matrix of coefficients whose $(l,k)^{th}$ element is b_{lk}. The intercept function $\alpha(t)$ can also be approximated by basis expansion on $\eta(t)$ as

$$\alpha(t) = \sum_{l=1}^{K_2} a_l \eta_l(t) = \eta^{\mathrm{T}}(t) a,$$

where a_l and a are a constant and a vector of coefficients, respectively. For model (2.5), we can obtain an estimate of $\beta(s,t)$ as $\hat{\beta}(s,t) = \phi^{\mathrm{T}}(s)\hat{B}\eta(t)$, by minimizing the function

$$J(a,B) = \int_0^T \sum_{i=1}^{n} \left\{ Y_i(t) - \eta^{\mathrm{T}}(t) a - \int_0^T X_i(s) \phi^{\mathrm{T}}(s) B \eta(t) ds \right\}^2 dt.$$

with respect to a and B. For more details, see Ramsay and Silverman (2005).

Based on the same expansion of the two-dimensional regression function on tensor product basis function, Antoch et al. (2008) applied a smoothing spline method by minimizing, with respect to B,

$$Q(B) = \frac{1}{n}\sum_{i=1}^{n}\int_0^T \left[Y_i^*(t) - \int_0^T X_i^*(s)\phi^{\mathrm{T}}(s)B\eta(t)ds\right]^2 dt + \rho f(m, B), \tag{2.6}$$

where $X_i^*(t)$ and $Y_i^*(t)$ are centralized versions of $X_i(t)$ and $Y_i(t)$, respectively, and

$$f(m, B) = \sum_{i=0}^{m} \frac{m!}{i!(m-i)!}\int_0^T \int_0^T \left[\frac{\partial^m}{\partial s^i \partial t^{m-i}}\phi^{\mathrm{T}}(s)B\eta(t)\right]^2 dsdt.$$

Here, m denotes the order of derivative to which the penalty is applied.

He et al. (2003), on the other hand, considered the expansion of the regression function using the eigenbasis of the predictor and response processes. The smooth random processes, $X(t)$ and $Y(t)$, under the regularity conditions, can be represented as

$$X(t) = \mu_X(t) + \sum_{m=1}^{\infty}\xi_m\psi_m(t),\ Y(t) = \mu_Y(t) + \sum_{k=1}^{\infty}\zeta_k\varphi_k(t)$$

where ξ_m and ζ_k are uncorrelated mean zero functional principal component scores with the second moments equal to the eigenvalues γ_m and η_k for $\sum_{m=1}^{\infty}\gamma_m < \infty$ and $\sum_{k=1}^{\infty}\eta_k < \infty$. Based on the eigenvalues and eigenfunctions obtained by the above decomposition, they came up with the following representation of the regression function,

$$\beta(s, t) = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{E\left\{\xi_m\zeta_k\right\}}{\gamma_m}\psi_m(s)\varphi_k(t). \tag{2.7}$$

They also discussed the existence and uniqueness of the expression given in (2.7) in their paper. For more detail, see He et al. (2003). Note that the cross covariance function of $X(t)$ and $Y(t)$, $G_{XY}(s,t)$, can be expanded as

$$G_{XY}(s,t) = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} E\left\{\xi_m \zeta_k\right\} \phi_m(s)\varphi_k(t).$$

Then, $E\left\{\xi_m \zeta_k\right\}$ has the representation

$$E\left\{\xi_m \zeta_k\right\} = \int_0^T \int_0^T \psi_m(s) G_X(s,t)\varphi_k(t) ds dt,$$

where $G_X(s,t)$ is the auto-covariance function of the predictor process. Using the above representation and finite number of eigenvalues and functions for the predictor and response processes, Yao et al. (2005a) proposed an estimator of the regression function

$$\hat{\beta}(s,t) = \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{\hat{\sigma}_{km}}{\hat{\gamma}_m} \hat{\psi}_m(s)\hat{\varphi}_k(t),$$

where $\hat{\psi}_m(s)$, $\hat{\varphi}_k(t)$ are estimates of $\psi_m(s)$ and $\varphi_k(t)$, $\hat{\sigma}_{km} = \displaystyle\int_0^T \int_0^T \hat{\psi}_m(s)\hat{G}_X(s,t)$ $\hat{\varphi}_k(t) ds dt$ and $\hat{G}_{XY}(s,t)$ is an estimator of $G_{XY}(s,t)$.

2.3 Historical Functional Linear Models

In the case of non-periodic data, from the feed-forward point of view, it is not plausible to use future values of the predictor process in predicting the current value of the response process. Malfait and Ramsay (2005) tailored the model in (2.5) to be applicable to situations where the response process $Y(t)$ depends only on the predictor $X(s)$ at time $s \leq t$ as

$$Y_i(t) = \alpha(t) + \int_{s_0(t)}^{t} \beta(s,t) X_i(s) ds + \varepsilon_i(t), \quad s \in [s_0(t), t],\ t \in [0,T], \tag{2.8}$$

where $s_0(t) = max(0, t-\delta)$ and $0 \leq \delta \leq T$. Without loss of generality, we can drop the intercept function by defining $Y_i^*(t) = Y_i(t) - \bar{Y}(t)$ and $X_i^*(s) = X_i(s) - \bar{X}(s)$

where $\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)/n$ and $\bar{X}(s) = \sum_{i=1}^{n} X_i(s)/n$. Hence the model can be written without the intercept function as

$$Y_i^*(t) = \int_{s_0(t)}^{t} \beta(s,t) X_i^*(s) ds + \varepsilon_i(t), \quad s \in [s_0(t), t], t \in [0, T]. \tag{2.9}$$

While the functional linear models had rectangular support, considering past, present and future values of the predictor process, the historical functional linear model has triangular support, considering the effect of the past and present of the predictor process. Malfait and Ramsay (2005) employed, a triangular basis function instead of a tensor product basis to expand the two dimensional regression function defined on triangular support. This expansion is known as the finite element method and is popular in the numerical solution of partial differential equation boundary value problems. Detailed construction of the basis function is provided in Appendix A.

Let $\phi_k(s,t),\ k = 1, \dots, K$ be the triangular basis functions. Expanding the regression function as $\sum_{k=1}^{K} b_k \phi_k(s,t)$, let us define $\psi_{ik}(t) = \int_{s_0(t)}^{t} X_i^*(s) \phi_k(s,t) ds$. Then we have the following alternative formulation of (2.9):

$$Y_i^*(t) = \sum_{k=1}^{K} b_k \psi_{ik}(t) + \int_{s_0(t)}^{t} X_i^*(s) \varepsilon_a(s,t) ds + \varepsilon_i(t) = \sum_{k=1}^{K} b_k \psi_{ik}(t) + \varepsilon_i'(t), \tag{2.10}$$

where $\varepsilon_a(s,t) = \beta(s,t) - \sum_{k=1}^{K} b_k \phi_k(s,t)$ and $\varepsilon_i'(t)$ is the sum of random and approximation error. For the estimation of $b_1, \cdots, b_K$, one minimizes the L^2 criterion,

$$SSE = \int_0^T \sum_{i=1}^{n} \left\{ \varepsilon_i'(t) \right\}^2 dt. \tag{2.11}$$

Malfait and Ramsay (2005) minimized SSE using multivariate approximation and recovered the regression function $\beta(s,t)$ as $\hat{\beta}(s,t) = \sum_{k=1}^{K} \hat{b}_k \phi_k(s,t)$.

Harezlak et al. (2007), on the other hand, proposed penalized methods for the estimation of $\beta(s,t)$ using the same basis function as Malfait and Ramsay (2003). Let us assume that the predictor processes are observed at a set of discretized points, $t_j, j = 1, \ldots, J$, and $\psi_{ik}(t_j)$'s are available. They minimized

$$\sum_{i=1}^{n}\sum_{j=1}^{J}\left[Y_i(t_j) - \sum_{k=1}^{K}\psi_{ik}(t_j)b_k\right]^2 + \lambda b^{\mathrm{T}}Kb,$$

with respect to b_k's, where $K = w_H K^H + w_V K^V + w_P K^P$ and K^H, K^V, and K^P denotes the penalties on the smoothness of the regression function to horizontal, vertical, and parallel directions. They also put penalties on the absolute values of differences of neighboring coefficients by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{J}\left[Y_i(t_j) - \sum_{k=1}^{K}\psi_{ik}(t_j)b_k\right]^2 + \lambda\|P(b)\|_1,$$

where $P(b) = w_H D^H(b) + w_V D^V(b) + w_P D^P(b)$ and D^H, D^V, and D_P denote functions of absolute values of differences of neighboring coefficients in the horizontal, vertical, and parallel directions. For the former approach, they translated the minimization problem into the maximization of the mixed model log-likelihood, and utilized existing algorithms for the estimation and inference of the estimated regression function. For the latter one, not having an analytical solution, they used a modified LASSO scheme.

2.4 Varying Coefficient Models

Longitudinal data arise in many fields where repeated measurements are taken on each subject. Not measured on a dense grid as functional data, new issues present themselves in modeling of longitudinal data, such as irregular design, missing values, and sparsity.

To model longitudinal data, many parametric and nonparametric techniques were proposed in the literature. Parametric models for longitudinal data such as generalized linear models, nonlinear models and mixed effect models are widely studied for the past two decades. A nice summary about the important results for parametric methods is given in Diggle et al. (1994) among many others. For more references, see Wu and Yu (2002). Parametric models usually provide simple interpretations and feasible computational efficiency in estimation and inference. However, if the model is not correctly specified, the estimation cannot be free from model bias and the conclusions can be misleading.

From the general knowledge about nonparametric methods, it is well known that they provide flexibility at the cost of computational complexity. In the early stage of adoption of nonparametric methods to longitudinal data, nonparametric models were mainly focused on estimation of the mean curve of the response process as a function of time using various nonparametric smoothing techniques. Hart and Wehrly (1986), Altman (1990), Hart (1991), Rice and Silverman (1991) studied the estimation and inference of the conditional mean of $Y(t)$ for given time point t.

Extending the nonparametric models to include covariates other than time, Zeger and Diggle (1994) and Cheng and Wei (2000) proposed the partially linear model. This model combines the marginal linear model with nonparametric regression so that it combines the advantages of both parametric and nonparametric models. This model belongs to a different class of models, in between fully nonparametric and parametric models, called structural nonparametric regression.

An important class of models for longitudinal data, varying coefficient models, was proposed by Hastie and Tibshirani (1993). For n randomly selected subjects, each repeatedly measured over time, the longitudinal sample of $(Y(t), t, X(t))$ is denoted by $\{Y_{ij}, t_{ij}, X_{ij} : i = 1 \cdots, n, j = 1, \cdots, n_i\}$, where t_{ij} is the j^{th} measurement time of the i^{th} subject, Y_{ij} and $X_{ij} = (X_{ij0}, \cdots, X_{ijk})^{\mathrm{T}}$ are observed response and covariate vector of size $K+1$ at time t_{ij}. Here, X_{ijk} denotes the k^{th} component of the covariate vector. The varying coefficient model can be written as

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta(t_{ij}) + \varepsilon_i(t_{ij}), \tag{2.12}$$

where the error term $\varepsilon_i(t_{ij})$ has mean 0 and is assumed to be independent for different subjects.

Two popular methods for estimation in varying coefficient models are local polynomial fitting and basis approximation. Local polynomial fitting relies on the Taylor expansion of the varying coefficients $\beta_l(t), l = 0, \cdots, k$. Suppose that the $(p+1)^{th}$ derivative of $\beta_l(t)$ at the point t_0 exists. A Taylor expansion for t in a neighborhood of t_0 gives

$$\beta_l(t) \approx \beta_l(t_0) + \beta_l'(t_0)(t-t_0) + \frac{\beta_l''(t_0)}{2!}(t-t_0)^2 + \cdots + \frac{\beta_l^{(p)}(t_0)}{p!}(t-t_0)^p,$$

where $f^{(m)}$ denotes the m^{th} derivative of the function f. This polynomial is fitted locally by a weighted least squares. (Fan and Gijbel, 1996)

For longitudinal data, Hoover et al. (1998) proposed the kernel type local polynomial estimator, which minimizes

$$L_p(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i \left[Y_{ij} - \sum_{l=1}^{k}\left\{X_{ij}^{(l)}\left(\sum_{r=1}^{p} b_{lr}(t_{ij}-t)^r\right)\right\}\right]^2 K\left(\frac{t_{ij}-t}{h}\right) \tag{2.13}$$

with respect to b_{lr}, where w_i are non-negative weights and $K(\cdot)$ is a kernel function. The case with $r=0$ and $r=1$ result in the kernel estimator and the local linear estimator respectively. More flexible versions of local least squares estimation have been studied by Fan and Zhang (2000), that allow for different bandwidths for different coefficient functions. Their method consists of two steps where in the first step raw estimators are obtained and in the second step they are refined. They considered the varying coefficient model (2.12) at a fixed time point t as a classical linear model and estimated the raw estimates at time t via the ordinary least squares (OLS) method. By applying OLS repeatedly for all available time points, they obtained a set of raw estimates for k coefficient functions. To make k varying coefficients smooth, smoothing techniques are employed for each coefficient function across the time support separately. They adopted local polynomial regression as a smoothing technique in the second step with the bandwidth selector from Ruppert et al. (1995). Two main advantages of the two-step estimator over other estimation methods for varying coefficient models are its ease in implementation and the availability of various smoothing techniques in the second step.

Another estimation approach, the method of basis expansions, can employ various basis systems, depending on the nature of the data, such as the spline basis, fourier basis, wavelet basis and truncated power basis. The spline basis is a flexible way of approximating a function based on piecewise polynomials, called splines, joined at a sequence of knots. Here, knots are the locations where the derivatives of the splines could have discontinuities. The B-spline basis is one of the most widely used spline bases since it is numerically more stable than the truncated power basis and faster than the fourier basis in function approximation. For more details about the comparison of B-spline basis with others, see Fan and Gijbel (1996).

Let $\{B_{rs}(t) : s = 1, \cdots, K_r\}$ be a set of basis functions and γ_{rs} be a constant. Then the varying coefficient function $\beta_r(t)$ can be approximated by $\displaystyle\sum_{s=1}^{K_r} \gamma_{rs} B_{rs}(t)$ and the approximation of (2.12) can be written as

$$Y_{ij} \approx \sum_{r=0}^{k} \sum_{s=1}^{K_r} X_{ijr} \gamma_{rs} B_{rs}(t) + \varepsilon_i(t_{ij}). \tag{2.14}$$

Note that if the coefficient function, $\beta_r(t)$, is spanned by a set of basis functions, $\{B_{rs}(t) : s = 1, \cdots, K_r\}$, the approximation can be replaced by equality. The least square estimators $\hat{\gamma}_{rs}$ of γ_{rs} can be obtained by minimizing

$$L(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - \sum_{r=0}^{k} \sum_{s=1}^{K_r} (X_{ijr} \gamma_{rs} B_{rs}(t_{ij})) \right]^2 \right\}, \tag{2.15}$$

with respect to $\gamma = (\gamma_0^{\mathrm{T}}, \cdots, \gamma_k^{\mathrm{T}})^{\mathrm{T}}$, where $\gamma_r^{\mathrm{T}} = (\gamma_{r1}, \cdots, \gamma_{rK_r})^{\mathrm{T}}$. When the solution of (2.15) exists, the $\beta_r(t)$ can be recovered by $\hat{\beta}_r(t) = \sum_{s=1}^{K_r} \hat{\gamma}_{rs} B_{rs}(t)$. In addition to the advantage of computational sparsity, the basis approximation method has two more major advantages. One is that it can be applied to various classes of models including parametric, semiparametric and nonparametric models. The other is that it can incorporate random effects. For more details about these properties, see James, Hastie and Sugar (2000) and Rice and Wu (2001).

A generalization of the varying coefficient models is proposed by Sentürk and Müller (2008) for longitudinal data. The proposed model generalizes the ordinary time varying coefficient model in two ways. First, in this model, not only the current value of covariates but also recent past values of predictor affect the current value of the response variable. Second, it allows for measurement error on both the response and the predictor variables. The demand for a new estimation procedure arises mainly due to the second part.

Under the general longitudinal set up given in equation (2.12), let $\{t_j, j = 1, \cdots, T\}$ be the distinct time points among $\{t_{ij}, j = 1, \cdots, T_i, i = 1, \cdots, n\}$. Define $X_i(t_j)$ and $Y_i(t_j)$ to be the underlying, unobserved predictor and response processes, respectively, and let $X_i'(t_j) = X_i(t_j) + \varepsilon_{xi}(t_j)$ and $Y_i'(t_j) = Y_i(t_j) + \varepsilon_{yi}(t_j)$ be their observed, distorted versions. Hence, $\varepsilon_{xi}(t_j)$ and $\varepsilon_{yi}(t_j)$ are independently and identically distributed mean zero additive measurement errors with variances σ_x^2 and σ_y^2, respectively. The underlying model can then be written as

$$Y_i(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j) X_i(t_{j-q-(r-1)}) + \varepsilon_i(t_j), \tag{2.16}$$

where p denotes the number of time points where the covariate has impact on the response, i.e., the window width into the past of covariates, and $\varepsilon_i(t_j)$ is the realization of a zero-mean stochastic process with covariance function $\delta(t', t) = cov\{\varepsilon(t'), \varepsilon(t)\}$. In this model, a time lag q is also included to allow for a lag in the effect of the covariates. The varying coefficient model is a special case of model (2.16) when $p = 1$ and $q = 1$. This model can be thought of as a historical linear model for longitudinal data since only past values of the covariates affect the current value of response process.

For fixed t_j, let us define $X'_{qpj} = (X'_{1,q,p,j}, \cdots, X'_{n_j,q,p,j})^{\mathrm{T}}$ and $Y'_j = (y'_{1j}, \cdots, y'_{n_j j})^{\mathrm{T}}$, where $X'_{i,q,p,j} = (1, X'_i(t_{j-q}), \cdots, X'_i(t_{j-q-p+1}))^{\mathrm{T}}$ and n_j denotes the number of subject available at the time point t_j. Also define $\epsilon(t_j) = (\varepsilon_1(t_j), \cdots, \varepsilon_{n_j}(t_j))'$ and I_{n_j} as the identity matrix of dimension $n_j \times n_j$. Then the model (2.16) at time t_j can be written in linear form as

$$Y_j = X_{qpj}\beta(t_j) + \epsilon(t_j), \tag{2.17}$$

where $\epsilon(t_j)$ satisfies $E(\epsilon(t_j)) = 0$ and $\mathrm{cov}\{\epsilon(t_j)\} = \delta(t, t)I_{n_j}$. Since the usual least squares estimator of $\beta(t_j)$ at time t_j does not target $\beta(t_j)$ due to measurement error in the predictors, Sentürk and Müller used the predictor measurements apart from t_{j-p} as instrumental variables. Their estimator for $\beta(t_j)$ is given as

$$b_{qp}(t_j) = (b_{0j}, b_{1j}, \cdots, b_{pj})^{\mathrm{T}} = \left({X'}^{\mathrm{T}}_{qpj-p} M_{j-p,j} {X'}_{qpj}\right)^{-1} {X'}^{\mathrm{T}}_{qpj-p} Y'_j,$$

where $M_{j-p,j}$ denotes a $n_{j-p} \times n_j$ matrix for which the (a, b)th entry equals 1 if the ath entry of Y'_{j-p} and the bth entry of Y'_j come from the same subject and equals 0 otherwise.

Chapter 3

The Recent History Functional Linear Model for Functional Data

3.1 Introduction

For the estimation of the functional linear models given in (1.1), a tensor product basis function is used for the two-dimensional basis expansion of the regression function due to the rectangular support property in the literature. For the historical functional linear models given in (1.2), Malfait & Ramsay (2003) proposed an estimation method that is based on a two-dimensional basis expansion over the triangular support via the finite element method.

In this chapter, we propose an estimation procedure for densely measured functional data that is based on a one-dimensional basis expansion instead of the two-dimensional ones proposed for the functional linear models and the historical functional linear models. While this simplification results in significant gain in computational costs, it is also easy to implement. To illustrate the cases where model (1.3) would be reasonable, we analyze the relationship between the measurements of rainfall and the flow level of Curlew creek in Florida. A plausible way to model the fluctuations of the flow level would be to assume that the measurements of rainfall from the recent past are affecting the flow level of the river at a given time point. Note that, for this data set, it would not make sense to think that rainfall from future affects the current flow level, nor does one need to consider the whole past of rainfall. It is also not plausible to assume that the current rainfall affects the river flow immediately and hence the proposed model provides a nice generalization and specification of the previously proposed models.

This chapter is organized as follows. In Section 2, we introduce the proposed estimation procedure for the recent history functional linear models and in Section 3, the proposed model parameter selection method is discussed. In Section 4, we compare, via simulations, the performance of the proposed estimation method with that of Malfait & Ramsay (2003) proposed for the historical functional linear models. Application to the flow level data is given in Section 5. We conclude with a remarks section.

3.2 Estimation in Recent History Functional Linear Models

For a sample of n subjects, the response process is modeled via

$$Y_i(t) = \alpha(t) + \int_{t-\delta_1}^{t-\delta_2} X_i(s)\beta(s,t)ds + \varepsilon_i(t), \quad t \in [\delta_1, T],\ 0 \leq \delta_2 \leq \delta_1. \tag{3.1}$$

The proposed estimation procedure makes use of the observation that the regression function, $\beta(s,t)$, is a univariate function in s for a fixed time point t. Fixing the time point t, we can expand the regression function with respect to s using the basis functions, $\phi_k(\cdot),\ k = 1, \ldots, K$, resulting in $\beta(s,t) \approx \sum_{k=1}^{K} b_k\phi_k(s)$, where $s \in [t-\delta_1, t-\delta_2]$. Here we assume a fixed known K for each time point t, which yields reasonable approximations at all the time points. In sections that follow, we will discuss criterions to choose K from the data in practice. Repeating this univariate expansion for different t leads us to the following two dimensional expansion,

$$\beta(s,t) \approx \sum_{k=1}^{K} b_k(t)\phi_{k,t}(s), \quad s \in [t-\delta_1, t-\delta_2], \tag{3.2}$$

where $b_k(t)$ denotes the coefficient of the k^{th} basis function at time t. Using the expansion in (3.2), we can rewrite the regression model in (3.1) as

$$Y_i(t) = \alpha(t) + \sum_{k=1}^{K} b_k(t) \int_{t-\delta_1}^{t-\delta_2} X_i(s)\phi_{k,t}(s)ds + \varepsilon_i'(t), \tag{3.3}$$

where $\varepsilon_i'(t)$ is the sum of the intrinsic and the approximation errors. Defining $\psi_{i,k}(t)$ as $\psi_{i,k}(t) = \displaystyle\int_{t-\delta_1}^{t-\delta_2} X_i(s)\phi_{k,t}(s)ds$, the proposed model in (3.1) simplifies to

$$Y_i(t) = \alpha(t) + \sum_{k=1}^{K} \psi_{i,k}(t)b_k(t) + \varepsilon_i'(t), \quad t \in [\delta_1, T]. \tag{3.4}$$

The model in (3.4) is a varying coefficient model where $\psi_{i,k}(t)$ are the time dependent predictors and $\alpha(t)$, $b_k(t), k = 1, \ldots, K$ are the varying coefficient functions to be estimated.

Outline of the proposed estimation algorithm is as follows. We start with identifying an appropriate choice of basis system $\phi(\cdot)$ for the expansion of the bivariate regression function. After a basis system is chosen, we estimate $\psi_{i,k}(t)$ based on numerical integration methods. Using the estimated $\psi_{i,k}(t)$ as the time-varying predictors, the varying coefficient functions in (3.4) are estimated. Finally estimates of the bivariate regression function $\hat{\beta}(s,t)$ are obtained using the expansions in (3.2).

Common choices of basis systems are the spline basis, Fourier basis, wavelet basis and truncated power basis. The B-spline basis is one of the most widely used spline bases since it is numerically more stable than the truncated power basis and faster than the Fourier basis in approximating processes. For more details about the comparison of the B-spline basis with others, see Fan and Gijbel (1996). The performance of the B-spline basis largely depends on the location and the number of knots. The knot sequence must be determined from the data since the location and the number of knots are closely related with the complexity of the process and the sparsity of data. Nevertheless, for the densely recorded functional data considered here, we will use an equally spaced knot sequence assuming that the smoothness of the whole process is uniform. The number of knots, however, still matters and this will be determined from the data. For a fixed time t, if there are k equally spaced interior knots over the range $[t-\delta_1, t-\delta_2]$, then there are $K=(k+4)$ B-spline basis functions provided that we use the cubic B-spline basis.

For numerical integration, we can apply a first-order approximation in the case of densely measured functional data. Suppose, without loss of generality, that the predictor process and the response process are measured on T equally spaced time points inside the time interval $[0,1]$ so that we have measurements on $t_j = \frac{j}{T}$, where $j=1,\ldots,T$. Then $\psi_{i,k}(t)$ is approximated by

$$\hat{\psi}_{i,k}(t) = \int_{t-\delta_1}^{t-\delta_2} X_i(s)\phi_{i,t}(s)ds \approx \frac{1}{\lfloor(\delta_1-\delta_2)T\rfloor} \sum_{t_j\in[t-\delta_1,t-\delta_2]} X_i\left(\frac{j}{T}\right)\phi_{k,t}\left(\frac{j}{T}\right), \tag{3.5}$$

where $\lfloor a \rfloor$ denotes the ground function that equals the largest integer smaller than a. The approximation in (3.5) is a variation of the usual rectangular rule using the function values evaluated on the knots instead of those evaluated between the knots.

After estimating $\psi_{i,k}(t)$, we next target the coefficient functions $\alpha(t)$ and $b_k(t)$ in (3.4). For the estimation of varying coefficient models in (3.4), we adopt the two step estimation of Fan and Zhang (2000). The estimation procedure proposed by Fan and Zhang (2000) rests on the observation that in varying coefficient models a different linear regression model holds between the response and the predictor at each fixed time point t. Based on this observation, they target the raw varying coefficient function estimates in the first step via ordinary least squares. The raw estimates are refined in the second step by one-dimensional smoothing procedures separately for each varying coefficient function. The method is shown to be fast and quite flexible allowing different degrees of smoothness for different varying coefficient functions in the refinement step. In this article, we use local linear fits for the refinement step. More specifically, let $\hat{\boldsymbol{\Psi}}(t)$ be the $n \times K$ matrix, which has $\hat{\psi}_{i,k}(t)$ as $(i,k)^{th}$ element and $\tilde{\boldsymbol{\Psi}}(t) = [\mathbf{1}, \hat{\boldsymbol{\Psi}}(t)]$. If we define $\boldsymbol{y}(t)$ as an n dimensional vector, which contains the n realizations of the response variable at time t, then, in the first step, we can get the raw estimates, $\tilde{\alpha}(t)$ and $\tilde{\boldsymbol{b}}(t) = [\tilde{b}_1(t), \ldots, \tilde{b}_K(t)]^{\mathrm{T}}$, as

$$\begin{pmatrix} \tilde{\alpha}(t) \\ \tilde{\boldsymbol{b}}(t) \end{pmatrix} = \left[\tilde{\boldsymbol{\Psi}}(t)^{\mathrm{T}}\tilde{\boldsymbol{\Psi}}(t)\right]^{-1} \tilde{\boldsymbol{\Psi}}(t)^{\mathrm{T}}\boldsymbol{y}(t), \quad t \in [\delta_1, 1] \tag{3.6}$$

Let J be the collection of indices j that satisfy $t_j \in [\delta_1, 1]$ and τ be the number of time points in J. A linear smoother for $b_k(t)$ at time point t is given by

$$\hat{b}_k(t) = \sum_{j=1}^{\tau} w_k(t_j, t)\tilde{b}_k(t_j),$$

where the weights $w_k(t_j, t)$ can be constructed by various smoothing techniques. For the local linear fit, let us suppose that the varying coefficient function $b_k(t)$ is two times continuously differentiable. Then the weights are constructed by

$$w_k(t_j, t) = \frac{K_h(t_j - t)\left\{S_{n,2} - (t_j - t)S_{n,1}\right\}}{S_{n,0}S_{n,2} - S_{n,1}^2}, \tag{3.7}$$

where $S_{n,j} = \sum_{i=1}^{n} K_h(t_i - t)(t_i - t)^j$ and $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$.

To recover the regression function $\beta(s,t)$, we plug the respective estimates obtained above into the expansion in (3.2).

3.3 The Choice of Model Parameters: The Number of Basis Functions and Lags

In the proposed model, we need to select two time lags δ_1 and δ_2 and the number of basis functions used, K. The choice of δ_1 and δ_2 determines the width of the window back in time where the predictor process has an impact on the response variable and the choice of K controls the precision of the approximation of $\beta(s,t)$. Following Malfait & Ramsay (2003), we employ a two step selection algorithm in choosing the two sets of parameters determining window size and number of basis functions. We first choose the number of basis functions for the maximum window size of interest, determining the resolution needed for a reasonable fit. Determining the number of basis functions is equivalent to determining the number of interior knots, denoted as k_0. Once the number of knots k_0 for the largest window is determined, the number of knots used to approximate the regression function for shorter windows is taken to be proportional to the window's relative length compared to the largest window size considered. For example, if 5 knots for the largest window size, say 10, is considered to be reasonable for the approximation of the regression function, then we take 3 knots for a shorter window of size 6.

In the proposed method, for the selection of k_0 at the largest window size, we adopt root squared prediction error (RSPE) from Müller & Zhang (2005). They use RSPE to decide the number of basis functions for the approximation of the two dimensional regression function. A modification of the RSPE criterion applicable to the proposed model at a given time point t can be written as

$$RSPE(t) = \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \widehat{Y}_i^{(-i)}(t) - Y_i(t) \right\}^2 \right]^{1/2},$$

where $\widehat{Y}_i^{(-i)}(t)$ denotes the estimated response value for the i^{th} subject measured at time t from the data excluding the i^{th} subject.

The size of the window in the proposed estimation procedure utilizing the connection to varying coefficient models determines the total number of predictors in the varying coefficient models in (3.4). Nevertheless, the choice of the window size is not a typical variable selection problem in that, for different δ_1 and δ_2 combinations, we have different basis functions, and hence, different predictors. Therefore, the choice of δ_1 and δ_2 is a model selection problem in varying coefficient models rather than a variable selection problem and we adopt a similar selection criterion to Fan, Yao & Cai (2003). For a model selection criterion of varying coefficient models at a given time point t, they use a pointwise AIC criterion utilizing the fact that the varying coefficient model reduces to a linear model at a fixed time point. In the adaptation of their idea to our model, we employ the pointwise AIC at time t, $AIC(t)$, defined by

$$AIC(t) = n\left(\log(2\pi) + \log(RSS(t)) + 1\right) + 2K,$$

where $RSS(t) = \sum_{i=1}^{n}(\tilde{Y}_i(t) - Y_i(t))^2$, $\tilde{Y}_i(t) = \tilde{\alpha}(t) + \sum_{k=1}^{K}\phi_k(t)\tilde{b}_k(t)$ and $\tilde{\alpha}(t)$, $\tilde{b}_k(t)$ are the raw estimates of the intercept and the k^{th} varying coefficient function respectively. Here K represents the number of basis functions used, and hence, the number of predictors. In practice, an initial set of δ_1 and δ_2 combinations are considered and the combination that results in the minimum $AIC(t)$ for most of the time points considered is picked.

3.4 Simulation Study

Goals of the simulation are to study the finite sample properies of the proposed estimator in comparison with that of Malfait & Ramsay (2003)'s estimation method (referred to as MR) and to study the performance of the proposed parameter selection criterion. One of the main advantages of the proposed algorithm is that it provides data-driven automatic selection of the window size and the optimal number of knots. On the other hand, the MR method does not provide automatic choice of these parameters. Instead, the parameters of the historical functional linear models are determined by monitoring the changes in R^2 as usually done in the linear models. Note also that the MR method does not allow for one of the lags, δ_2. Due to these restrictions, the two goals outlined at the beginning of this section need to be addressed by two separate simulation designs. For the model assessment, we define NIE (normalized integrated error) utilizing one subject leave out cross validation criterion as

$$NIE = \frac{\int \sqrt{\sum_{i=1}^{n} \{\hat{Y}_i^{(-i)}(t) - Y_i(t)\}^2 dt}}{\int \sqrt{\sum_{i=1}^{n} \{Y_i(t)\}^2 dt}}. \tag{3.8}$$

For the first part of the simulation, we compare the proposed method with the MR method assuming that the true window size and the model parameters are known. The true window size is fixed as $\delta_1 = 0.1$ and $\delta_2 = 0$. Note that in the MR method, the resolution is controlled by the number of subintervals (denoted N) that the two dimensional regression function support is divided into, which in turn specifies the location and number of nodes of the two dimensional basis expansion. We consider the performance of the MR method at different resolutions and compare NIEs at 10, 20 and 40 subintervals to those obtained from the proposed method with one interior knot for three different number of subjects, $n = 30, 100$ and 300. The computational intensiveness of the two algorithms is also compared via the computing time in seconds under the same computing environment. The second part of the simulation studies the performance of the model selection criterion proposed for the choice of δ_1 and δ_2 under the true window size of $\delta_1 = 0.1$ and $\delta_2 = 0.05$.

3.4.1 Data generation

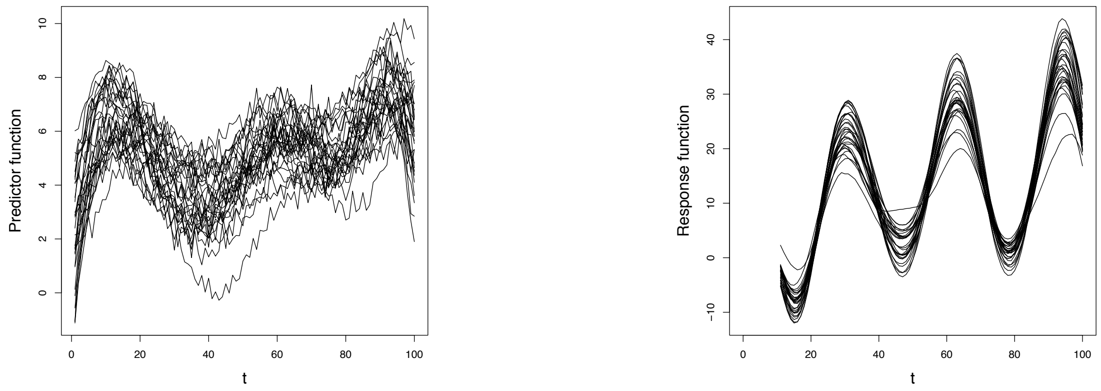
Data is generated from the model

$$Y_i(t) = \alpha(t) + \int_{t-\delta_1}^{t-\delta_2} \beta(s,t)X_i(s)ds + \varepsilon_i(t), \quad t \in [\delta_1, 1] \tag{3.9}$$

on 100 equally spaced grid points in [0,1]. The intercept function is taken to be $\alpha(t) = 10\sin(2t)$ for $\delta_1 \leq t \leq 1$ while the true bivariate regression function, $\beta(s,t)$, is defined to be $\beta(s,t) = 1.5\sin(0.2s - 10) + 0.3\cos(0.2t) - 1$, where $s \in [t - \delta_1, t - \delta_2]$ and $t \in [0,1]$. For the generation of the predictor trajectory, we use 8 cubic B-spline basis functions with 4 equally spaced interior knots in [0,1] weighed according to 8 randomly generated coefficients from the multivariate normal distribution with mean $\mu = [3,8,6,2,7,4,9,5]^{\mathrm{T}}$ and variance $\sigma^2_{coef}\boldsymbol{I}$, $\sigma^2_{coef} = 3$. The variance of the coefficients is chosen so that the predictor trajectories do not lose the overall pattern but still have a certain amount of variability. A small amount of independent error generated from normal distribution with mean 0 and variance 0.1 is added to the predictor trajectories to prevent singularities. A typical set of the generated predictor trajectories when $n = 30$ is given in Figure 3.1a. The error processes, $\varepsilon_i(t), i = 1, \ldots, n$, are generated by the basis expansion, as in the predictor process case, with $\mu = [0,0,0,0,0,0,0,0]^{\mathrm{T}}$ and $\Sigma_\varepsilon = 5 \cdot \boldsymbol{I}$. The response trajectories are displayed in Figure 3.1b.

3.4.2 Simulation results

Due to the computational intensiveness, we consider a modified NIE using 10-fold cross validation instead of one-subject-leave-out. For the first comparison simulation the average NIEs from 500 Monte Carlo runs are reported in Table 4.1. Note that, in Table 4.1, N denotes the number of subintervals for the MR method determining precision and the standard deviation of the NIEs are given in paranthesis. This table suggests that the proposed estimation method outperforms



(a) Predictor Trajectories

(b) Response Trajectories

Figure 3.1: A typical set of predictor and response trajectories when $n = 30$.

Table 3.1: NIE for the proposed and the MR method.

		$n = 30$	$n = 100$	$n = 300$
Proposed method		0.1192 (0.0069)	0.1154(0.0038)	0.114(0.002)
MR method	$N = 10$	0.1464 (0.0111)	0.1384(0.0058)	0.1357(0.0031)
	$N = 20$	0.1526(0.0121)	0.142(0.0062)	0.1388(0.0035)
	$N = 40$	0.1545(0.0142)	0.1457(0.0062)	0.142(0.0034)

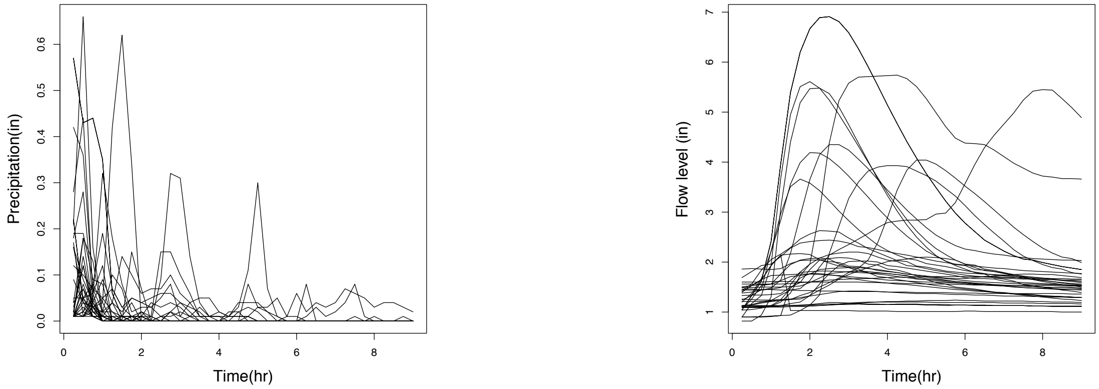
the MR method in terms of NIE. We also compare the average computing times, in seconds, for the two methods under the same computing environment. Note that computing time is defined as the total time to obtain NIEs, which includes multiple model fittings. As the number of subjects or the number of subintervals increase, the average computing time increases. In the simulation with 30 subjects, the MR method with 10 subintervals takes about six times more than the proposed method. For 300 subjects, the MR method with 40 subintervals takes about fifty times more time than the proposed method. This result is expected since while the MR method is based on two dimensional basis expansions, the proposed method uses only one dimensional expansions and smoothing procedures.

The performance of the proposed parameter selection is studied through the second simulation. Under the same data generation scheme, we use $\sigma_\varepsilon^2 = 5$ for the error variance and $\delta_1 = 0.1$ and $\delta_2 = 0.05$ for window size. Six different combinations, $(20, 0), (20, 5), (20, 10), (10, 0), (10, 5)$ and $(5, 0)$, of (δ_1, δ_2) are considered as candidates and one of the six combinations yielding the smallest AIC for most of the time points is chosen. As the number of subjects increases, the proportion of the correct choice also increases. Starting from 0.727 for $n = 30$, the proportion of the correct selection estimated via 500 monte-carlo runs increase to 0.837, 0.988 for $n = 100$ and $n = 300$, respectively.

3.5 Data Analysis

We apply the proposed model to investigate the relationship between the flow level of Curlew creek in Florida and precipitation. From June 7th to October 21st of 2008, both precipitation and flow level are recorded every 15 minutes in inches. We select raining periods for our analysis where the rain lasts for more than 45 minutes and we consider the 9-hour-long data segment from the start of the rain for both the precipitation and flow level. Between the above given dates, there are 33 periods that satisfy this description. Since the flow level goes back to normal quickly after the raining period and the precipitation in a raining period does not seem to affect the flow level of the following raining period, each raining period can be thought of as an independent subject. The collected data are presented in Figure 5.3.

Note that, for the expansion of the regression function, a linear B-spine basis is used instead of a cubic B-spline basis since it yields a better fit in terms of NIE. This implies that the complexity of the regression function can be captured well through the linear approximation. The maximum window size that we consider in this study is $(\delta_1, \delta_2) = (20, 0)$, that is, the flow level of the river at a given time point is assumed to be affected by the patterns of precipitation from the past five hours. The optimal number of knots for the maximum window size is chosen to be



(a) Precipitation data from 33 raining periods.

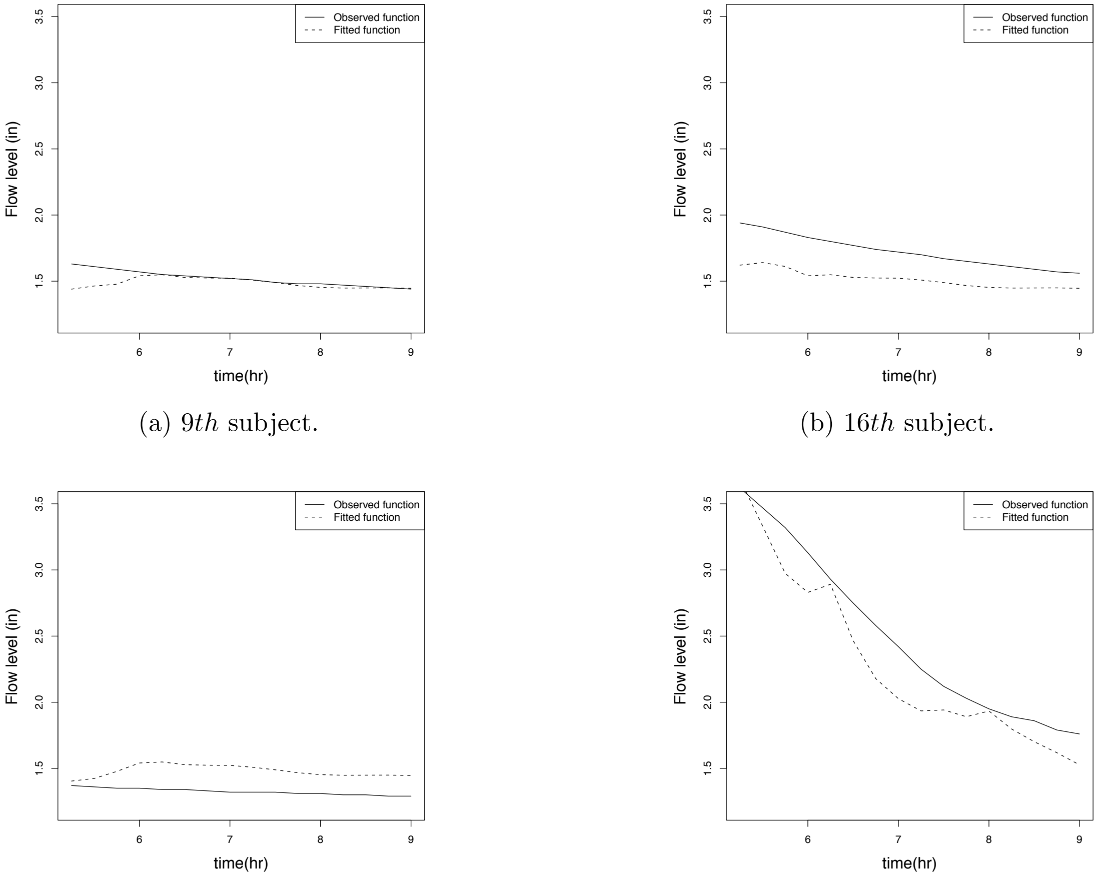
(b) Flow level of Curlew creek, Florida from 33 raining periods.

Figure 3.2: Precipitation and Flow level of Curlew creek, Florida.

1 and the corresponding average RSPE is 0.6642. Therefore, for smaller window sizes considered, we also apply one interior knot for the basis approximation. We tried 8 different sets of δ_1 and δ_2 combinations: $(20, 15)$, $(20, 10)$, $(20, 5)$, $(20, 0)$, $(15, 10)$, $(15, 5)$, $(15, 0)$, $(10, 5)$, $(10, 0)$ and $(5, 0)$. The optimal combination of δ_1 and δ_2 obtained by the proposed criterion is $(20, 5)$, implying that precipitation from the past 5 to 1.25 hours has an effect on the flow level. For the selected combination, the resulting NIE is 0.258 which implies that about 74 % of total variation in the response function is explained by the fitted model. Figure 4.3 displays flow levels for four randomly selected subjects along with their predicted trajectories.

The estimated regression surface is given in Figure 3.4. Displayed in Figure 3.4b are 15 profiles of the bivariate regression surface $\hat{\beta}(t-l, t)$, for $l \in [1.25, 4.75]$ and $t \in [5, 9]$ hours.

Profiles are slices of the regression function in Figure 3.4a parallel to the diagonal and getting further from it by l hours. These profiles represent the influence of the predictor process on the response process at l amount back in time from the current time point t. In this plot, the dominating profile corresponds



(a) $9th$ subject.

(b) $16th$ subject.

(c) $22nd$ subject.

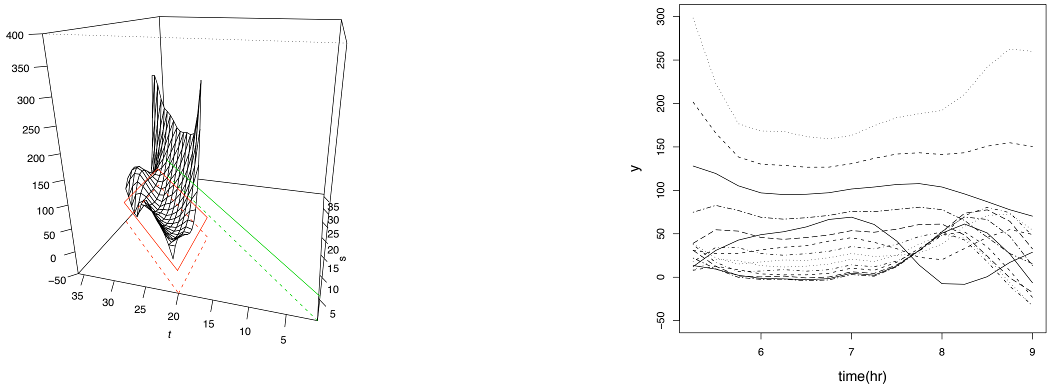
(d) $32nd$ subject.

Figure 3.3: The prediction results for 4 selected subjects. The solid line represents the observed response trajectory and the dashed line denotes the predicted response trajectory.

to $l = 1.25(hr)$ and the magnitude of the profile decreases as the size of the lag increases. This implies that the current value of the response process is affected most by the predictor process at the closest time point.

3.6 Concluding Remarks

In this chapter, we propose a variant of the historical functional linear model, which models the historical influence of the predictor process on the response pro-



(a) The estimated regression function.

(b) The sliced regression function in the direction of the diagonal.

Figure 3.4: Estimated regression function and profiles parallel to the line $s = t$.

cess flexibly via the use of two lag parameters. For the estimation of this new class of models, we proposed a new estimation method based on a one-dimensional expansion of the regression function utilizing the connection between functional linear and varying coefficient models. Note that the proposed estimation method assumes densely measured functional data. This assumption plays a key role especially in approximating integrals over the repeated measurements of each subject such as estimating the $\psi_{i,k}(t)$'s. In future research we will consider alternative estimation algorithms that can also accommodate sparse longitudinal data with irregular measurement times and few number of observations per subject.

Chapter 4

The Recent History Functional Linear Model for Longitudinal Data

4.1 Introduction

In this chapter, we propose an estimation method for the recent history functional linear model given as

$$Y_i(t) = \alpha(t) + \int_{t-\delta_1}^{t-\delta_2} \beta(s,t)X_i(s)ds + \varepsilon_i(t),\ s \in [t-\delta_1, t-\delta_2],\ t \in [0,T],\ i = 1,\ldots,n, \tag{4.1}$$

for the case where a moderate number of observations per subject is taken at irregular time points. In Chapter 3, we proposed an estimation method for the recent history functional linear model for functional data using one-dimensional basis expansions and local polynomial smoothing techniques highlighting the connection of the functional linear model to the varying coefficient models. For the varying coefficient model, we employed the two-step estimator proposed by Fan and Zhang (2003), which utilizes the property of the varying coefficient model being the classical linear model on a fixed time point t.

This approach, however, is not directly applicable for longitudinal data with measurements taken at irregular time points across the subjects, since the estimation of the linear model at a given time point may fail due to lack of data. As a remedy, one may consider binning to include enough number of subjects and fit linear regression model assuming that the linear relationship between the predictor and response variables does not change within a fixed bin. It is, however, not straightforward how to choose bin size and it may not be reasonable to assume that the linear relationship does not change within the bin if the bin size is large. At the same time, the irregularity in measurement time and insufficiency of measurements make it difficult to approximate the integrations involved in the model. Therefore, for more flexible treatment of irregular measurements, we consider the local polynomial fit proposed by Hoover et al. (1998) instead of the two-step estimator of Fan and Zhang (2003).

This chapter is organized in the following manner. In Section 4.2, we introduce the proposed model and outline its estimation method. Parameter selection methods are discussed in Section 4.3. In Section 4.4, we investigate the finite sample properties of the estimator via simulation studies.

4.2 Estimation

Let $(X_i, Y_i),\ i = 1, \dots, n$, be the realizations of square integrable random processes (X, Y) with mean μ_X and μ_Y, and variance $G_X(s,t)$ and $G_Y(s,t)$, respectively. Here, s and t denote time indices defined on the finite and closed interval, $[0, T]$. For longitudinal data, what we observe for the i^{th} realization of the predictor and response processes are not the entire processes but a small number, say n_i, of measurements taken irregularly at time points $T_{i1}, \dots, T_{in_i}$, and we have data of the following form

$$(X_{ij}, Y_{ij}, T_{ij}),\ i = 1, \dots, n, \quad j = 1, \dots, n_i. \tag{4.2}$$

In what follows, let us denote by Δ_t the time interval $[t-\delta_1, t-\delta_2]$. By projecting the true regression function, $\beta(\cdot, t)$, for a given time point t, on the space spanned by K basis functions, $\phi_k(\cdot),\ k = 1, \ldots, K$, the true regression function can be approximated by $\beta(s,t) \approx \sum_{k=1}^{K} \phi_k(s) b_k(t)$, where $b_k(t)$ is the coefficient of the k^{th} basis function. Then, the model given in (4.1) can be written as

$$Y_i(t) = \alpha(t) + \sum_{k=1}^{K} b_k(t) \int_{\Delta_t} X_i(s)\phi_k(s-t+\delta_1)ds + \varepsilon'_i(t) = \alpha(t) + \sum_{k=1}^{K} b_k(t)\psi_{ik}(t) + \varepsilon'_i(t) \tag{4.3}$$

where $\psi_{ik}(t) = \int_{\Delta_t} X_i(s)\phi_k(s-t+\delta_1)ds$ and $\varepsilon'_i(t)$ is the sum of the intrinsic error process for the i^{th} subject and the approximation error from the basis expansion of the regression function at time point t. In this expression, K controls the model complexity and should be estimated from data.

To estimate the model given in (4.3), one has to estimate $\psi_{ik}(t)$ first. For longitudinal data, it is not guaranteed to have good approximations of $\psi_{ik}(t)$'s for all of the time points, $T_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i$, since measurements in the sliding window, $\Delta_{T_{ij}}$, may not be sufficient for numerical integration. As a consequence, we only include the time points where the approximation of $\psi_{ik}(t)$'s are available in the estimation procedure. After the approximation of $\psi_{ik}(t)$'s, let us note that the model in (4.3) reduces to the varying coefficient model with $\psi_{ik}(t)$'s as the time varying covariates forming induced data triples, $(\psi_{ij}, Y_{ij}, T_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, n'_i$, where $\psi_{ij} = [\psi_{i1}(T_{ij}), \ldots, \psi_{iK}(T_{ij})]^{\mathrm{T}}$ and n'_i is the number of time points where approximated ψ_{ij}'s are available. Note that, n'_i is less than or equal to n_i, since we exclude the time points where the approximation of $\psi_{ik}(T_{ij})$ are not available.

For the expansion of the regression function, a suitable set of basis functions needs to be determined considering the nature of data. Among the common choices, the Fourier basis function is suitable for the expansion of the periodic data and the polynomial basis function is useful for the approximation of the curves that change globally on the entire support. On the other hand, the piecewise polynomials, such as truncated power basis and B-spline basis, are frequently used to capture the local feature of the function keeping the global pattern of the function unchanged. The B-spline basis function is one of the most widely used one due to the numerical stability and the fast computation rather than the truncated power basis and Fourier basis. In this chapter, we apply the B-spline basis function for the approximation of the regression function. For more discussion on the choice of the basis, see Fan and Gijbel (1996) and Ramsay and Silverman (2005). For those who are interested in B-spline, see Schumaker (1987).

The numerical integration for the product of the predictor and the basis functions needs to be evaluated to estimate $\psi_{ik}(T_{ij})$ defined in (4.3). Note that the induced varying coefficients are evaluated only at the time points where there are enough observations within $\Delta_{T_{ij}}$ to perform the numerical integration with a certain level of precision. The number of measurements needed for the numerical approximation can be manually determined depending on the window size of interest and the properties of the predictor process.

Based on the Taylor expansion, we can approximate the intercept and varying coefficient functions, $\alpha(t_0)$ and $\beta_k(t_0)$, $k = 1, \ldots, K$, using polynomials of order p via $\alpha(t_0) \approx \sum_{l=0}^{p} \alpha_l^t (t - t_0)^l$ and $\beta_k(t_0) \approx \sum_{l=0}^{p} \beta_{kl}^t (t - t_0)^l$ at the vicinity of t_0. Based on this approximation, the varying coefficient model given in (4.3) can be rewritten as

$$Y_i(t) \approx \sum_{l=0}^{p} \alpha_p^t (t - t_0)^l + \sum_{k=1}^{K} \int_{\Delta_t} X_i(s)\phi_k(s - t + \delta_1)ds \left\{ \sum_{l=0}^{p} \beta_{kl}^t (t - t_0)^l \right\} + \varepsilon_i'(t).$$

To obtain a local polynomial fit, let $K_h(\cdot)$ be a kernel function with the bandwidth h defined by $K_h(t) = K(t/h)/h$ on a compact support where $K(\cdot)$ is a function that satisfies $\int K(t)dt = 1$ and $\int tK(t)dt = 0$. Also let us define a parameter vector, $\theta_t = [\alpha_0^t, \ldots, \alpha_p^t, b_{10}^t, \ldots, b_{1p}^t, b_{20}^t, \ldots, b_{2p}^t, \ldots, b_{K0}^t, \ldots, b_{Kp}^t]^{\mathrm{T}}$. The local polynomial estimate of order p for θ_t can be obtained by minimizing

$$Q(\theta_t) = \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left[Y_{ij} - \sum_{l=0}^{p} \left\{ \alpha_l^t + \sum_{k=1}^{K} \hat{\psi}_{ik}(t_{ij}) b_{lk}^t \right\} (t_{ij} - t)^l \right]^2 K_h\left(t_{ij} - t\right),$$

where w_i is a weight applied on the i^{th} subject that satisfies $\sum_{i=1}^{n} w_i n_i = 1$. The weight $w_i = 1/(\sum_{i=1}^{n} n_i)$ puts equal weight to all of the measurements, i.e., it gives more weight for the subject with many measurements. One can set $w_i = 1/(nn_i)$ if uniform weight for each subject is desired.

Local linear fit, in the case of $p = 1$, is preferred over local constant fit ($p = 0$) due to its automatic carpentry property for the boundary effect and over higher order fit ($p > 1$) due to its simplicity. The local linear estimator of $[\alpha(t), b_1(t), \ldots, b_k(t)]^{\mathrm{T}}$ can be constructed in the following manner. For a given t, let $B_i(t)$ be a matrix of size $n_i \times 2$, where the l^{th} column of $B_i(t)$ is defined as $[1, (t_i - t)]$ and $t_i = [t_{i1}, \ldots, t_{in_i'}]^{\mathrm{T}}$ and $W_i(t)$ be a diagonal matrix of which the j^{th} diagonal entity is $w_i K_h\left(t_{ij} - t\right)$ for $j = 1, \ldots, n_i'$. Defining $\mathcal{A}_i(t) = [A_i(t_{i1}), \ldots, A_i(t_{in_i'})]^{\mathrm{T}}$, where $A_i(t_j) = [1, (t_j - t), \psi_{i1}(t), \psi_{i1}(t)(t_j - t), \ldots, \psi_{iK}(t), \psi_{iK}(t)(t_j - t)]$, the least squares estimator of θ_t is given by

$$\hat{\theta}_t = \left\{ \sum_{i=1}^{n} \mathcal{A}_i^{\mathrm{T}}(t) W_i(t) \mathcal{A}_i(t) \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathcal{A}_i^{\mathrm{T}}(t) W_i(t) Y_i \right\}$$

where $\hat{\theta}_t = [\hat{\alpha}_0^t, \hat{\alpha}_1^t, b_{01}^t, \hat{b}_{11}^t, \ldots, \hat{b}_{0K}^t, \ldots, \hat{b}_{1K}^t]^{\mathrm{T}}$. The local linear estimator of intercept function $\alpha(t)$ is $\hat{\alpha}_0^t$ and that of the regression function can be obtained by the equation $\hat{\beta}(s,t) = \sum_{k=1}^{K} \phi_k(s) e_{2k+1}^{\mathrm{T}} \hat{\theta}_t$, where e_q is the $2(K+1)$-dimensional vector with 1 as its q^{th} element and zero elsewhere.

4.3 Parameter Selection

We have three sets of parameters used in the estimation procedure: K, h, and (δ_1, δ_2) combinations. The number of basis functions and bandwidths for the estimation of the varying coefficient model given in (4.3) control the resolution of the fit. As the bandwidth h gets smaller and the number of basis functions K used for the expansion of the regression function gets larger, we can get an estimate with better resolution. The (δ_1, δ_2) combination, on the other hand, determines the size of window where predictor processes have effects on the response processes. Observing different roles of different parameters, we apply a sequential method to select the optimal parameters, that is, we define the resolution of the fit for the maximum window considered by deciding h and K first, and select the best combination of lags, (δ_1, δ_2), under the same resolution next.

To select the optimal combination for the number of basis functions, K, and the bandwidth, h, we use a modified RSPE defined in Chapter 3. Once the number of basis functions K is determined for the maximum window size, the number of basis functions for shorter lags can be chosen proportional to the maximum window size to keep the resolution in the same level. For example, if $K = 8$ for the window of size 10 is chosen based on a modified average RSPE, one can apply $K = 4$ for the window of size 5. In the case of cubic B-spline basis functions, the smallest K can be taken is 4 for windows shorter than 5. On the other hand, the selected bandwidth h by the algorithm can be used for all the window sizes of interest as it only controls the local property of the estimator that has little connection to the size of window. In practice, we try different combinations of K and h, and select the combination that minimizes the RSPE. Since K can only assume integer values, this procedure is not computationally expensive and is feasible by trying a moderate number of candidate bandwidths.

For the choice of the (δ_1, δ_2) combination, we took advantage of the fact that the varying coefficient model at a fixed time point is a linear model and hence compare pointwise AICs for candidate combinations of (δ_1, δ_2) in Chapter 3. Due to the irregularity of measurement time points, it is not guaranteed that there are enough observations at a given time point t for pointwise AIC. To deal with this, we divide the support into a number of bins so that there are enough subjects with at least one measurement inside each bin and obtain the bin-wise AIC instead of pointwise AIC. To make sure that the measurements inside a bin are not correlated to each other, for the subjects that have multiple measurements inside the bin, we select one observation that is the closest to the center of the bin. The combination of (δ_1, δ_2) that is chosen to be the best for the most bins is considered as the optimal window size. To fix the idea, let n_I be the number of subjects that have at least one measurement in the I^{th} bin and those subjects are denoted by $i_1, \ldots, i_{n_I}$. The AIC at a given bin, I, is defined as

$$AIC(I) = n_I \left[\log(2\pi) + \log\{RSS(I)\} + 1\right] + 2K,$$

where $RSS(I) = \sum_{j=1}^{n_I} (\tilde{Y}_{i_j} - Y_{i_j})^2$, Y_{i_j} is the j^{th} subject that belongs to the I^{th} bin, $\tilde{Y}_{i_j} = \tilde{\alpha}^I + \hat{\psi}_{i_j}\tilde{\beta}^I$, and $\tilde{\alpha}^I$, $\tilde{\beta}^I$ are the least squares estimates of the intercept and the vector of the coefficient for the linear model in the I^{th} bin, respectively. Here, $\hat{\psi}_{i_j}$ is the K-dimensional vector with $\psi_{i_j k}(t)$ as its k^{th} element.

4.4 Simulation Studies

In this section, we investigate the performance of the estimator in comparison with other estimators that ignore the sliding window property of the recent history functional linear model and the performance of the criterion for the window combination, (δ_1, δ_2). For model assessment, we use the NIE (normalized integrated error) defined in Chapter 3. Throughout the simulation study, we do not consider the intercept function, as it can be easily recovered by the formula

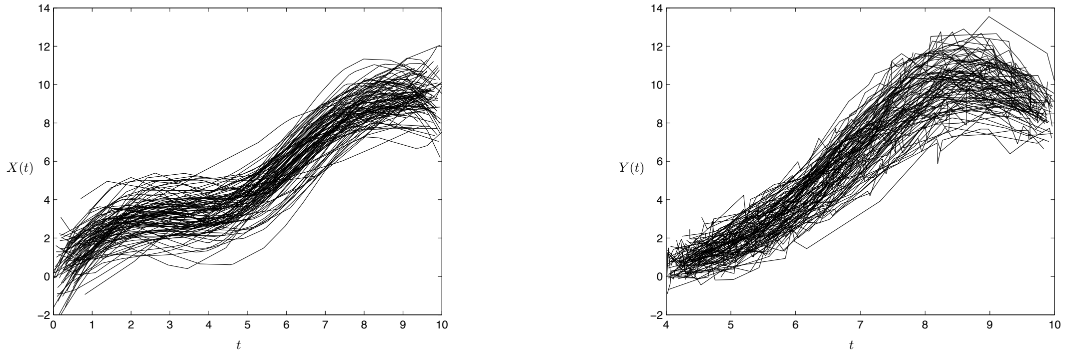
$$\hat{\alpha}(t) = \hat{\mu}_Y(t) - \int_{\Delta_t} \hat{\mu}_X(s)\hat{\beta}(s,t)ds.$$

4.4.1 Data generation

For n subjects, the number of measurements, n_i, of the i^{th} data triple is generated from the discrete uniform distribution, [20,25]. The design points, $T_{i1}, \ldots, T_{in_i}$, are randomly selected from the uniform distribution $[0, 10]$. On design point t, the predictor trajectory is generated around the mean function $t + \sin(t)$ using the B-spline basis of order 4 with knot sequence $[2, 4, 6, 8]$ on the support of $[0, 10]$ as $X(t) = t + \sin(t) + \sum_{q=1}^{8} b_q B_q(t),\ t \in [0, 10]$. Note that there are 8 basis functions involved in the generation of the predictor processes. The eight coefficients, $b_q,\ q = 1, \ldots, 8$, are independently generated from a normal distribution with mean 0 and variance σ^2. The regression function is generated by two basis functions, $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$ and $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, via the equation, $\beta(s,t) = \sum_{i=1}^{2}\sum_{j=1}^{2} c_{ij}\phi_i(s)\phi_j(t),\ s \in [t-\delta_1, t-\delta_2]$, where $c_{11} = 2, c_{12} = 2, c_{13} = 1$ and $c_{14} = 2$. The error process is generated by the basis functions used for the generation of the predictor trajectories with coefficients sampled from the normal distribution with mean 0 and σ_ε^2 independently. The response function is generate by the equation $\displaystyle\int_{\Delta_t} \beta(s,t)X(s)ds + \varepsilon(t)$. In this simulation study, the set of two true lags, (δ_1, δ_2), is set to be $(2, 0)$; the variances for the predictor and error functions are set to be 2 and 0.25, respectively. Typical predictor and response trajectories are depicted in Figure 4.1 for $n = 100$.

4.4.2 Simulation result

In this section, we present estimation results based on simulated data with $n = 100$ to study the performance of the proposed estimation procedure. In the first step of the estimation procedure, we estimate $\psi_{ik}(t)$'s. The measurement time points available in the original data and those available after the estimation are given

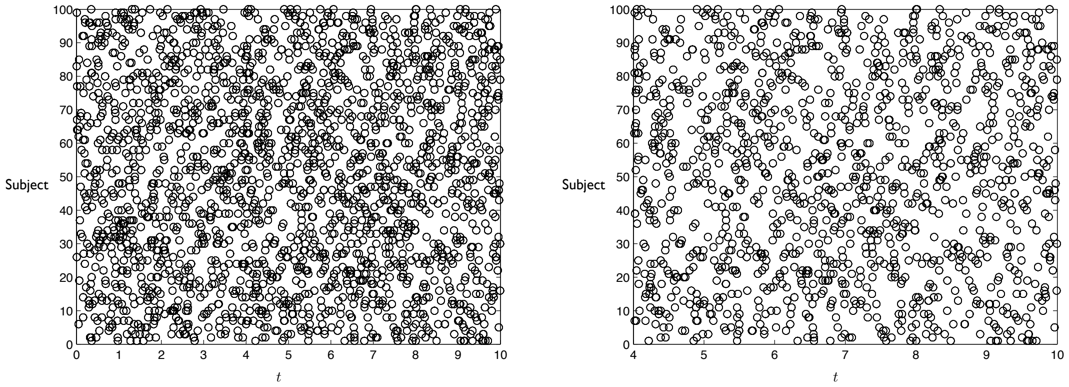


(a) Predictor functions

(b) Response functions

Figure 4.1: Generated predictor and response processes for $n = 100$.

in Figure 4.2. Only time points where there are more than 3 measurements in the sliding window back in time are considered in the estimation procedure to maintain the precision of the approximation. From Figure 4.2, we can see that the approximation procedure reduces the total number of observations in dataset by about 50% yielding around 10 observations per subject after calculating induced covariates. After approximating the induced covariates, one has to decide the number of basis functions and the bandwidth for the varying coefficient model.



(a) Original design points

(b) Induced design points

Figure 4.2: The time points where (X_{ij}, Y_{ij}, T_{ij}) and $(\psi_{ij}, Y_{ij}, T_{ij})$ are available. The Y-axis depicts the subjects, $n = 1, \ldots, 100$, and the X-axis represents the time interval, [0,10].

In this particular simulation study, the bandwidth, $h = 1$, and the number of basis functions, $K = 7$, gives us the smallest RSPE for the maximum window size $(\delta_1, \delta_2) = (4, 0)$. Finally, we use 5 basis functions, i.e., cubic B-spline basis function with 1 interior knot at the center of the window, since the true window size in this simulation study is $(\delta_1, \delta_2) = (2, 0)$. In Figure 4.3, we present the predicted response trajectories in Figure 4.1 based on these parameters.

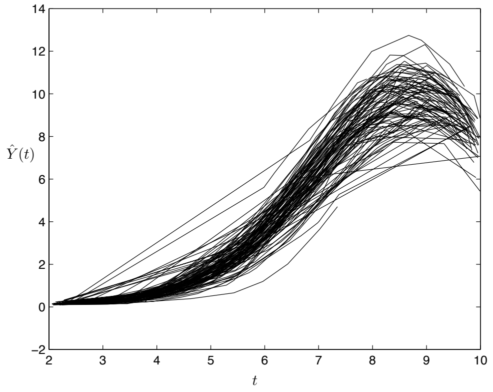


Figure 4.3: Predicted response trajectories for the simulated response trajectories given in Figure 4.1.

We study the importance of correctly specifying the sliding window in the estimation procedure. For this, we apply the functional principal component regression proposed by Yao et al. (2005a), which assumes the whole predictor process to have effects on the response, and the varying coefficient model, which models pointwise relationship between the predictor and response, to the simulated data with sliding window, $(\delta_1, \delta_2) = (2, 0)$. All parameters required for the models are chosen by their own automatic criteria such as CV, GCV, and functional R^2. For summary statistics, we report the average NIEs from 200 Monte Carlo simulations in Table 4.1. The standard error of NIE is also provided in the parenthesis. Since the mean NIE from the recent history functional linear model is significantly lower than the other two in Table 4.1, we can see that it is important to figure out the correct window size for the case where only a fraction of the predictor process has

Table 4.1: NIEs from the recent history functional linear model (RHFLM), the varying coefficient model (VCM), and functional principal component regression model (FPCR) for the simulation data with $(\delta_1, \delta_2) = (2, 0)$

	RHFLM	VCM	FPCR
Mean NIE	0.1088	0.2826	0.2282
(Std. dev.)	(0.0056)	(0.0382)	(0.0436)

effect on the response process.

To study the performance of the window selection algorithm, we try 8 different sets of the window combinations. The true window combination is $(2, 0)$ as before and we report the ratio of the correct choice for the window combination among 600 Monte Carlo simulations in Table 4.2 under different simulation setups. The order of basis functions used for this study is $m = 1$ and different choices for the order of basis system result in the same pattern. Two parameters K and h are automatically chosen by RSPE criterion and bin size is taken as one tenth of the entire support considering the computational costs. The results displayed in Table 4.2 indicate that the window selection criterion performs better as noise level decreases and as the number of subject increases.

Table 4.2: The performance of window size selection criterion. The number that appears in each cell is the proportion of selecting the true window size and that in parenthesis denotes the number of replications.

	$\sigma_\varepsilon = 0.2$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 0.8$
$n = 30$	0.9448(598)	0.8512(598)	0.32(600)
$n = 50$	1(600)	1(600)	0.765(600)

Chapter 5

The Recent History Functional Linear Model for Sparse Longitudinal Data

5.1 Introduction

We address estimation in regression modeling of sparse longitudinal data. Sparse designs where the repeated measurements taken on each subject are irregular and the number of repetitions per subject is small, are encountered commonly in applications. An example is the longitudinal primary biliary liver cirrhosis data collected by the Mayo Clinic (see Appendix D of Fleming and Harrington, 1991). Due to missed visits the data is sparse and highly irregular where each patient visited the clinic at different times. We consider estimation in regression models where the sparse longitudinal predictor process only from the recent past has an effect on the sparse response trajectory.

An intermediate model considered in this dissertation between the functional linear model with global support and the varying coefficient model with point-wise support is the recent history functional linear model

$$E\{Y(t)|X(s),\ s \in \Omega_t\} = \alpha(t) + \sum_{k=1}^{K} b_k(t) \int_{\Omega_t} X(s)\phi_k(s)ds, \tag{5.1}$$

where the two dimensional regression function $\beta(s,t)$ of the functional linear model is approximated by the product form $\sum_{k=1}^{K} b_k(t)\phi_k(s)$ of one dimensional functions. In (5.1), the predictor process from the recent past in a sliding window is assumed to affect the current response value, i.e. $\Omega_t = [t-\delta_1, t-\delta_2]$ for $0 < \delta_2 < \delta_1 < T$. The two delay parameters help define the sliding window support where δ_2 denotes the delay for the predictor process to start affecting the response, and δ_1 is the delay beyond which the predictor process has no effect. Other regression models have also been proposed with a sliding window support such as the generalized varying coefficient model of Şentürk and Müller (2008) and the functional varying coefficient model of Şentürk and Müller (2010), where authors argue that the sliding support is useful in many applications where the response is affected by recent trends in the predictor process. In Chapter 3, we consider the regression of river flow on rainfall as a motivating example where the river flow level would depend on the recent rainfall but not current rainfall or rainfall from distant past.

We proposed an estimation procedure for the recent history functional linear model for densely recorded functional data in Chapter 3. In this chapter, we propose a new estimation procedure for the recent history functional linear model specifically tailored for sparse longitudinal data. Sparsity is a real challenge in modeling longitudinal data, since nonparametric methods cannot feasibly explain a single trajectory for sparse designs. In addition, while standard semiparametric estimation approaches to longitudinal data have been studied extensively for irregular designs, they are not particularly designed to address sparsity issues and can yield inconsistent results for sparse noise contaminated longitudinal data.

Most recently there have been a number of proposals in the literature to broaden the reach of functional data analysis in general and functional linear models in particular, to include sparsely observed longitudinal data. James et al. (2000) proposed an estimation method based on reduced rank mixed-effect model emphasizing the sparse data case among others. Yao et al. (2005a) proposed an estimation procedure for the functional linear model with $\Omega_t = [0, T]$, which is applicable to sparse longitudinal data based on functional principal component analysis. The main idea is that information across all the subjects can still be pooled effectively even for sparse data in the estimation of mean functions and covariance surfaces of the random processes, which are needed for functional principle components analysis.

The challenge of estimation based on sparse longitudinal data is intensified when one considers sliding window supports for the regression function such as the support $[t - \delta_1, t - \delta_2]$ of the recent history functional linear model. That is even if the entire observed process is not very sparse, observations in the sliding window can easily get sparse fast as the window size decreases. We address this challenge by drawing connections between the proposed recent history functional linear model and the varying coefficient model and basing the estimation algorithm mainly on the estimation of auto- and cross-covariances following the proposals of Yao et al. (2005a). More specifically, note that for a set of K predetermined basis functions $\phi_k(s)$, the recent history functional linear model in (5.1) reduces to a multiple varying coefficient model with K induced predictors $\int_{\Omega_t} X(s)\phi_k(s)ds$. The unknown varying coefficient functions $b_k(t)$ are then targeted based on a set of covariance representations which can be estimated efficiently based on sparse longitudinal data, pooling information across subjects.

The chapter is organized as follows. In Section 5.2, we introduce the recent history functional linear model. The proposed estimation method is outlined and uniform consistency of the proposed estimators are established in Section 5.3. The prediction of the response trajectories is also proposed in Section 5.3 utilizing

Gaussian assumptions, where asymptotic distributions are derived leading to pointwise asymptotic confidence bands. In Section 5.4, we discuss numerical issues in implementation along with the choice of model parameters. We study the finite sample properties of the proposed estimators through simulations given in Section 5.5. We apply the proposed method to a primary biliary liver cirrhosis longitudinal data in Section 5.6, to study the dynamic relationship between serum albumin concentration and prothrombin time. Concluding remarks are given in Section 5.7 and technical details are assembled in an Appendix.

5.2 Data and Model

Taking a functional approach we view the observed longitudinal data as noise-contaminated realizations of a random process that produces smooth trajectories. We will reflect sparsity in the following representation through a random number of repeated measurements per each subject at random time points. Let $(X_i, Y_i),\ i = 1, \ldots, n$, be the pairs of square integrable predictor and response trajectories, which are realizations of the smooth random processes (X, Y), defined on a finite and closed interval, $\mathcal{T} = [0, T]$. The smooth random processes have unknown smooth mean functions $\mu_X(t) = EX(t)$ and $\mu_Y(t) = EY(t)$, and auto-covariance functions $G_X(s,t) = \text{cov}\{X(s), X(t)\}$ and $G_Y(s,t) = \text{cov}\{Y(s), Y(t)\}$. Throughout this chapter, s and t refer to time indices defined on $\mathcal{T}$. The observed trajectories of the i^{th} subject, $X_{ij} = X(T_{ij}) + \varepsilon_{ij}, Y_{ij} = Y(T_{ij}) + e_{ij},\ j = 1, \ldots, N_i$ are noise contaminated realizations of the random processes X and Y measured at i.i.d. random time points, T_{ij}. Here, ε_{ij} and e_{ij} denote the additive i.i.d. zero mean finite variance measurement errors of the predictor and the response trajectories, respectively. The numbers of random time points per subject, N_i, are i.i.d. realizations of the random variable N where $P(N > 1) > 0$.

Under mild conditions, auto-covariance functions defined above have orthogonal expansions in terms of eigenfunctions $\psi_m(\cdot)$ and $\varphi_p(\cdot)$ with non-increasing eigenvalues γ_m and η_p,

$$G_X(s,t) = \sum_{m=1}^{\infty} \gamma_m \psi_m(s)\psi_m(t), \quad G_Y(s,t) = \sum_{p=1}^{\infty} \eta_p \varphi_p(s)\varphi_p(t), \quad \text{for } s,t \in [0,T]. \tag{5.2}$$

Then, based on the Karhunen-Loéve expansion (Ash and Gardner, 1975), the observations X_{ij} and Y_{ij} can be represented as

$$X_{ij} = \mu_X(T_{ij}) + \sum_{m=1}^{\infty} \zeta_{im}\psi_m(T_{ij}) + \varepsilon_{ij}, \quad Y_{ij} = \mu_Y(T_{ij}) + \sum_{p=1}^{\infty} \xi_{ip}\varphi_p(T_{ij}) + e_{ij}, \tag{5.3}$$

where ζ_{im} and ξ_{ik} denote the mean zero functional principal component scores with second moments equal to the corresponding eigenvalues γ_m and η_p for $\sum_{m=1}^{\infty} \gamma_m < \infty$ and $\sum_{p=1}^{\infty} \eta_p < \infty$.

Let Δ_t denote the interval, $[t-\delta_1, t-\delta_2]$, where $0 < \delta_1 < \delta_2 < T,\ t \in [\delta_1, T]$, and let Δ denote the interval $[0, \delta_1 - \delta_2]$. Note that the first lag δ_1 is the time point beyond which the predictor process does not have an effect on the response process and the second lag δ_2 allows a delay for the predictor process to start having an effect on the response process. The recent history functional linear model where $\beta(s,t) = \sum_{k=1}^{K} b_k(t)\phi_k(s-t+\delta_1)$, can be given as

$$\begin{aligned} E\{Y(t)|X(s),\ s \in \Delta_t\} &= \alpha(t) + \sum_{k=1}^{K} b_k(t) \int_{\Delta_t} X(s)\phi_k(s-t+\delta_1)ds \\ &= \alpha(t) + \sum_{k=1}^{K} b_k(t)\widetilde{X}_k(t), \end{aligned} \tag{5.4}$$

for K predetermined basis functions $\phi_k(t)$ defined on Δ. In (5.4) $\widetilde{X}_k(t) = \int_{\Delta_t} X(s)$ $\phi_k(s-t+\delta_1)ds,\ k = 1, \dots, K$, are the induced covariates and $b_k(t),\ k = 1, \dots, K$, are the unknown time varying coefficient functions of interest. Defining $\widetilde{X}(t) = [\widetilde{X}_1(t), \dots, \widetilde{X}_K(t)]^{\mathrm{T}}$ and $b(t) = [b_1(t), \dots, b_K(t)]^{\mathrm{T}}$, the model in (5.4) can be rewritten in vector form as $E\{Y(t)|X(s),\, s \in \Delta_t\} = \alpha(t) + b^{\mathrm{T}}(t)\widetilde{X}(t)$.

In (5.4), K controls the resolution of the fit and should be chosen based on the data. Depending on the specific features of the regression function, various basis functions such as Fourier, truncated power, eigen and B-spline bases can be used in (5.4). Because of their fast computation and good properties, we will use the B-spline basis in the following calculations. For more discussions on the B-spline basis, see Fan and Gijbels (1996) and Ramsay and Silverman (2005).

5.3 Estimation and Asymptotic Properties

5.3.1 Proposed estimation algorithm

In the proposed estimation procedure, we utilize connections of the proposed model to varying coefficient models. There are three main estimation methods for varying coefficient models proposed in the literature: local polynomial smoothing (Wu et al., 1998; Hoover et al., 1998; Fan and Zhang, 2000; 2008; Kauermann and Tut, 1999), polynomial spline (Huang et al., 2002; 2004; Huang and Shen, 2004) and smoothing spline (Hastie and Tibshirani, 1993; Hoover et al., 1998; Chiang et al., 2001).

Note that the previously proposed methods cannot be directly employed here, since the induced covariates, $\widetilde{X}_k(t)$'s in (5.4) cannot be estimated well for sparse designs due to the difficulty in numerically approximating the integral in their definition. In fact, the induced covariates may not be well approximated not only in sparse designs but also in longitudinal data in general, since the integration involved is over a narrow window into the past, where there may not be enough points. We propose to base the estimation on the covariance structure to address this difficulty, which will be shown to adjust for measurement error in predictors as well.

Let $\widetilde{X}_{ik}(t)$ denote observation of the k^{th} induced covariate in (5.4) for the

i^{th} subject taken at time point t, that is, $\widetilde{X}_{ik}(t) = \int_{\Delta_t} X_i(s)\phi_k(s)ds$, and $\widetilde{X}_i(t) = [\widetilde{X}_{i1}(t), \ldots, \widetilde{X}_{iK}(t)]^{\mathrm{T}}$. The proposed estimation rests on the following equality that follows from (5.4)

$$\nu(t)b(t) = \theta(t), \tag{5.5}$$

where $\theta(t)$ is the $K \times 1$ vector with the k^{th} element equal to $\text{cov}\{\widetilde{X}_k(t), Y(t)\}$, $\nu(t)$ is the $K \times K$ matrix with the $(k,\ell)^{th}$ element equal to $\text{cov}\{\widetilde{X}_k(t), \widetilde{X}_\ell(t)\}$ and $b(t)$ is the $K \times 1$ vector of K varying coefficient functions. The elements $\nu_{k\ell}(t)$ and $\theta_k(t)$ can be given in terms of auto- and cross-covariance functions of the predictor and response processes as

$$\nu_{k\ell}(t) = \text{cov}\{\widetilde{X}_k(t), \widetilde{X}_\ell(t)\} = \int_{\Delta_t}\int_{\Delta_t} G_X(s_1, s_2)\phi_k(s_1 - t + \delta_1)\phi_\ell(s_2 - t + \delta_1)ds_1 ds_2 \tag{5.6}$$

and

$$\theta_k(t) = \text{cov}\{\widetilde{X}_k(t), Y(t)\} = \int_{\Delta_t} G_{XY}(s,t)\phi_k(s - t + \delta_1)ds, \tag{5.7}$$

where $G_{XY}(s,t) = \text{cov}\{X(s), Y(t)\}$. The estimation of the components given in (5.6) and (5.7) begins with estimation of the mean functions, $\mu_X(t)$ and $\mu_Y(t)$, via locally smoothing the aggregated data (T_{ij}, X_{ij}) and (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. For the estimation of the auto- and cross-covariance functions, we apply two-dimensional local linear smoothing to the raw auto- and cross-covariances defined as, respectively,

$$\begin{aligned} G_{X,i}(T_{ij}, T_{ij'}) &= \{X_{ij} - \hat{\mu}_X(T_{ij})\}\{X_{ij'} - \hat{\mu}_X(T_{ij'})\}, \\ G_{XY,i}(T_{ij}, T_{ij'}) &= \{X_{ij} - \hat{\mu}_X(T_{ij})\}\{Y_{ij'} - \hat{\mu}_Y(T_{ij'})\}, \end{aligned}$$

for $i = 1, \ldots, n$, $j, j' = 1, \ldots, N_i$. To obtain smooth estimates, the raw covariances are fed into a two dimensional local smoothing algorithm where special care needs to be taken in estimating the auto-covariance surface. Since the only terms in the raw auto-covaraince matrix that are perturbed by the additive measurement error on the predictors are along the diagonal, we remove the diagonal before the application of two dimensional smoothing, following (Yao et al., 2005a). Explicit forms of mean and covariance function estimators are given in Appendix D. Plugging the covariance function estimators into equations (5.6) and (5.7), and performing numerical integrations, we can obtain the estimators of $\theta(t)$ and $\nu(t)$, $\hat{\theta}(t)$ and $\hat{\nu}(t)$, respectively. Then the estimator for the varying coefficient vector, $b(t)$, is defined as

$$\hat{b}(t) = \hat{\nu}^{-1}(t)\hat{\theta}(t),$$

leading to the final estimator $\hat{\beta}(s,t) = \sum_{k=1}^{K} \hat{b}_k(t)\phi_k(s)$ for the regression surface $\beta(s,t) = \sum_{k=1}^{K} b_k(t)\phi_k(s)$.

There are two major advantages of the proposed estimation procedure. The first advantage is that, since it depends on estimation of the mean functions and covariance surfaces which are estimated from the entire data, it enables us to surmount the sparsity of the design by pooling information. The second is that it naturally adjusts for measurement error in the predictor process by considering the covariance structure and removing the diagonal terms before smoothing. In addition note that we can get estimates of the covariance surfaces on a fine set of grid points through smoothing, which allows precise numerical integration approximations in the estimation of $\theta(t)$ and $\nu(t)$.

For a given number of components, K, uniform consistency of the proposed estimator is established by the below Theorem.

Theorem 1. *Under Assumptions (A.1)-(A.5) given in Appendix B,*

$$\sup_{s,t \in \Delta_t \times [\delta_1, T]} |\hat{\beta}(s,t) - \beta(s,t)| = O_p\left\{ \frac{1}{\sqrt{n}} \left(\frac{1}{h_X^2} + \frac{1}{h_1 h_2} \right) \right\} \quad \textit{as } n \to \infty.$$

Here, h_X is the bandwidth used in obtaining the smooth auto-covariance surface of the predictor process and, h_1 and h_2 are used in obtaining the cross-covariance surface between the response and predictor processes. The set of bandwidths depend on n and are all required to converge to 0 as $n \to \infty$, for further details see B.

5.3.2 Prediction of response trajectories

We will establish the prediction of a new response trajectory, Y^*, based on the sparse predictor trajectory X^*. In what follows, let us define $P_m(t) = \int_{\Delta_t} \beta(s,t) \times \psi_m(s)ds = \sum_{k=1}^{K} b_k(t) \int_{\Delta_t} \phi_k(s)\psi_m(s)ds$, where $\psi_m(s)$ are the eigenfunctions of the auto-covariance of the predictor process defined in (5.2). From the functional representation of the predictor trajectory given in (5.3), the predicted response trajectory can be given as

$$E\left\{Y^*(t)|X^*(s),\ s \in \Delta_t\right\} = \mu_Y(t) + \sum_{m=1}^{\infty} \zeta_m^* P_m(t), \tag{5.8}$$

where $\zeta_m^* = \int_{\mathcal{T}} \left\{X^*(s) - \mu_X(s)\right\} \psi_m(s)ds$ is the m^{th} functional principal component score of the predictor process X^*. For estimation of (5.8), $\mu_Y(t)$ can be estimated from aggregated data as described above and estimators of the eigenfunctions, $\psi_m(t)$, can be obtained from the eigen-decomposition of the estimated auto-covariance surface. Details on these estimation procedures are given in Appendix D. For estimation of ζ_m^* in sparse designs, we invoke Gaussian assumptions following Yao et al. (2005b).

Let us define the collection of N^* error contaminated observations from the predictor trajectory as $X'^* = (X_1^*, \ldots, X_{N^*}^*)^{\mathrm{T}}$, where $X_\ell'^*$ is the ℓ^{th} measurement observed at time point T_ℓ^*. We assume that $(\zeta_m^*, \varepsilon_\ell^*)$, for $\ell = 1, \ldots, N^*$, are jointly Gaussian. Further define $X^* = \{X^*(T_1^*), \ldots, X^*(T_{N^*}^*)\}^{\mathrm{T}}$, $\mu_X^* = \{\mu_X^*(T_1^*), \ldots,$ $\mu_X^*(T_{N^*}^*)\}^{\mathrm{T}}$, $\psi_m^* = \{\psi_m^*(T_1^*), \ldots, \psi_m^*(T_{N^*}^*)\}^{\mathrm{T}}$ and $T^* = (T_1^*, \ldots, T_{N^*}^*)^{\mathrm{T}}$. Under the Gaussian assumption, the best linear predictor for ζ_m^* given X'^*, N^* and T^* is obtained by

$$\tilde{\zeta}_m^* = \gamma_m \psi_m^{*T} \Sigma_{X'^*}^{-1} (X'^* - \mu_X^*), \tag{5.9}$$

where $\Sigma_{X'^*} = \mathrm{cov}(X'^* | N^*, T^*) = \mathrm{cov}(X^* | N^*, T^*) + \sigma_\varepsilon^2 I_{N^*}$ with I_{N^*} denoting the $N^* \times N^*$ identity matrix and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon)$ denoting the variance of the measurement error on the predictor process. Defining $\hat{\mu}_X^*$ and $\hat{\psi}_m^*$ to be the estimators of μ_X^* and ψ_m^* respectively, the estimator of $\tilde{\zeta}_m^*$ can be given as

$$\hat{\zeta}_m^* = \hat{\gamma}_m \hat{\psi}_m^{*T} \widehat{\Sigma}_{X'^*}^{-1} (X'^* - \hat{\mu}_X^*),$$

where the $(i,j)^{th}$ element of $\widehat{\Sigma}_{X'^*}$ is defined as $\widehat{\mathrm{cov}}\{X^*(T_i^*), X^*(T_j^*)\} + \hat{\sigma}_\varepsilon^2 \delta_{ij}$ where $\hat{\sigma}_\varepsilon^2$ is the estimator of the measurement error variance and $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. The explicit form of $\hat{\sigma}_\varepsilon^2$ is deferred to Appendix D. Hence the predicted response trajectory is obtained by

$$\widehat{Y}_M^*(t) = \hat{\mu}_Y(t) + \sum_{m=1}^{M} \hat{\zeta}_m^* \widehat{P}_m(t), \tag{5.10}$$

where $\widehat{P}_m(t) = \int_{\Delta_t} \hat{\beta}(s,t) \hat{\psi}_m(s) ds = \sum_{k=1}^{K} \hat{b}_k(t) \int_{\Delta_t} \phi_k(s) \hat{\psi}_m(s) ds$. The number M of eigenfunctions used in the decomposition of the predictor auto-covariance surface, given in (5.10) can be selected by leave-one-curve-out cross validation, generalized cross validation (GCV), or the Akiake information criterion (AIC). For more details on the selection of M, see Yao et al. (2005a). A consistency result of the predicted trajectory for the target trajectory $\widetilde{Y}^*(t) = \mu_Y(t) + \sum_{m=1}^{\infty} \tilde{\zeta}_m^* P_m(t)$ is established in Theorem 2.

Theorem 2. *Under (A1)-(A5), (B1)-(B3) of Appendix B, given T^* and N^*, for all $t \in [\delta_1, T]$, the predicted trajectories satisfy*

$$\lim_{n \to \infty} \widehat{Y}_M^*(t) = \widetilde{Y}^*(t) \qquad \textit{in probability.}$$

Note that, the number M of eigenfunctions used in the eigen-decomposistion of the predictor process is a function of n and tends to infinity as $n \to \infty$.

5.3.3 Asymptotic confidence bands for the predicted response trajectories

In this section, we construct point-wise asymptotic confidence intervals for the predicted response trajectory, $\widehat{Y}^*_M(t)$. For $M \geq 1$, let us define $\zeta^{*M} = (\zeta^*_1, \ldots, \zeta^*_M)^{\mathrm{T}}$, $\tilde{\zeta}^{*M} = (\tilde{\zeta}^*_1, \ldots, \tilde{\zeta}^*_M)^{\mathrm{T}}$, where $\tilde{\zeta}^*_m$ is as defined in (5.9). Note that $\mathrm{cov}(\zeta^*_m, X'^*) = \gamma_m \psi^*_m$. Defining the matrix $H = \mathrm{cov}(\zeta^{*M}, X'^* | T^*, N^*) = (\gamma_1 \psi^*_1, \ldots, \gamma_M \psi^*_M)^{\mathrm{T}}$ of size $M \times N^*$, the covariance matrix of $\tilde{\zeta}^{*M}$ can be given in terms of H as

$$\mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) = H \Sigma^{-1}_{X'^*} H^{\mathrm{T}}.$$

Observing that $\mathrm{cov}(\tilde{\zeta}^{*M}, \zeta^{*M} | T^*, N^*) = H \Sigma^{-1}_{X'^*} \mathrm{cov}\ (X'^*, \zeta^{*M}) = H \Sigma^{-1}_{X'^*} H^{\mathrm{T}}$, we have

$$\begin{aligned} \mathrm{cov}(\tilde{\zeta}^{*M} - \zeta^{*M} | T^*, N^*) &= \mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) + \mathrm{cov}(\zeta^{*M} | T^*, N^*) \\ &\quad - 2\mathrm{cov}(\tilde{\zeta}^{*M}, \zeta^{*M} | T^*, N^*) \\ &= \mathrm{cov}(\zeta^{*M} | T^*, N^*) - \mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) \equiv \Omega_M, \end{aligned}$$

where $\Omega_M = D - H \Sigma_{X'^*} H^{\mathrm{T}}$ with $D = \mathrm{diag}\{\gamma_1, \ldots, \gamma_M\}$. Hence, under the Gaussian assumption, conditioning on T^* and N^*, $\tilde{\zeta}^{*M} - \zeta^{*M}$ is distributed as $N(0, \Omega_M)$.

Let $\hat{\zeta}^{*M} = (\hat{\zeta}^*_1, \ldots, \hat{\zeta}^*_M)^{\mathrm{T}}$ and $\widehat{\Omega}_M = \widehat{D} - \widehat{H} \widehat{\Sigma}^{-1}_{X'^*} \widehat{H}^{\mathrm{T}}$, where $\widehat{D} = \mathrm{diag}\{\hat{\gamma}_1, \ldots, \hat{\gamma}_M\}$ and $\widehat{H} = (\hat{\gamma}_1 \hat{\psi}^*_1, \ldots, \hat{\gamma}_M \hat{\psi}^*_M)^{\mathrm{T}}$. Defining $\widehat{P}(t) = \{\widehat{P}_1(t), \ldots, \widehat{P}_M(t)\}^{\mathrm{T}}$, Theorem 3 gives the asymptotic distribution of the predicted response trajectory $\widehat{Y}^*_M(t) = \hat{\mu}_Y(t) + \widehat{P}^{\mathrm{T}}(t) \hat{\zeta}^{*M}$.

Theorem 3. *Under (A1)-(A5), (B1)-(B3), and (C1) of Appendix B, given* N^*

and T^*, *for all* $t \in [\delta_1, T]$ *and* $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left\{ \frac{\widehat{Y}_M^*(t) - E\{Y^*(t)|X^*(s) \in \Delta_t\}}{\sqrt{\widehat{\omega}_M(t,t)}} \le x \right\} = \Phi(x), \tag{5.11}$$

where $\omega_M(t,t) = P^{\mathrm{T}}(t)\Omega_M P(t)$, $\hat{\omega}_M(t,t) = \widehat{P}^{\mathrm{T}}(t)\widehat{\Omega}_M\widehat{P}(t)$ *and* Φ *denotes the Gaussian cdf.*

From Theorem 3, it follows that ignoring the bias from the truncation in $\widehat{Y}_M^*$ at M eigen components, the $(1-\alpha)100(\%)$ asymptotic pointwise confidence interval for $E\{Y^*(t)|X^*(s), s \in \Delta_t\}$ is

$$\widehat{Y}_M^*(t) \pm \Phi\left(1 - \frac{\alpha}{2}\right)\sqrt{\widehat{P}_M^{\mathrm{T}}(t)\widehat{\Omega}_M\widehat{P}_M(t)}.$$

5.4 Numerical Issues in Implementation and Parameter Selection

An important issue in the implementation of the proposed estimation algorithm is the inversion of the matrix $\hat{\nu}(t)$ in equation (5.5). To obtain stable estimators for β, penalized solutions have been studied in literature (Cardot et al., 2003) minimizing the penalized least squares

$$\{\hat{\nu}(t)b(t) - \hat{\theta}(t)\}^{\mathrm{T}}\{\hat{\nu}(t)b(t) - \hat{\theta}(t)\} + \lambda b^{\mathrm{T}}(t)Qb(t),$$

where Q is a $K \times K$ matrix that determines the type of penalty used. Here, λ is a tuning parameter that controls the amount of regularization and should be chosen from data balancing the stability and validity of the resulting estimator. The penalized estimator is then given by $\hat{b}_\lambda(t) = \{\hat{\nu}(t) + \lambda Q\}^{-1}\hat{\theta}(t)$. A common choice of Q is the $K \times K$ identity matrix, which leads to the ridge solution, $\hat{b}_\lambda(t) = \{\hat{\nu}(t) + \lambda I\}^{-1}\hat{\theta}(t)$. Another common choice of penalty used is on the degree of smoothness of the penalized solution where Q is chosen such that its $(i,j)^{th}$ element is equal to $\int_\Delta \phi_i^{(m)}(s)\phi_j^{(m)}(s)ds$ with $\phi_k^{(m)}(t)$ denoting the m^{th} derivative of the k^{th} predetermined basis function. In the following applications, we use the ridge solution.

Hence, the proposed estimation procedure for the recent history functional linear model includes three sets of parameters with different roles: K, λ, and (δ_1, δ_2). The number of predetermined basis functions used, K, controls the resolution of the fit; the tuning parameter, λ, controls the stability of the estimates; and the window, (δ_1, δ_2), determines the model used via controlling the predictor window affecting the response. An important observation in the current estimation set-up is that the proposed estimation procedure is not sensitive to the choice of K provided that there are enough basis functions used in the estimation, since the penalized solution employed via λ prevents over-fitting. Instead the choice of the tuning parameter λ is a more important choice controlling the stability of the estimates. This fact has been pointed out before in similar functional estimation problems, see Cardot et al. (2003) for further details. We run multiple simulation studies comparing the estimated regression surfaces obtained with different choices of K and λ to confirm this observation, where the results are summarized in Section 5.5.2. Based on the above argument, following Cardot et al. (2003), we fix K at a value that guarantees good precision and concentrate on the choice of λ and (δ_1, δ_2). For example, K is fixed at 10 in both the simulation studies and the data example that follow, to include 10 B-spline basis functions of order 4 with 6 interior equi-distance knots in the window Δ, which prove to provide good precision.

For the selection of λ and (δ_1, δ_2), consider the following two criteria. Define the normalized prediction error (NPE) as

$$\text{NPE}\{\lambda; (\delta_1, \delta_2)\} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i^\delta} \frac{|\widehat{Y}_{ij} - Y_{ij}|}{|Y_{ij}|},$$

where $\widehat{Y}_{ij}$ is the predicted value for the j^{th} measurement on the i^{th} response trajectory obtained using λ and (δ_1, δ_2), N_i^δ is the number of observations from the i^{th} subject obtained in the interval, $[\delta_1, T]$, and $N = \sum_{i=1}^{n} N_i^\delta$. Note that NPE measures the relative absolute prediction error which will also be used in evaluating the finite sample performance of the proposed estimates in the simulations given in Section 5.5. We also define leave-one-curve-out cross validation squared prediction error as

$$CV\{(\delta_1, \delta_2); \lambda\} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{N_i^\delta} \{\widehat{Y}_{ij}^{(-i)} - Y_{ij}\}^2},$$

where $\widehat{Y}_{ij}^{(-i)}$ is the fitted response trajectory value of the i^{th} subject at time point T_{ij} obtained from the data excluding the i^{th} subject. Note that in both criteria, NPE and the cross validation score, the fitted response values are obtained via the prediction methods proposed in Section 5.3.2.

We propose to estimate λ and (δ_1, δ_2) in a hierarchical manner, where first λ is chosen for a set of (δ_1, δ_2) values using NPE and finally the set of (δ_1, δ_2) values are compared to yield the final choice of (δ_1, δ_2) with minimum cross validation score. This hierarchical approach has the advantage that the computationally faster NPE criterion is used within the inside loop of choosing λ and that the relatively more refined cross validation score is utilized for the selection of (δ_1, δ_2), which is similar to model selection in an outer loop. More specifically, the algorithm can be outlined as follows. Let Λ and D be the predetermined sets of λ and (δ_1, δ_2) values considered. For a fixed $(\delta_1, \delta_2)_0 \in D$, NPE values are calculated for all $\lambda \in \Lambda$. The λ that gives the smallest NPE value is selected as the optimal λ for the given $(\delta_1, \delta_2)_0$ and is thus used for calculating the cross validation score for $(\delta_1, \delta_2)_0$. Repeating the above steps for all $(\delta_1, \delta_2) \in D$, the optimal (δ_1, δ_2) is selected to be the one with the minimum cross validation score.

5.5 Simulation Studies

In this section, we investigate the finite sample properties of the proposed estimators through two simulation studies. In the first simulation, we study efficiency of the NPE criterion for selecting the tuning parameter λ for a fixed (δ_1, δ_2) choice, along with the finite sample performance of the proposed estimator based on the selected λ from the algorithm. In the second simulation, we evaluate the performance of the cross validation method for choosing (δ_1, δ_2) with varying sample size. The following simulation results are reported based on 500 Monte Carlo runs.

5.5.1 Data generation

For n subjects, the number of measurements made on the i^{th} predictor and response processes, N_i, are uniformly selected from 3, 4, and 5. The design points, $T_{i1}, \ldots, T_{iN_i}$ are randomly selected from the uniform distribution between 0 and 50. At a given time point t, the predictor trajectory is evaluated around the mean function $t + \sin(t)$ using two principal components, $\psi_1(t) = -\cos(\pi t/50)/\sqrt{5}$ and $\psi_2(t) = \sin(\pi t/50)/\sqrt{5}$ for $t \in [0, 50]$. The corresponding two eigen component scores, ζ_1 and ζ_2, are independently sampled from the Gaussian distributions with mean 0 and the variances, $\rho_1 = 4$ and $\rho_2 = 1$, respectively. Hence, the predictor process at time t is generated via $X(t) = t + \sin(t) + \sum_{m=1}^{2} \zeta_m \psi_m(t)$. In addition, i.i.d. measurement error, ε, simulated from the Gaussian distribution with mean 0 and variance $\sigma_\varepsilon^2 = 0.025$, is added to the predictor observations in accordance with (5.3). The regression function is generated based on the same basis functions used for the predictor process via, $\beta(s,t) = \sum_{i=1}^{2} \sum_{j=1}^{2} c_{ij} \psi_i(s) \psi_j(t),\ \ s \in \Delta_t$, where $c_{11} = 2, c_{12} = 2, c_{21} = 1$, and $c_{22} = 2$. The true window combination, (δ_1, δ_2), is set at $(20, 0)$. The error process, $\epsilon(t)$, is also generated using the same basis functions as the predictor process with eigen component scores independently sampled from the Gaussian distributions with mean 0 and variances 0.025, 0.004, respectively.

The response trajectory, $Y(t)$ is generated by the equation, $\int_{\Delta_t} \beta(s,t)X(s)ds+\epsilon(t)$ using the numerical integration procedure. To obtain a noisy version of the response trajectory, we add i.i.d measurement error, e, generated from the Gaussian distribution with mean 0 and variance 0.025.

5.5.2 Simulation results

To evaluate the performance of the λ selection criterion, NPE, let us define relative square deviation (RSD) at time t as

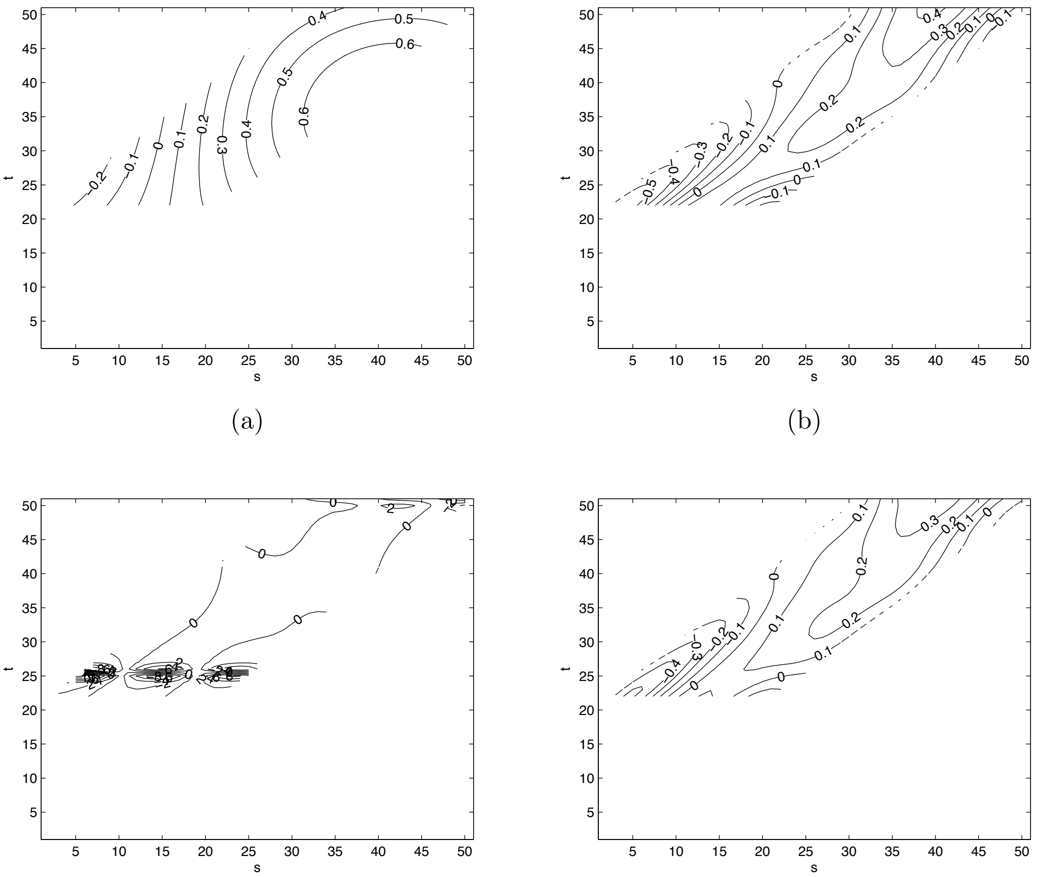
$$\text{RSD}(t) = \frac{\int_{\Delta_t}\{\hat{\beta}(s,t) - \beta(s,t)\}^2 ds}{\int_{\Delta_t} \beta(s,t)^2 ds}.$$

The RSD measures the relative size of the squared difference between the estimated and true regression functions at time t. The RSD integrated over the entire support, $t \in [\delta_1, T]$, will be called integrated RSD, denoted IRSD.

Before reporting on results from the two simulation set-ups, we present the true and estimated regression functions with different λ and K choices to demonstrate the relative importance of λ over K in the estimation procedure discussed in Section 5.4. The true and estimated regression functions are plotted in Figure 5.1 from three Monte-Carlo simulation runs at $n = 200$ with (K, λ) equal to $(5, 7.4)$ (Figure 5.1b, for the NPE minimizer ($\lambda = 7.4$)), $(5, 1)$ (Figure 5.1c) and $(10, 7.4)$ (Figure 5.1d), respectively.

While the difference between the estimators plotted in Figures 1b and 1d is small corresponding to doubling of the K choice, the estimator given in Figure 5.1c at the wrong λ value of 1 cannot recover the true regression function with IRSD$= 298$. This confirms that the choice of λ plays a more important role in the proposed estimation algorithm than the choice of K.

The estimated regression function $\hat{\beta}(s,t)$ with the tuning parameter λ equal to the minimizer of IRSD can be thought of as the "optimal estimate", since



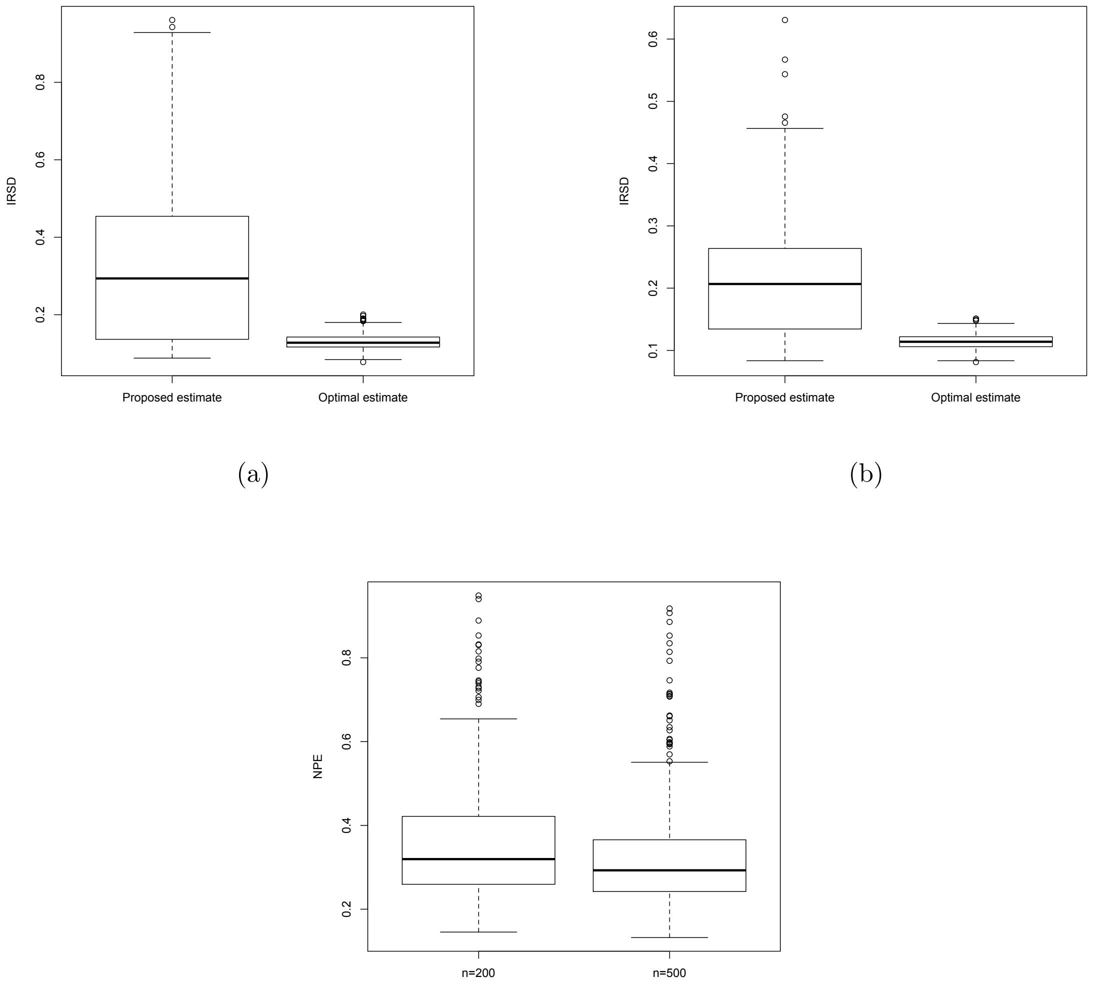
(a) (b)

(c) (d)

Figure 5.1: (a) The true regression function defined on $\Delta = [0, 20]$. The estimated regression function with (K, λ) equal to $(5, 7.4)$ (plot b, estimated IRSD= 0.3632), $(5, 1)$ (plot c, estimated IRSD= 298) and $(10, 7.4)$ (plot d, estimated IRSD= 0.3274).

information on the true regression function is utilized in the comparison. Hence, this estimate can only be obtained in a simulation setting where the true regression function is known. We compare this choice of λ with the minimizer of NPE, which is the only estimate that can be obtained from the data in reality. The performance of the two estimators is compared in Figure 5.2 in terms of IRSD, where boxplots of estimated IRSD values are given for the optimal and proposed estimates at $n = 200$ and 500 from 500 Monte-Carlo simulation runs. In the displayed boxplots, 22 and

12 outliers are removed for $n = 200$ and $n = 500$, respectively.



(a)

(b)

(c)

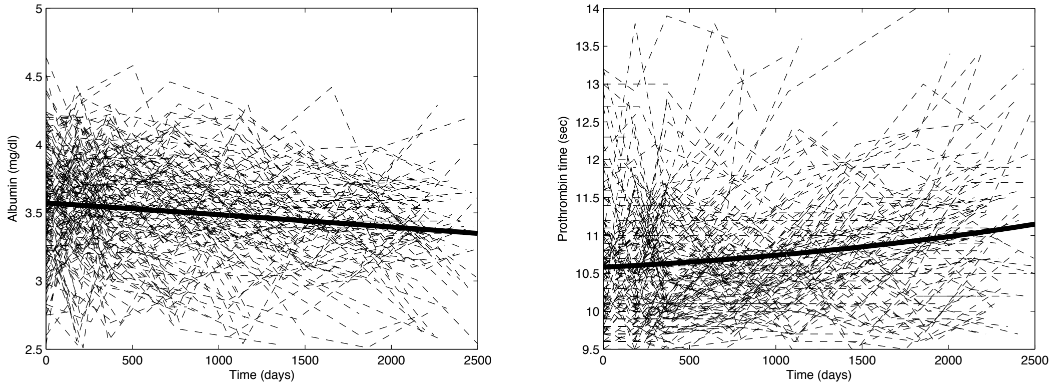
Figure 5.2: Boxplots of the estimated IRSD values of the proposed and optimal estimators for $n = 200$ (a) and $n = 500$ (b). (c) Boxplot of NPE values of the proposed estimators for $n = 200$ and $n = 500$.

Figure 5.2 suggests that the estimator with the λ choice selected via NPE performs close to the one with the optimal choice of λ. Note that the performance of both estimators improve with increasing sample size. For the proposed estimators with λ chosen by NPE, the median estimated IRSD is 0.2937 at $n = 200$ and is 0.20670 at $n = 500$, which implies that the estimated regression surface is close to the true one. We also give in Figure 5.2 the boxplots of NPE values of the proposed estimator with λ chosen by NPE. The median NPE value at $n = 200$ is 0.3196 and the median NPE drops to 0.2928 for $n = 500$.

The performance of leave-one-curve-out cross validation score for the selection of δ is studied through the second simulation. For the computational efficiency, we use 10-fold cross validation. The true value of (δ_1, δ_2), is set to be $(20, 0)$, where 5 candidate (δ_1, δ_2) pairs are considered at $(30, 0), (30, 5), (20, 0)$,
$(20, 5)$, and $(10, 5)$. The correct (δ_1, δ_2) choice ratio out of 500 Monte Carlo runs are 0.8929 and 0.9214 for $n = 200$ and 500, respectively. There seems to be improvement with increasing sample size.

5.6 Data Analysis

To demonstrate the proposed method, we include an application to the longitudinal primary biliary liver cirrhosis data collected between January 1974 and May 1984 by the Mayo Clinic. In the study design, patients were scheduled to visit the clinic at six months, one year and annually thereafter post diagnosis, where certain blood characteristics were recorded. However, due to missed visits, the data is sparse and highly irregular where each patient visited the clinic at different times. We explore the dynamic relationship between serum albumin level in mg/dl (predictor) and prothrombin time in seconds (response). Both variables are used as an indicator of the liver function, where a decrease in serum albumin levels and elevated prothrombin times are typically associated with malfunctioning of the liver (Murtaugh et al., 1994). We include 201 female patients in the analysis where predictor and response measurements before 2500 days are considered. The number of observations per subject ranges from 1 to 9, with a median of 5 measurements. Individual trajectories of the serum albumin level and prothrombin time



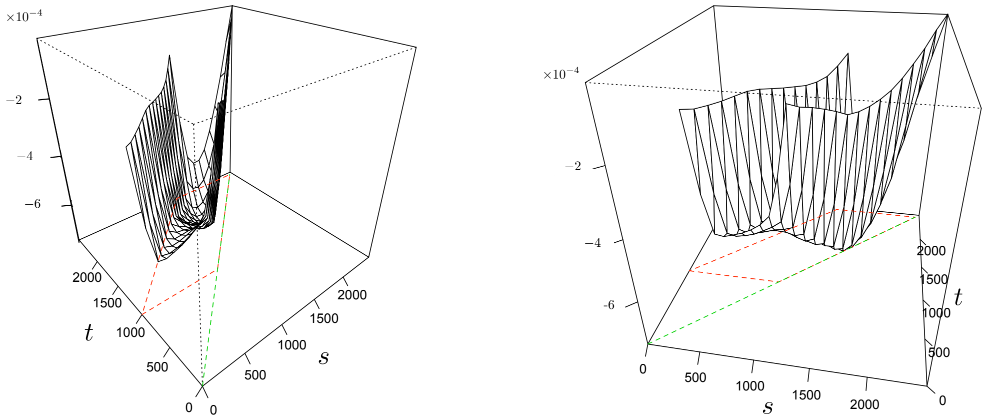
(a) (b)

Figure 5.3: (a) Individual predictor trajectories (dashed) overlaying the estimated cross-sectional mean of the predictor process (solid). (b) Individual response trajectories (dashed) overlaying the cross-sectional mean of the response process (solid).

overlaying their respective estimated mean functions are given in Figure 5.3. The estimated mean functions indicate opposite patterns as expected, where there is a decreasing trend for the predictor process and an increasing trend for the response.

We next fit the proposed recent history functional linear model to the data with $K = 10$ B-spline functions. Among the five candidate (δ_1, δ_2) choices considered, $[1000, 0]$, $[500, 0]$, $[700, 200]$, $[1000, 200]$, and $[1000, 500]$, the minimizer of cross-validation error was $[1000, 0]$, and the NPE choice of λ was 6502. To save on computational time, we report results from 10 fold cross validation.

The estimated regression surface is displayed in Figure 5.4 viewed from two different angles. Most of the estimated regression surface is negative, stressing the general opposing trends also observed in the literature between serum albumin levels and prothrombin time. For a given time point, the albumin concentration has the strongest effect in magnitude on prothrombin time with a delay of about 500 days where the effect decays as the lag increases. Note also that the observed negative effect of the past albumin concentration levels on the current prothrombin



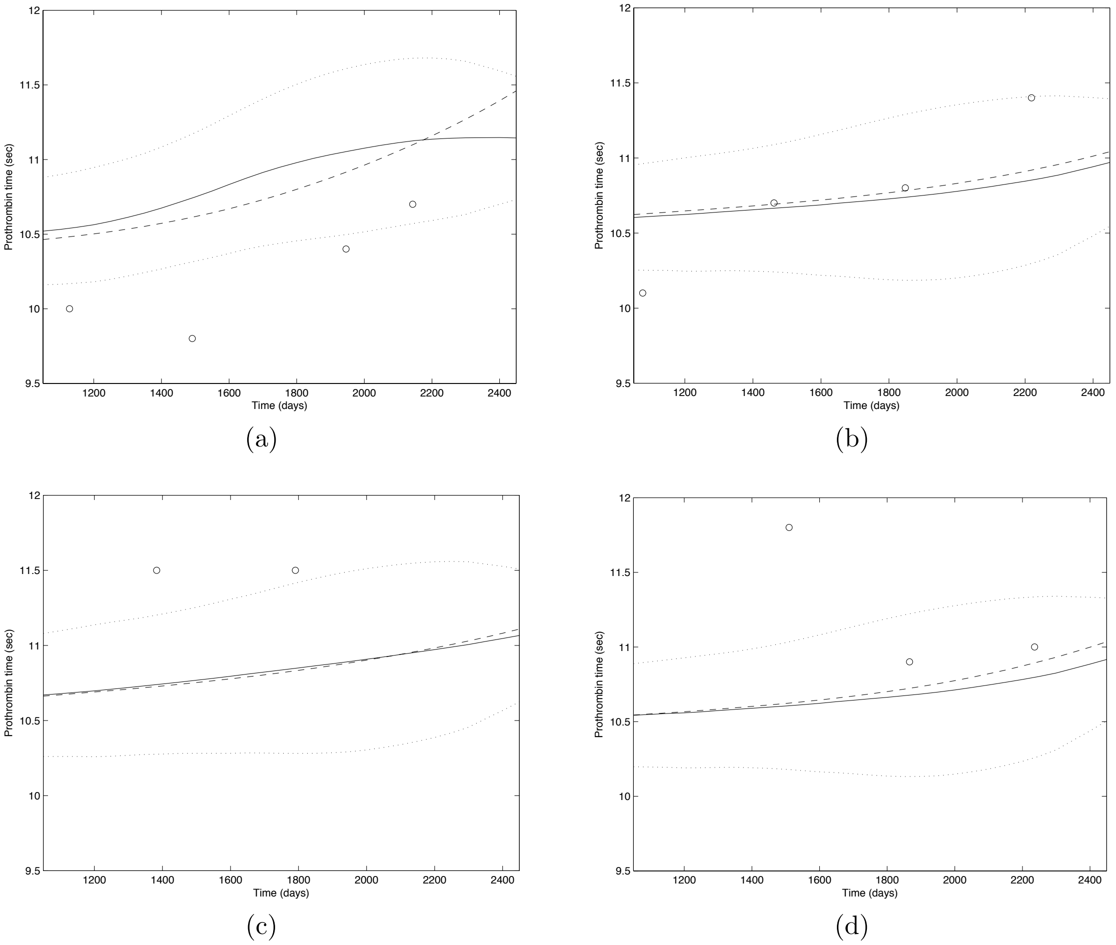
(a)

(b)

Figure 5.4: The estimated regression surface defined on $[t-1000, t] \times [1000, 2500]$ in the longitudinal primary biliary liver cirrhosis data obtained by $\lambda = 6502$ and $K = 10$.

time seems to get more pronounced towards the later stages of the study and hence the disease.

Predicted response trajectories for four randomly selected subjects obtained from the proposals of Section 5.3.2 and 5.3.3 and the corresponding 95% asymptotic confidence bands are given in Figure 5.5. We also include for comparison predicted trajectories obtained from the functional linear model proposed by Yao et al. (2005a). The predicted trajectories seem to be quite close to those from a functional linear model fit which uses the entire predictor trajectory including observations from future and distant past measurement times in the predictions as well. Hence, we conclude that the recent history functional linear model, which models the effects of the predictor process from the recent past of 1000 days till the present time, provides a reasonable model for the data in terms of prediction. Given its ease in interpretation due to its restricted regression support when compared with the full functional linear model, the proposed model emerges as a viable



(a) (b) (c) (d)

Figure 5.5: Predicted trajectories for 4 randomly selected subjects. In each plot, circles denote the original response observations, the solid line is the response trajectory predicted via the proposed method, the dashed line is the predicted trajectory using the functional linear fit of Yao et al. (2005a), and the dotted lines represent the 95% asymptotic confidence bands.

alternative for the analysis of the current data set.

5.7 Discussion

We proposed an estimation algorithm for the recent history functional linear models which are useful in applications. The sliding window support of the recent history functional linear model strikes a useful balance between the global support of the functional linear models and the point wise support of the varying coefficient models. The assumption that only the predictor process from the recent past has an effect on the response rather than the future predictor values or only the current predictor value, is useful in many applications where changes in the response process can be explained using recent trends of the predictor process. In addition the product form assumed for the regression surface uses only one dimensional smooth functions considerably easing and speeding estimation.

Our proposal is geared towards sparse longitudinal data where the estimation procedure proposed also accommodates measurement error in variables. Sparsity and measurement error are both commonly encountered in longitudinal designs. We provide asymptotic properties of our estimators that enable the estimation of the predicted response trajectories and that lead to asymptotic confidence bands. Choice of model parameters is also addressed, where favorable properties of the proposed estimators are demonstrated in simulations and data applications.

Chapter 6

Future Work

In this dissertation, we proposed the recent history functional linear model and developed estimation procedures geared towards three different types of data: functional, longitudinal, and sparse longitudinal data. For functional and longitudinal data with a moderate number of observations per subject, we made use of two estimation methods for the varying coefficient model. In the case of sparse longitudinal data, not having enough observations for the approximation of integrals involved in the proposed model, we utilized the covariance structure in estimation to effectively pool information across subjects. Based on the covariance function, we proposed a ridge type solution to cope with the inverse problem of the covariance function. We established the consistency of the proposed estimator and normality of the predicted values relying on the normality assumption for the functional principal component scores of the predictor processes.

In this chapter we present three possible directions of future study. One is the development of more precise asymptotic theory for the estimator proposed in Chapter 5. More technically, in Chapter 5, we established consistency of the proposed estimator assuming the number of basis functions is fixed and ignoring the approximation error of the expansion. We also simplified the problem by assuming invertibility of the covariance function. Hence, it is of interest to study the asymptotic behavior of the estimator as the number of basis functions, K, increases including the convergence rate of the proposed estimator and the relationship between K and n. Considering the tuning parameter, ρ, in asymptotic study will be of interest as well.

The second and third directions of future work are to the extensions of the proposed model: 1) to include multiple predictor processes and 2) to include random functional effects. In Chapters 3, 4, and 5, we discussed the recent history functional linear model defined by

$$Y_i(t) = \alpha(t) + \int_{\Delta_t} X_i(s)\beta(s,t)ds + \varepsilon_i(t), \quad t \in [\delta_1, T],\ 0 \leq \delta_2 \leq \delta_1.$$

This model can be extended to one with P predictor processes as

$$Y_i(t) = \alpha(t) + \sum_{p=1}^{P} \int_{\Delta_t} X_{ip}(s)\beta_p(s,t)ds + \varepsilon_i(t), \quad t \in [\delta_1, T],\ 0 \leq \delta_2 \leq \delta_1.$$

The functional linear model with multiple predictor processes for global support was proposed by Han et al. (2007) for the market valuation of companies based on historical, present, and future financial ratio information such as adjusted debt ratio, Q-ratio, and CFRoI. They proposed an estimation method based on a backfitting algorithm using the regression function estimator proposed by Yao et al. (2005a). We, however, cannot apply their method directly to the recent history functional linear model with multiple predictors due to the non-rectangular support property. Instead, we can extend the proposed estimator to this model for functional data and longitudinal data with a moderate number of observations.

The recent history functional linear model only deals with single-level functional or longitudinal data, i.e., all subjects belong to a single population. A common way of analyzing the observations from multiple subgroups is to include random effects in the model. There has not been any work on functional linear model with random effects proposed in the literature to the best of our knowledge. Di et al. (2009) developed multilevel functional principal component analysis (MFPCA) to analyze the group specific mean functions of hierarchical longitudinal data, particularly, Sleep Heart Health Study (SHHS). For the recent history functional linear model, the random effects can be included as

$$Y_{ij}(t) = \alpha(t) + \int_{\Delta_t} X_{ij}(s)\beta(s,t)ds + \int_{\Delta_t} Z_{ij}(s)\gamma_i(s,t)ds + \varepsilon_{ij}(t) \tag{6.1}$$

for the j^{th} observed processes in the i^{th} group, where $\beta(s,t)$ is the regression function for the entire population and $\gamma_i(s,t)$ is the i^{th} subject specific regression function. One way of modeling $\gamma_i(s,t)$ is to expand it on a basis function system, $\phi_k(\cdot)$, $k = 1, \ldots, K_2$, as $\gamma_i(s,t) = \sum_{k=1}^{K_2} \phi_k(s)u_{ik}(t)$, where $u_i = [u_{i1}, \ldots, u_{iK_2}]^T \sim N(0,V)$. Here, V is the covariance matrix of u_i. The expansion of the regression function, $\beta(s,t)$, on another set of basis function systems, $\varphi_l(\cdot)$, $l = 1 \ldots, K_1$, enables us to approximate the model given in (6.1) as

$$Y_{ij}(t) \approx \alpha(t) + \sum_{k=1}^{K_1} \psi_{ijk}^{\mathrm{T}}(t)b_k(t) + \sum_{l=1}^{K_2} \varphi_{ijl}^{\mathrm{T}}(t)u_{ijl}(t) + \varepsilon_{ij}(t),$$

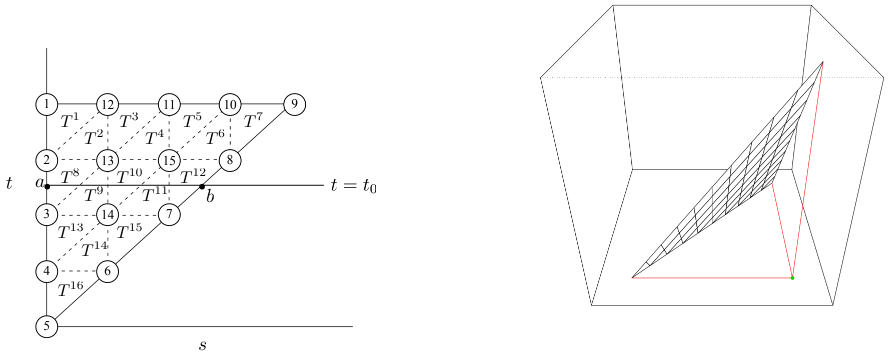
where $\psi_{ijk}(t) = \int_{\Delta_t} X_{ij}(s)\varphi_k(s)ds$ and $\varphi_{ijl}(t) = \int_{\Delta_t} Z_{ij}(s)\psi_l(s)ds$. This model can be considered as an extension of the random varying coefficient model proposed by Wu and Liang (2004). In their paper, they also proposed an estimation method based on a backfitting algorithm, and this method can be applied to the functional linear model with random effects with a slight modification.

Appendix A

Finite Element Method

Malfait & Ramsay (2003) used a computational technique called finite element method (FEM). FEM is a method that approximates a function by the union of lower order polynomial approximations on a finite number of finely divided regions of the support. They applied FEM instead of the tensor product basis function introduced in Chapter 2 to their model mainly because it does not have rectangular support due to the exclusion of future values of predictor process in the evaluation of the current response function value. In their application, they divided the support into small triangular elements and constructed two dimensional basis function on those elements.

To construct the basis function, first let us partition the support into n^2 triangular elements by dividing each axis into n subintervals of length λ and divide each rectangular area once again into two triangular elements by the line parallel to $s = t$ as shown in Figure A.1a. We call the grid point resulting from the division of both axes a node. Note that we have $K = (n+1)(n+2)/2$ nodes for the entire triangular support, which implies that the whole past values of the predictor are used to predict the current value of the response function. Considering the lag δ from where the predictor function starts to have impact on the current value of response function, we discretize the lag δ to $m\lambda$ and discard all triangles contain-



(a) Support of $\beta(s, t)$ divided into 16 triangular elements by 5 subintervals resulting in 15 nodes.

(b) Piecewise linear plane on a triangular elements when the right bottom vertex is the node for which we are constructing the basis fucntion $\phi_k(s, t)$.

Figure A.1: Triangular Support of coefficient function $\beta(s, t)$ and the piecewise linear plane on a signle triangular element.

ing points more than $m\lambda$ units from the diagonal in the horizontal direction. This leaves $m(2n - m)$ triangular elements covering the domain of $\beta(s, t)$ and we have $K = (m + 1)(n + 1 - m/2)$ nodes.

Node k inside the support has six adjacent triangular elements. For example, node 13 in Figure A.1a has $T^2, T^3, T^4, T^{10}, T^9$ and T^8 as triangular elements surrounding it. These six triangular elements form the hexagon support for the basis function at node k. The k^{th} basis function $\phi_k(s, t)$ at node k is defined as a tent shaped function on the hexagon support around node k. This basis function can be obtained by patching six pieces of flat plane on six triangular elements around node k. If bottom right vertex is the node for which we are constructing a basis, then the plane defined on a triangular element is given in Figure A.1b.

Each plane on the triangular element must have $\phi_k(s, t) = 1$ at node k and $\phi_k(s, t) = 0$ on two opposite vertices. Moreover, two planes on adjacent triangular elements must agree on the side they share.

Let us consider how the linear plane on triangular elements can be constructed. Note that each point (s, t) on the support can be evaluated on three vertices around it so that we can have three positive basis functions at a certain point. For example given in Figure A.1a, any point (s, t) in triangular element T^{10} will have positive basis function value for node 13, 14 and 15 and 0 for the rest. The basis functions ϕ_ν^e associated with the vertices of triangle T^e can be defined by the area coordinates for that triangle. For an arbitrary point $(s, t) \in T^e$ shown in Figure A.2, the basis functions $\phi_\nu^e, \nu = 1, 2, 3$ are defined as the ratio of the triangular subarea A_ν^e to the whole area A^e as

$$\phi_\nu^e(s, t) = \frac{A_\nu^e(s, t)}{A^e(s, t)}. \tag{A.1}$$

Note that the value of basis function $\phi_\nu^e(s, t)$ on triangulare element equals 1 at vertex ν but zero at two other vertices, and for all $(s, t) \in T^e$, $\sum_{\nu=1}^{3} \phi_\nu^e(s, t) = 1$.

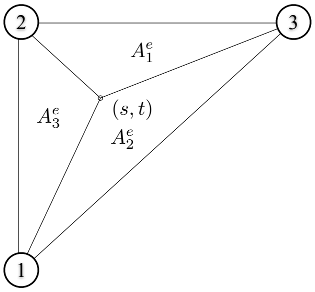


Figure A.2: The basis functions on node $\nu = 1, 2, 3$ associated with point (s, t) in a triangular element are the ratios of each of the subareas A_ν^e to the total area A^e of the triangle.

Appendix B

Assumptions

We present the assumptions in three groups, assumptions (A) are needed for all three Theorems in Chapter 5, assumptions (B) are needed for the consistency and asymptotic normality of the predicted response trajectories given in Theorems 2 and 3 respectively, and assumption (C1) is only used in Theorem 3.

The data (T_{ij}, X_{ij}) and (T_{ij}, Y_{ij}), $i = 1, \ldots, n, j = 1, \ldots, N_i$, are assumed to be the i.i.d. samples from the joint densities, $g_1(t, x)$ and $g_2(t, y)$. Assume also that the observation times T_{ij} are i.i.d. with marginal densities $f_{\mathcal{T}}(t)$. Let T_1 and T_2 be two different time points, and X_1 (respectively Y_1) and X_2 (respectively Y_2) be the repeated measurements of X (respectively Y) made on the same subject at times T_1 and T_2. The predictor (and response) measurements made on the same subject at different times are allowed to be dependent. Assume $(T_{ij}, T_{il}, X_{ij}, X_{il})$, $1 \leq j \neq l \leq N_i$, is identically distributed as (T_1, T_2, X_1, X_2) with joint density function $g_{XX}(t_1, t_2, x_1, x_2)$ and analogously for $(T_{ij}, T_{il}, Y_{ij}, Y_{il})$ with identical joint density function $g_{YY}(t_1, t_2, y_1, y_2)$. The following regularity conditions are assumed on $f_{\mathcal{T}}(t), g_1(t, x), g_2(t, y), g_{XX}(t_1, t_2, x_1, x_2)$ and $g_{YY}(t_1, t_2, y_1, y_2)$. Let p_1, p_2 be non-negative integers with $0 \leq p_1 + p_2 \leq 4$.

(A1) The derivative $(d^p/dt^p) f_{\mathcal{T}}(t)$ exists and is continuous on $[0, T]$ with $f_{\mathcal{T}}(t) > 0$ on $[0, T]$, $(d^p/dt^p) g_1(t, x)$ and $(d^p/dt^p) g_2(t, y)$ exist and are continuous on

$[0,T] \times \mathbb{R}$, and $\{d^p/(dt^{p_1}dt^{p_2})\}g_{XX}(t_1,t_2,x_1,x_2)$ and $\{d^p/(dt^{p_1}dt^{p_2})\}g_{YY}(t_1,t_2,x_1,x_2)$ exist and are continuous on $[0,T]^2 \times \mathbb{R}^2$ for $p_1 + p_2 = p, 0 \leq p_1, p_2 \leq p$.

(A2) The number of measurements N_i made for the i^{th} subject is a random variable such that $N_i \overset{i.i.d}{\sim} N$, where N is a positive discrete random variable with $P(N > 1) > 0$. The observation times and measurements are assumed to be independent of the number of observations for any subset $J_i \in \{1, ..., N_i\}$ and for all $i = 1, \ldots, n$, i.e. $\{T_{ij}, X_{ij}, Y_{ij} : j \in J_i\}$ is independent of N_i .

Let $K_1(\cdot)$ and $K_2(\cdot,\cdot)$ be the nonnegative univariate and bivariate kernel functions for smoothing the mean functions μ_X and μ_Y, and auto-covariance surface, G_X, and cross-covariance surface, G_{XY}. Assume that K_1 and K_2 are densities with zero means and finite variances on a compact support.

(A3) The Fourier transformations of $K_1(u)$ and $K_2(u,v)$, defined by $\kappa_1(t) = \int e^{-iut}K_1(u)du$ and $\kappa_2(t,s) = \int\int e^{-(iut+ivs)}K_2(u,v)dudv$ are required to be absolutely integrable, i.e. $\int |\kappa_1(t)|dt < \infty$ and $\int\int |\kappa_2(t,s)|dtds < \infty$.

Let h_X and h_Y be the bandwidths used for estimating μ_X and μ_Y, respectively. Also let h_G be the bandwidth used for estimating G_X, and let (h_1, h_2) be bandwidths used in the estimation of G_{XY}.

(A4) As $n \to \infty$, the following are assumed about the bandwidths.

(A4.1) $h_X \to 0$, $h_Y \to 0$, $nh_X^4 \to \infty$, $nh_Y^4 \to \infty$, and $nh_X^6 < \infty$, $nh_Y^6 < \infty$.

(A4.2) $h_G \to 0$, $nh_G^6 \to \infty$, and $nh_G^8 < \infty$.

(A4.3) Without loss of generality, $h_1/h_2 \to 1$, and $nh_1^6 \to \infty$, $nh_1^8 < \infty$.

(A5) Assume that the fourth moments of Y and X are finite.

(B1) The number of eigenfunctions used in (5.10), $M = M(n)$, is an integer valued sequence that depends on sample size n and satisfies the rate conditions given in assumption (B5) of Yao et al. (2005a).

(B2) The number and locations of measurements for a given subject does not change as the sample size $n \to \infty$.

(B3) For all $1 \leq i \leq n, m \geq 1$ and $1 \leq \ell \leq N_i$, the functional principal component scores ζ_{im} and the measurement errors $\varepsilon_{i\ell}$ in (5.3) are jointly Gaussian.

(C1) There exists a continuous positive definite function $\omega(s,t)$ such that $\omega_M(s,t)$ $\to \omega(s,t)$, as $M \to \infty$.

Appendix C

Proofs

Proof of Theorem 1: Uniform consistency of $\widehat{G}_X(s,t)$ is given in Theorem 1 of Yao et al. (2005b) and that of $\widehat{G}_{XY}(s,t)$ is given in Lemma A.1 of Yao et al. (2005a). Consistency of $\hat{\nu}_{kl}(t)$ and $\hat{\theta}_k(t)$ for ν_{kl} and $\theta_k(t)$ follow from uniform consistency of $\widehat{G}_X(s,t)$ and $\widehat{G}_{XY}(s,t)$. This implies consistency of $\hat{b}(t)$ for $b(t)$, and hence that of $\hat{\beta}(s,t)$.

Proof of Theorem 2: For fixed M, define $\widetilde{Y}^*_M(t) = \mu_Y(t) + \sum_{m=1}^{M} \tilde{\zeta}^*_m P_m(t)$, where $\tilde{\zeta}_m$ is as defined in (5.9) and $P_m(t) = \int_{\Delta_t} \beta(s,t)\psi_m(s)ds$. Then, it follows that

$$|\widehat{Y}^*_M(t) - \widetilde{Y}^*(t)| \leq |\widehat{Y}^*_M(t) - \widetilde{Y}^*_M(t)| + |\widetilde{Y}^*_M(t) - \widetilde{Y}^*(t)| = Q_1 + Q_2.$$

The convergence of Q_2 to 0 as $n \to \infty$ follows from Lemma A.3 in Yao et al. (2005a). Note that, for Q_1,

$$Q_1 = |\widehat{Y}^*_M(t) - \widetilde{Y}^*_M(t)| \leq |\hat{\mu}_Y(t) - \mu_Y(t)| + \sum_{m=1}^{M} |\hat{\zeta}^*_m \widehat{P}_m - \tilde{\zeta}^*_m P_m(t)|.$$

Uniform consistency of $\hat{\mu}_Y(t)$ for $\mu_Y(t)$ follows from Theorem 1 in Yao et al. (2005b), and consistency of $\hat{\zeta}^*_m$ for $\tilde{\zeta}^*_m$ follows from Theorem 3 in Yao et al. (2005b). From uniform consistency of $\hat{\beta}(s,t)$ established in Theorem 1 of Section 5.3.1 and that of $\hat{\psi}_k(t)$ shown in Yao et al. (2005a), uniform consistency of $\widehat{P}_m(t)$ follows. Combining these results, we have

$$\sup_{t\in[\delta_1,T]} |\widehat{Y}^*_M(t) - \widetilde{Y}^*_M(t)| \xrightarrow{p} 0 \;\text{ as }\; n \to \infty,$$

and by Slutsky's Theorem, Theorem 2 follows.

Proof of Theorem 3: Under the Gaussian assumption, for any fixed $M \geq 1$, we have $\tilde{\zeta}^*_M - \zeta^*_M \sim \mathcal{N}(0, \Omega_M)$. In the proof of Theorem 2, it is shown that

$$\lim_{n\to\infty} \sup_{t\in\mathcal{T}} |\widehat{Y}^*_M(t) - \widetilde{Y}^*_M(t)| \xrightarrow{p} 0.$$

Observing that $\widehat{Y}^*_M(t) - Y^*_M(t) = \widehat{Y}^*_M(t) - \widetilde{Y}^*_M(t) + \widetilde{Y}^*_M(t) - Y^*_M(t)$, we have $\{\widehat{Y}^*_M(t) - Y^*_M(t)\} \xrightarrow{\mathcal{D}} Z_M \overset{\mathcal{D}}{=} \mathcal{N}(0, \omega_M(t,t))$. Under assumption (C1), letting $M \to \infty$ leads to $Z_M \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0, \omega(t,t))$. From the Karhunen-Loéve Theorem, $|Y^*_M(t) - Y^*(t)| \xrightarrow{p} 0$ as $M \to \infty$. Therefore, $\lim_{M\to\infty} \lim_{n\to\infty} |\widehat{Y}^*_M(t) - Y^*(t)| \overset{\mathcal{D}}{=} Z$. From the convergence of $\hat{\psi}(t)$, $\hat{\zeta}^*_m$, $\hat{\gamma}_m$ and $\widehat{P}_m$ for $\psi(t)$, ζ^*_m, γ_m and $\widetilde{P}_m$, we can deduce $\hat{\omega}_M(t,t) \xrightarrow{p} \omega_M(t,t)$ as $n \to \infty$. Under the assumption (C1), it follows that $\lim_{M\to\infty} \lim_{n\to\infty} \hat{\omega}_M(t,t) = \omega(t,t)$ in probability. Applying Slutsky's theorem, (5.11) follows.

Appendix D

Estimation Procedures

In this section, we provide explicit forms for the local polynomial smoothing procedures used in estimating the mean functions and covariance surfaces. Eigen-decompositions for the estimated covariance surfaces and the explicit form of the measurement error variance estimator are also provided.

The estimator of the mean function for the predictor process, $\hat{\mu}_X(t)$, can be obtained by local linear regression via minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i} K_1\left(\frac{T_{ij}-t}{h_X}\right)\{X_{ij}-\eta_0-\eta_1(t-T_{ij})\}^2$$

with respect to η_0, η_1, which leads to $\hat{\mu}_X(t) = \hat{\eta}_0$. Estimation of $\mu_Y(t)$ follows similarly.

For estimation of the cross-covariance surface G_{XY}, the two dimensional local linear smoother is fitted to the raw covariances by minimizing

$$\sum_{i=1}^{n}\sum_{1\le j,\ell\le N_i} K_2\left(\frac{T_{ij}-s}{h_1},\frac{T_{i\ell}-t}{h_2}\right)[G_{XY,i}(T_{ij},T_{i\ell})-f\{\eta,(s,t),(T_{ij},T_{i\ell})\}] \qquad \text{(D.1)}$$

with respect to $\eta = (\eta_0, \eta_1, \eta_2)$, yielding $\widehat{G}_{XY}(s,t) = \hat{\eta}_0$. Estimation of the auto-covariance surface of the predictor process can be obtained similarly where the diagonal elements of the raw auto covariance matrix are not included in the smoothing as described in Section 5.3.1. Hence the second sum in (D.1) is taken over $1 \leq j \neq \ell \leq N_i$, excluding the diagonal terms.

In order to obtain the estimator for the measurement error variance σ_ε^2, we first estimate the diagonal elements of the auto-covariance surface $G_X(t,t)$ excluding the diagonal raw covariances contaminated with error, by applying a local linear smoother along the diagonal and local quadratic smoother along the direction perpendicular to the diagonal. The resulting estimators of the diagonal elements are denoted by $\widetilde{G}_X(t)$. This estimator is then compared to a linear smoother fit only to the diagonal raw covariance terms $\{T_{ij}, G_{X,i}(T_{ij}, T_{ij})\}$, estimating $G_X(t,t) + \sigma_\varepsilon^2(t)$. This estimator is denoted by $\widehat{V}(t)$. We estimate the error variance by these two estimators, yielding

$$\hat{\sigma}_\varepsilon^2 = \frac{2}{T} \int_{\mathcal{T}_1} \{\widehat{V}(t) - \widetilde{G}(t)\} dt,$$

where $\mathcal{T}_1 = [T/4, 3T/4]$. Here, the integration is taken over the middle half of $\mathcal{T}$ in order to remove the boundary effect of the local polynomial smoother.

The eigenfunctions and eigenvalues of the estimated auto-covariance surface, $\widehat{G}_X(s,t)$, for the predictor process are the solutions, $\hat{\psi}_k$ and $\hat{\gamma}_k$, of the eigenequation given by

$$\int_{[0,T]} \widehat{G}_X(s,t)\hat{\psi}_k(s) ds = \hat{\gamma}_k \hat{\psi}_k(t),$$

where $\int_{[0,T]} \hat{\psi}_k^2(t) dt = 1$ and $\int_{[0,T]} \hat{\psi}_m(t)\hat{\psi}_k(t) dt = 0$ for $m \neq k$. For numerical solutions, discretization of the smoothed covariance function can be used following Rice and Silverman (1991).

Bibliography

[1] Marron, J. S., H. G. Müller, J. Rice, J. L. Wang, N. Wang, and Y. Wang (2004) "Discussion of nonparametric and semiparametric regression," *Statistica Sinica*, **14**, pp. 615–629.

[2] Rice, J. (2004) "Functional and longitudinal data analysis: Perspectives on smoothing," *Statistica Sinica*, **14**, pp. 631–647.

[3] Müller, H. G. (2005) "Functional Modeling and Classification of Longitudinal Data*," *Scandinavian Journal of Statistics*, **32**(2), pp. 223–240.

[4] Bosq, D. (2000) *Linear processes in function spaces: theory and applications*, Springer Verlag.

[5] Cardota, H., F. Ferraty, and P. Sardab (1999a) "Functional linear model," *Statistics and Probability Letters*, **45**, pp. 11–22.

[6] Ramsay, J. and C. Dalzell (1991) "Some tools for functional data analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**, pp. 539–572.

[7] Ramsay, J. and B. Silverman (2002) *Applied functional data analysis: methods and case studies*, Springer Verlag.

[8] ——— (2005) *Functional Data Analysis*, Springer-Verlag, New York.

[9] Malfait, N. and J. Ramsay (2003) "The historical functional linear model," *The Canadian Journal of Statistics*, **31**(2), pp. 115–128.

[10] Hastie, T. and R. Tibshirani (1993) "Varying-coefficient models," *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, pp. 757–796.

[11] Staniswalis, J. and J. Lee (1998) "Nonparametric Regression Analysis of Longitudinal Data." *Journal of the American Statistical Association*, **93**(444), pp. 1403–1404.

[12] Dauxois, J., A. Pousse, and Y. Romain (1982) "Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference," *Journal of Multivariate Analysis*, **12**(1), pp. 136–154.

[13] Cardot, H., F. Ferraty, A. Mas, and P. Sarda (2003) "Testing hypotheses in the functional linear model," *Scandinavian Journal of Statistics*, pp. 241–255.

[14] Cardot, H., F. Ferraty, and P. Sarda (2003) "Spline estimators for the functional linear model," *Statistica Sinica*, **13**(3), pp. 571–592.

[15] Marx, B. and P. Eilers (1999) "Generalized linear regression on sampled signals and curves: a P-spline approach," *Technometrics*, **41**(1), pp. 1–13.

[16] Crambes, C., A. Kneip, and P. Sarda (2009) "Smoothing splines estimators for functional linear regression," *Annals of Statistics*, **37**(1), pp. 35–72.

[17] Cardot, H. (2006) "Conditional functional principal components analysis," *Scandinavian journal of statistics*, **34**(2), pp. 317–335.

[18] Hall, P. and J. Horowitz (2007) "Methodology and convergence rates for functional linear regression," *Annals of Statistics*, **35**(1), p. 70.

[19] Cardot, H. and J. Johannes (2010) "Thresholding projection estimators in functional linear models," *Journal of Multivariate Analysis*, **101**(2), pp. 395–408.

[20] Antoch, J., L. Prchal, M. Rosa, and P. Sarda (2008) "Functional linear regression with functional response: application to prediction of electricity consumption," *Functional and Operatorial Statistics*, pp. 23–29.

[21] He, G., H. Muller, and J. Wang (2000) "Extending correlation and regression from multivariate to functional data," *Asymptotics in statistics and probability: papers in honor of George Gregory Roussas*, p. 197.

[22] Yao, F., H. G. Müller, and J. Wang (2005a) "Functional linear regression analysis for longitudinal data," *Annals of Statistics*, **33**, pp. 2873–2903.

[23] Harezlak, J., B. Coull, N. Laird, S. Magari, and D. Christiani (2007) "Penalized solutions to functional regression problems," *Computational statistics and data analysis*, **51**(10), pp. 4911–4925.

[24] Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2002) *Analysis of longitudinal data*, Oxford University Press, USA.

[25] Wu, C. and K. Yu (2002) “Nonparametric varying-coefficient models for the analysis of longitudinal data,” *International Statistical Review/Revue Internationale de Statistique*, **70**(3), pp. 373–393.

[26] Hart, J. and T. Wehrly (1986) “Kernel regression estimation using repeated measurements data,” *Journal of the American Statistical Association*, **81**(396), pp. 1080–1088.

[27] Altman, N. (1990) “Kernel smoothing of data with correlated errors,” *Journal of the American Statistical Association*, **85**(411), pp. 749–759.

[28] Hart, J. (1991) “Kernel regression estimation with time series errors,” *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(1), pp. 173–187.

[29] Rice, J. and B. Silverman (1991) “Estimating the mean and covariance structure nonparametrically when the data are curves,” *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(1), pp. 233–243.

[30] Zeger, S. and P. Diggle (1994) “Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters,” *Biometrics*, **50**(3), pp. 689–699.

[31] Cheng, S. and L. Wei (2000) “Inferences for a semiparametric model with panel data,” *Biometrika*, **87**(1), p. 89.

[32] Fan, J. and I. Gijbels (1996) *Local polynomial modelling and its applications*, CRC Press.

[33] Hoover, D., J. Rice, C. Wu, and L. Yang (1998) “Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data,” *Biometrika*, **85**(4), pp. 809–822.

[34] Fan, J. and J. Zhang (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), pp. 303–322.

[35] Ruppert, D., S. Sheather, and M. Wand (1995) “An effective bandwidth selector for local least squares regression,” *Journal of the American Statistical Association*, **90**(432), pp. 1257–1270.

[36] James, G., T. Hastie, and C. Sugar (2000) “Principal component models for sparse functional data,” *Biometrika*, **87**, pp. 587–602.

[37] Rice, J. and C. Wu (2001) “Nonparametric mixed effects models for unequally sampled noisy curves,” *Biometrics*, **57**(1), pp. 253–259.

[38] Sentürk, D. and H. G. Müller (2008) "Generalized varying coefficient models for longitudinal data," *Biometrika*, pp. 653–666.

[39] Müller, H. and Y. Zhang (2005) "Time-Varying Functional Regression for Predicting Remaining Lifetime Distributions from Longitudinal Trajectories," *Biometrics*, **61**(4), pp. 1064–1075.

[40] Fan, J., Q. Yao, and Z. Cai (2003) "Adaptive varying-coefficient linear models," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **65**(1), pp. 57–80.

[41] Schumaker, L. (2007) *Spline functions: basic theory*, Cambridge Univ Pr.

[42] Fleming, T. and D. Harrington (1991) *Counting processes and survival analysis*, Wiley, New York.

[43] Sentürk, D. and H. G. Müller (2010) "Functional varying coefficient models for longitudinal data," *Technical report, The Pennsylvania State University.*

[44] Ash, R. and M. Gardner (1975) *Topics in stochastic processes*, Academic Press, New York.

[45] Wu, C., C. Chiang, and D. Hoover (1998) "Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data." *Journal of the American Statistical Association*, **93**(444), pp. 1388–1389.

[46] Fan, J. and W. Zhang (2008) "Statistical methods with varying coefficient models," *Statistics and its interface*, **1**(1), pp. 179–195.

[47] Kauermann, G. and G. Tutz (1999) "On model diagnostics using varying coefficient models," *Biometrika*, **86**(1), pp. 119–128.

[48] Huang, J., C. Wu, and L. Zhou (2002) "Varying-coefficient models and basis function approximations for the analysis of repeated measurements," *Biometrika*, **89**(1), pp. 111–128.

[49] ——— (2004) "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, **14**(3), pp. 763–788.

[50] Huang, J. and H. Shen (2004) "Functional coefficient regression models for non-linear time series: a polynomial spline approach," *Scandinavian journal of statistics*, **31**(4), pp. 515–534.

[51] Chiang, C., J. Rice, and C. Wu (2001) "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables," *Journal of the American Statistical Association*, **96**(454), pp. 605–619.

[52] Yao, F., H. G. Müller, and J. Wang (2005b) "Functional Data Analysis for Sparse Longitudinal Data." *Journal of the American Statistical Association*, **100**, pp. 577–591.

[53] Murtaugh, P., E. Dickson, G. Van Dam, M. Malinchoc, P. Grambsch, A. Langworthy, and C. Gips (1994) "Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits," *Hepatology*, **20**(1), pp. 126–134.

[54] Han, S., N. Serban, and W. Rouse (2007) "Novel Perspectives on Market Valuation of Firms via Functional Regression," *Technical report, H. Milton Stewart School of Industrial Systems and Engineering Georgia Institute of Technology.*

[55] Di, C., C. Crainiceanu, B. Caffo, and N. Punjabi (2009) "Multilevel functional principal component analysis," *The annals of applied statistics*, **3**(1), p. 458.

[56] Wu, H. and H. Liang (2004) "Backfitting random varying-coefficient models with time-dependent smoothing covariates," *Scandinavian Journal of Statistics*, pp. 3–19.

Vita

Kion Kim

Kion Kim was born in Daegu, South Korea on December 3, 1977. After completing high school at Keisung High School, South Korea in 1996, he attended the Yonsei University in Seoul, South Korea from 1996 to 2003. While attending college, he performed his military service in the Republic of Korea army. He graduated with a Bachelor of Art degree in Business Administration and Applied Statistics in 2003. From 2003 to 2005, he attended the graduate school of Yonsei University and graduated with a Master of Art degree in Applied Statistics in 2005. He then entered the Pennsylvania state University in the fall of 2005. He earned a Ph.D. in statistics from the Pennsylvania State University in December 2010.